

基于 Web 2.0 的边与节点同时增长网络模型^{*}

熊 菲[†] 刘 云 司夏萌 丁 飞

(北京交通大学通信与信息系统北京市重点实验室, 北京 100044)

(2009 年 11 月 20 日收到; 2010 年 2 月 3 日收到修改稿)

模拟了 Web 2.0 网络的发展过程并研究其拓扑结构, 分析某门户网站实际博客数据的度分布、节点度时间变化, 发现与先前的无标度网络模型有所差别. 根据真实网络的生长特点, 提出了边与节点同时增长的网络模型, 包括随机连接及近邻互联的网络构造规则. 仿真研究表明, 模拟的网络更接近实际, 在没有优先连接过程时, 模型能得到幂率的度分布; 并且网络有更大的聚类系数以及正的度相关性.

关键词: 复杂网络, Web 2.0 网络, 博客, 幂率分布

PACC: 0590, 0175

1. 引 言

复杂系统科学的发展及其在通信、自动控制、生物技术等方面的应用, 使其获得了广泛的关注. 科学合作网络^[1,2]、生物网络^[3,4]、蛋白质弹性网络^[5]、互联网络^[6,7]、电阻器网络^[8]等复杂网络的研究不断涌现. Erdos 和 Renyi 建立了随机图理论, 节点之间随机连接, 网络度分布为 Poisson 分布, 平均最短距离较小^[9]. Watts 和 Strogatz 提出了小世界网络模型, 将规则网络的边随机重连, 网络的性质介于随机网络和规则网络之间, 因此能找到适当的重连概率, 使网络同时拥有大的聚类系数和小的平均最短路径长度^[10]. 鉴于较多的实际网络表现为幂率的度分布, Barabasi 和 Albert (BA) 研究了一种无标度网络的构造方法, 包括节点增长过程及优先连接过程. 这两个步骤的存在, 使网络的度分布为幂率分布, 幂指数恒为 -3 ^[11,12], 且度较大处邻居度之间无相关性^[13]. 相继出现了大量 BA 网络的扩展模型, 修改了节点增长或连接等规则^[14-20]. 文献[14]提出了节点数加速增长网络模型, 并分析了节点数增长率与网络平均最短距离及聚类系数的关系. 文献[15]在优先连接的基础上, 引入了三角连接规则,

提高了网络聚类系数, 但网络仍然有幂率的度分布.

由于观点动力学和传染病模型对网络传播介质的依赖, 在研究个体状态演化时也不能忽略网络的演化, 从而得到了不同的系统终态^[21,22]. Balazs 在 Deffuant 模型基础上, 考查了观点与网络拓扑同时演化机理, 网络的边不断调整, 仿真研究表明, 网络拓扑簇数不断增加, 最终趋于稳定, 并滞后于观点的演化^[23,24]. Zhou 等考虑了网络拓扑对谣言传播的影响, 用解析的方法将网络的度分布加入到状态方程中, 并分析了网络度的相关性及其与传播进程的关系^[25-27].

信息采集及处理能力的提高, 使得对大规模复杂网络的实证研究成为了可能. 吴晔等^[28,29]分析手机短信网络, 网络论坛等实际数据, 统计了网络度的变化, 并建立了相应的网络生长模型, 数值模拟的网络具有与实际网络相似的性质.

本文研究以博客网络为首的 Web 2.0 网络. Web 2.0 网络具有高度的自组织性, 网络的演化过程和经典无标度网络有显著区别, 且节点的度分布、度与加入时间的关系、平均度随时间变化与传统网络不尽相符. 根据博客网络的发展特点, 构建了网络边与节点同时增长的网络模型, 模拟网络的度分布尤其是分布的尾部、度的时间变化与实际数

^{*} 国家自然科学基金(批准号:60972012)、北京市自然科学基金(批准号:4102047)、教育部哲学人文社会科学研究重大课题(批准号:08WL1101)、北京市教育委员会学科建设与研究生建设项目(批准号:JXKJD20090001)和科技人员服务企业计划(2009GJA00048)资助的课题.

