



微博转发网络中的节点特征和传播模型

李晋¹, 杨子龙²

(1. 北京信息科技大学, 北京 100192;

2. 解放军电子工程学院网络系, 安徽 合肥 230037)

摘要: 微博作为重要的社交网络平台, 具有传播快速、平台影响大的特点。微博用户的节点特征决定了其网络影响力。研究了微博转发网络中节点的度值特征和传播模型。首先通过区分信息流动方向构建了微博转发网络; 其次分别讨论了出度—入度的均值和方差, 明确二者的差异, 并分析了考虑节点度值特征的信息传播过程; 最后通过仿真验证可以看出: 边的有向性对信息传播有着显著的影响, 在有向条件下, 渗流阈值增加, 同样概率下传播范围变小, 信息传播更为困难。

关键词: 转发网络; 微博; 度值特征; 信息传播

中图分类号: TN393

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2016006

Node characteristic and propagation model in microblog forwarding network

LI Jin¹, YANG Zilong²

1. Beijing University of Information Science and Technology, Beijing 100192, China

2. School of Network, Electronic