[†] E-mail: xfl985@eyou.com

据较好的契合,因而更好地描述了真实的博客网络.

2. BA 无标度网络

BA 无标度网络具有幂率的度分布得益于网络构造算法的两个过程:节点增长和节点的优先连接.

(1) 节点增长:从具有 m 个节点的网络开始,每个时步向网络添加 1 个新节点以及从该节点出发的 n 条边,分别连接到已有的节点上,满足 $n \leq m$.

(2) 节点的优先连接:新节点连接到已有节点的概率,与该已有节点的度成正比. 即新节点连接到节点 i 的概率为

$$\prod (k_i) = k_i / \sum_j k_j. \quad (1)$$

按照上述两个规则所构造网络的度服从幂率分布 $P(k) \propto k^{-\gamma}$, 幂指数 $\gamma = -3$, 如图 1 所示.

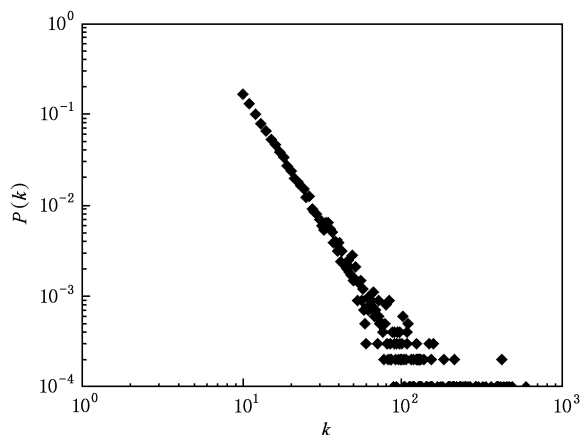


图 1 BA 网络的度分布

网络边的增长完全由节点的增加所导致,因此那些早期加入的节点,具有更大的度,其吸附新节点的能力也越强,呈现“富者越富”的现象. 在平均场的角度,节点的度随时间幂率增加,若 t_i 时步节点 i 加入网络,在此后的任意 t 时步,节点 i 的度为

$$k_i(t) = n \sqrt{t/t_i}. \quad (2)$$

任意时步 t , 节点的度按其加入网络的时间幂率衰减. 同时, BA 网络为度不相关网络,其平均邻居度分布如图 2 所示.

BA 网络的幂率度分布及度幂率增加的性质,符合一些真实复杂网络的特征. 但对于以博客为代表的 Web 2.0 网络,不满足 BA 网络的上述特性,且 BA 网络的节点增长过程不足以描述 Web 2.0 网络

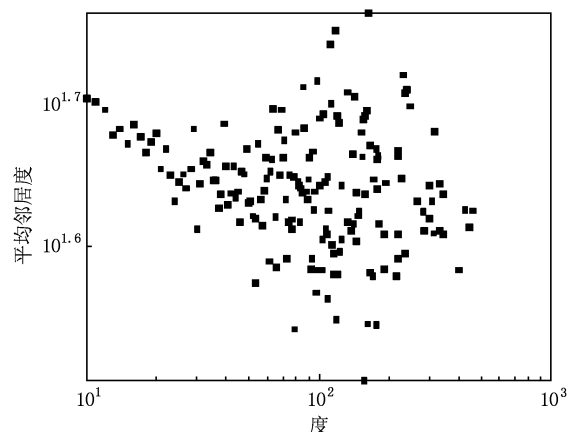


图 2 BA 网络平均邻居度

的演化.

3. Web 2.0 网络

近年来, Web 2.0 应用出现并得到了迅速发展, Web 2.0 网站大量涌现,已经成为互联网的重要组成部分. Web 2.0 是以 Blog, RSS, wiki 等社会软件的应用为核心、六度分隔、xml, ajax 等新理论和技术实现的互联网新一代模式. Web 2.0 以人为本,提供个性化的服务,用户即作者,既是网站内容的消费者,也是网站内容的制造者. 典型的 Web 2.0 网站主要包括一些以博客和社会网络应用为主的网站,尤其以博客网站发展最为迅速,影响力也更大. 数据显示有 76.8% 的用户使用过博客,视频分享和网络社区的月度覆盖人数超过 1 亿.

用户创造内容的 Web 2.0 网络,具有较强的互动性和自组织性. 作为代表的博客网站,网络发展是用户自愿的行为. 每个用户均有自己的主页,用户在自己的主页上通过博文发表意见,并评论其他用户的文章,网站并无集中的控制. 用户在访问他人的主页后,对感兴趣的或意见相近的用户,添加其为好友. 用户主页上通常会显示好友动态,好友能获得更多的关注,因而在两个用户之间形成了强联系. 用户作为网络的节点,若节点 i 主页有指向节点 j 的链接,相应地,节点 j 主页存在着通往节点 i 的路径, i 和 j 之间的联系可用一条无权无向的边表示. 博客网络组织形式如图 3 所示.

分析真实博客网络的发展过程,与 BA 网络构造相异,具有以下一些实际特性: (1) 网络持续增长. 网站运行期间,不断有新用户注册加入网络,并

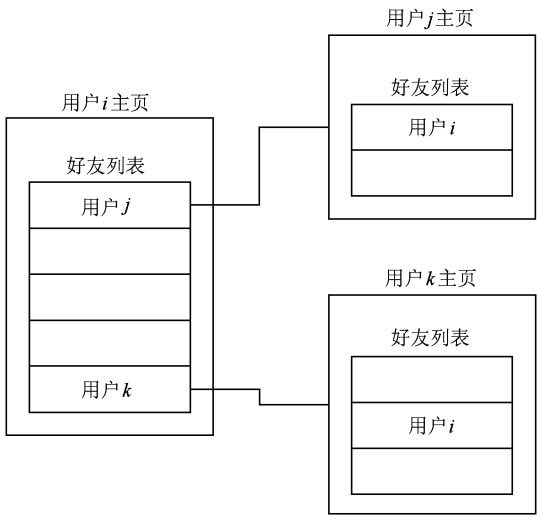


图3 博客网络用户组织方式

带来与其他用户的连接。(2)连接的随机性. 新用户加入网络后,随机访问一些用户,与部分用户建立好友关系. 由于不存在全局的监控者,新用户仅了解少数的网站推荐用户的信息,在访问其他用户之前,无法确知其的热度. 因此用户间的访问与好友的建立并非是优先连接,而更多是随机的选择。(3)网络连通性持续增强. 即使没有新用户加入,已有用户通过随机的交互,使网络的边不断增加。(4)近邻互联. 已有用户间的交互,主要以共同的好友为中介. 用户访问好友的主页,并通过其页面上提供的链接,随机访问好友的好友。

选取国内某著名门户网站的博客数据,从2007年9月以来,跨度两年的20多万份网页,总计14002个用户,并获取各用户之间的连接. 以用户为节点,用户的好友链接为边,去掉用户自环、重复等冗余连接,统计网络性质。

图4为博客网络的度分布. 可见博客网络的度服从幂率分布,但网络度分布的“长尾效应”不如BA网络明显,度分布尾部不平坦,有明显的截断效果. 并且度分布的幂率指数为 $\gamma = -1.5786 \pm 0.0119$,与BA网络的幂指数 $\gamma = -3$ 差别较大。

BA网络节点度随加入时间幂率衰减,但博客网络如图5,节点的度与节点加入时间不再有严格的关系,而是随时间剧烈波动. 大量的节点拥有很小的度,少量的节点度较大,节点度随加入时间存在一定的局部聚集性. 度大节点的出现比较均匀,在网络演化进程中,不断有较大幅度的节点产生。

以月为单位,跨度两年,计算博客网络平均度

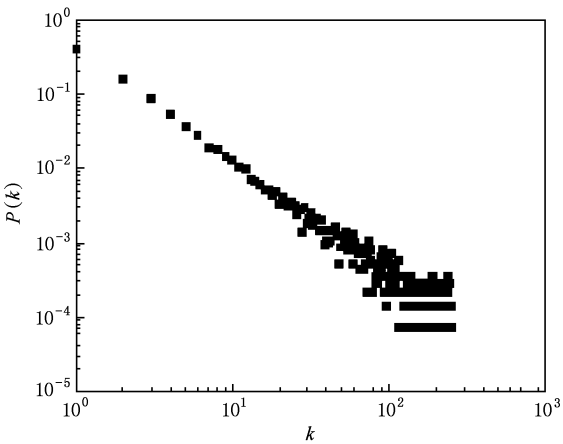


图4 博客网络的度分布

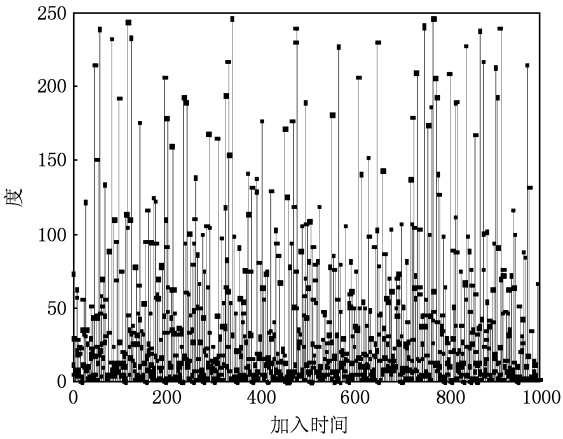


图5 博客网络节点度与加入时间关系

随时间的变化,如图6所示. 从图6可以看出,网络的平均节点度随时间持续增长. 在网络节点较少时,平均度增长快,随着新节点的不断加入,增长速度逐渐放缓. 这与BA网络平均度的时不变性质明显不同。

由此可见,博客网络的度分布与度的时变性质,均与BA网络有显著差别,网络生长特点也与BA网络有所不符,因而BA网络构造算法对博客网络建模不够完备。

4. 边与节点同时增长网络模型

基于以博客为首的Web 2.0网络的生长过程及用户交互关系,建立边与节点同时增加的网络模型

(1)节点增长. 初始有一 m 个由节点组成的网络,每个时步,将一个新节点添加至网络中,与已有

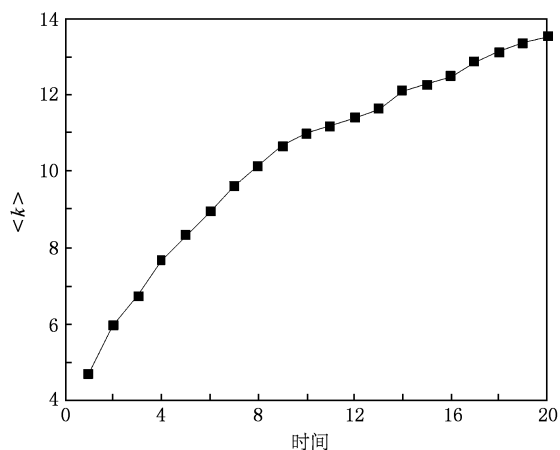


图6 博客网络平均度的增长

节点相连.

(2) 随机连接. 在已经存在的节点中, 随机选择 n 个, 与新节点建立连接, 满足 $n \leq m$.

(3) 边自发增长. 每个时步, 随机选择一个节点 i 及其邻居节点 j , i 以 p 的概率连接 j 的邻居.

从网络模型的 3 个增长规则可以看出, 网络边的增长, 来源于两个方面: 新节点的引入带来的从其出发的连接, 以及已经存在节点通过公共邻居互相连接. 网络连通度的增大不依赖于节点的加入, 在停止引入节点后, 网络的边仍然随时间持续增加. 由于连接的随机性, 节点的度不会随加入网络时间严格衰减. 网络的生长过程与实际博客网络相似, 对应了博客网络用户随机访问, 通过共同好友互相认识, 网络链接与用户的增加不同步等特性.

将上述边与节点同时增长网络模型进行数值模拟, 分析网络性质.

4.1. 度分布

从图 7 可以看出, 模拟网络的度服从幂率分布, 幂指数 $\gamma = -1.6184$. 度分布的尾部有截断现象, 与实际博客网络较好的相符. 即使网络的构建中并无优先连接过程, 网络仍然表现出指数可调的幂率度分布. 内插图为幂指数 γ 与近邻互联概率 p 的关系. 注意到 $p < 0.1$ 时, 节点增长起主要作用, 网络表现为随机网络的特征; 而 $p > 0.4$ 时, 网络连通性过大, 在度较大之处已偏离了幂率分布. 同时 m, n 变化对网络度分布的影响不明显.

图 8 为 1000 连续加入节点的度与加入时间的关系. 与 BA 网络不同, 仿真的网络节点度随加入时间的起伏很大, 存在着局部聚集, 较好的符合实际

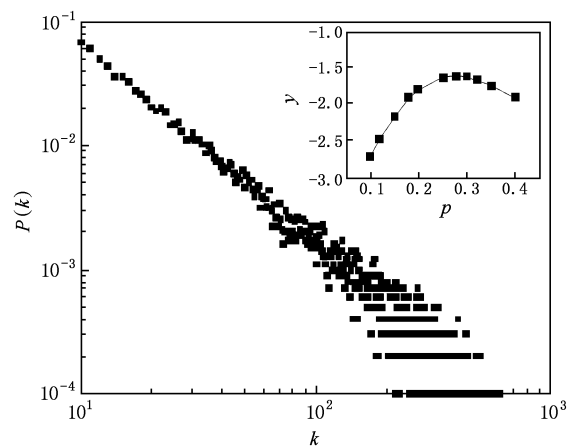


图7 同时增长模型的度分布

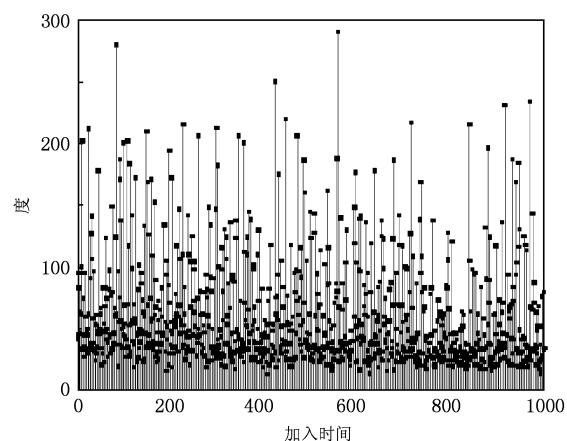


图8 模拟网络节点度与加入时间关系

网络.

图 9 为模拟网络的平均节点度, 等间隔取样. 可见, 在网络演化初期, 网络平均度增长快, 达到一定系统规模后, 缓慢增加, 与实际网络基本一致.

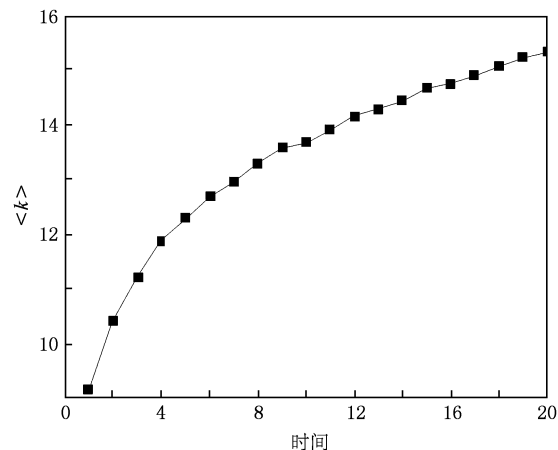


图9 模拟网络平均度的增长

4.2. 聚类系数

聚类系数衡量了网络节点聚集程度,反映了节点邻居之间的连通性.若节点 i 的度为 k_i ,则节点 i 的聚类系数 c_i 为

$$c_i = \frac{\sum_j a_{ij} \sum_l a_{il} a_{jl}}{k_i(k_i - 1)}. \quad (3)$$

若节点 i 与 j 之间有连接, $a_{ij} = 1$, 否则 $a_{ij} = 0$. 节点总数为 N 的网络平均聚类系数 C 为

$$C = \frac{1}{N} \sum_i c_i, \quad (4)$$

计算模型的平均聚类系数与 p 、 n 的关系,如图 10 所示.

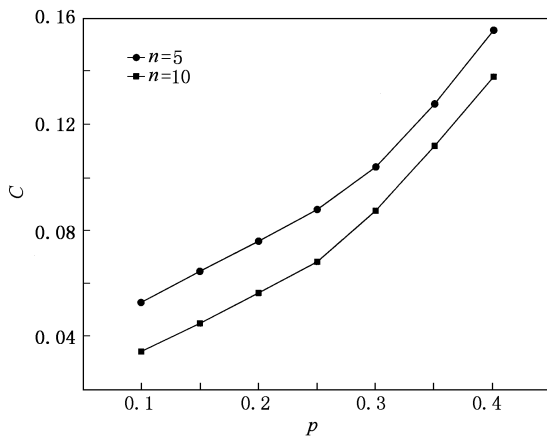


图 10 模拟网络的聚类系数

节点总数 5000, BA 网络在 $n = 5$, $n = 10$ 时聚类系数分别为 0.0125, 0.0177, 且随着系统规模的增大迅速减小. 如图 10 所示, 边与节点同时增长的网络, 与 BA 网络相比, p 适当时具有更大的聚类系数. p 的增大, 将使聚类系数非线性增长, 减小 n , 也能提高聚类系数.

4.3. 度的相关性

度的相关性是网络的重要性质之一, 描述了网络的相邻节点分布特征. 度的相关性可由平均最近邻居度表示. 度为 k_i 的节点 i 的平均邻居度 $k_{i, \text{neigh}}(k_i)$ 为

$$k_{i, \text{neigh}}(k_i) = \frac{\sum_j a_{ij} k_j}{k_i}. \quad (5)$$

则网络度 k 节点的平均邻居度 $k_{\text{neigh}}(k)$ 为

$$k_{\text{neigh}}(k) = \frac{\sum_i k_{i, \text{neigh}}(k)}{N_k} = \sum_{k'} k' P(k' | k). \quad (6)$$

计算模拟网络的平均邻居度分布, 如图 11 所示.

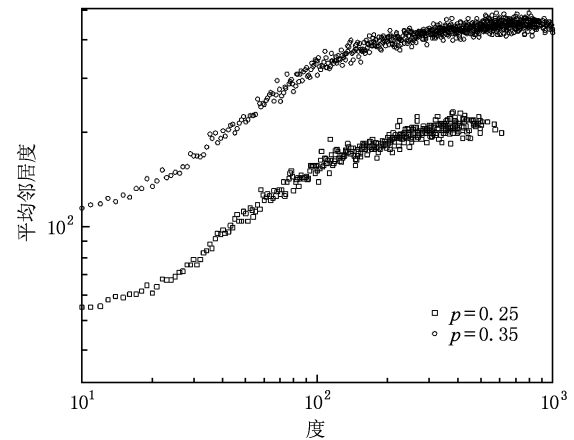


图 11 模拟网络的平均邻居度

从图 11 可以看出, 度较大的节点, 平均邻居度也较大, 网络具有正的度相关性, 因此出现了接近实际“富人俱乐部”的现象. 例如博客网络中影响力大的用户常常因为互相吸引而成为好友. 且增大 p , 平均邻居度整体提高, 邻居度的变化率在度相同处近似相等, 伴随着更大度节点的出现. 而在 BA 网络中, 度相关系数为 0, 平均邻居度分布较随机.

进一步地, 对于节点 i , 度的增加来自于三个方面: (1) 新节点的加入, 带来 n 条边, 节点 i 被随机选择与新节点相连; (2) 节点 i 被选择与其邻居的邻居相连; (3) 节点 i 邻居的邻居被选择, 以一定的概率连接节点 i . 因此, 节点 i 的度 k_i 增长率为

$$\begin{aligned} \frac{\partial k_i}{\partial t} &= \frac{n}{t+m} + \frac{p}{(t+m)k_i} \sum_{j \in \Gamma(i)} k_j \\ &\quad + \frac{p}{t+m} \sum_{j \in \Gamma(i)} \sum_{l \in \Gamma(j)} \frac{1}{k_l} \\ &\propto \frac{1}{k_i} \sum_{j \in \Gamma(i)} k_j = k_{i, \text{neigh}}. \end{aligned} \quad (7)$$

$\Gamma(i)$ 为 i 的邻居节点集. 可见, k_i 的增长率不与 k_i 直接相关, 而正比于平均最近邻居度, 邻居节点度的增大促使 k_i 增加, 从而产生了正的相关度.

5. 结 论

本文分析了以博客网站为代表的 Web 2.0 网络性质, 比较实际数据与 BA 无标度网络在度分布、度的时间变化等方面的不同. 根据博客网络用户的行为特征和网络演化规律, 建立了边与节点同时增

长的网络模型. 节点随机连接, 网络边的增长不依赖于节点的增加. 对此模型进行数值仿真, 网络的

度分布更好的符合了博客网络, 具有可调的度分布指数及聚类系数, 并有接近实际的正相关度.

-
- [1] Newman M E J 2001 *Phys. Rev. E* **64** 016132
 - [2] Newman M E J 2001 *Proc. Nat. Acad. Sci.* **98** 404
 - [3] Jeong H, Tombor B, Albert R, Oltvai Z N, Barabasi A L 2000 *Nature* **407** 651
 - [4] Lee T I, Rinaldi N J, Robert F, Odom D T, Bar-Joseph Z, Gerber G K, Hannett N M, Hannett C T, Thompson C M, Simon I, Zeitlinger J, Jennings E G, Murray H L, Gordon D B, Ren B, Wyrick J J, Jean-Bosco T, Volkert T L, Fraenkel E, Gifford D K, Young R A 2002 *Science* **298** 799
 - [5] Dietz H, Rief M 2008 *Phys. Rev. Lett.* **100** 098101
 - [6] Adamic L A, Huberman B A, Barabasi A, Albert R, Jeong H, Bianconi G 2000 *Science* **287** 2115
 - [7] Pastor-Satorras R, Vazquez A, Vespignani A 2001 *Phys. Rev. Lett.* **87** 258701
 - [8] Juanico B, Sanejouand Y H 2007 *Phys. Rev. Lett.* **99** 238104
 - [9] Erdos P, Renyi A 1959 *Publ. Math.* **6** 290
 - [10] Watts D J, Strogatz S H 1998 *Nature* **393** 440
 - [11] Barabasi A L, Albert R 1999 *Science* **286** 509
 - [12] Albert R, Barabasi A L 2002 *Phys. Rev. Mod.* **74** 47
 - [13] Huang Z X, Wang X R, Zhu H 2004 *Chin. Phys.* **13** 273
 - [14] Li J, Xiao J H, Jiang P Q, Zhou T, Wang W X 2006 *Acta Phys. Sin.* **55** 4051 (in Chinese) [李 季、肖井华、蒋品群、周涛、王文旭 2006 物理学报 **55** 4051]
 - [15] Pan Z F, Wang X F 2006 *Acta Phys. Sin.* **55** 4058 (in Chinese) [潘灶烽、汪小帆 2006 物理学报 **55** 4058]
 - [16] Hwang D U, Boccaletti S, Moreno Y, Lopez-Ruiz R 2005 *Math. Biosci. Engng.* **2** 317
 - [17] Klemm K, Eguiluz V M 2002 *Phys. Rev. E* **65** 036123
 - [18] Wang W X, Wang B H, Hu B, Yan G, Ou Q 2005 *Phys. Rev. Lett.* **94** 188702
 - [19] Jost J, Joy M P 2002 *Phys. Rev. E* **66** 036126
 - [20] Hu M B, Jiang R, Wu Y H, Wang W X, Wu Q S 2008 *Physica A* **387** 4967
 - [21] Li P P, Zheng D F, Hui P M 2006 *Phys. Rev. E* **73** 056128
 - [22] Yan G, Fu Z Q, Ren J, Wang W X 2007 *Phys. Rev. E* **75** 016108
 - [23] Kozma B, Barrat A 2008 *Phys. Rev. E* **77** 016102
 - [24] Kozma B, Barrat A 2007 *Physica A* **41** 224020
 - [25] Zhou J, Liu Z H, Li B W 2007 *Phys. Lett. A* **368** 458
 - [26] Nekovee M, Moreno Y, Bianconi G, Marsili M 2007 *Physica A* **374** 457
 - [27] Moreno Y, Nekovee M, Pacheco A F 2004 *Phys. Rev. E* **69** 066130
 - [28] Wu H, Xiao J H, Wu Z Y, Yang J Z 2007 *Acta Phys. Sin.* **56** 2037 (in Chinese) [吴 晔、肖井华、吴智远、杨俊忠 2007 物理学报 **56** 2037]
 - [29] Zhang L, Liu Y 2008 *Acta Phys. Sin.* **57** 5419 (in Chinese) [张 立、刘 云 2008 物理学报 **57** 5419]

Network model with synchronously increasing nodes and edges based on Web 2.0^{*}

Xiong Fei[†] Liu Yun Si Xia-Meng Ding Fei

(Key Laboratory of Communication and Information Systems-Beijing Jiaotong University,
Beijing Municipal Commission of Education, Beijing 100044, China)

(Received 20 November 2009; revised manuscript received 3 February 2010)

Abstract

We investigate the growing process and topological features of Web 2.0 networks. By analyzing the network's degree distribution, average degree and time evolution of the node degree of an actual blog on portal website, we found these properties are different from those of the former scale-free network models. According to the growth characteristics of actual networks, we put forward a new type of network with synchronously increasing nodes and edges, including construction algorithms of randomly linking and connection between close neighbours. The simulation results show that the networks generated from our model have power-law degree distribution in case of absence of the preferential attachment process, and the clustering coefficient increases and the connectivity correlations are assortative.

Keywords: complex networks, Web 2.0 network, blog, power-law distribution

PACC: 0590, 0175

^{*} Project supported by the National Natural Science Foundation of China (Grant No. 60972012), the Beijing Natural Science Foundation (Grant No. 4102047), the Major Program for Research on Philosophy and Humanity Social Sciences of the Ministry of Education of China (Grant No. 08WL1101), the Academic Discipline and Postgraduate Education Project of Beijing Municipal Commission of Education (Grant No. JXKJD20090001) and the Service Business of Scientists and Engineers Project (Grant No. 2009GJA00048).

[†] E-mail: xf1985@eyou.com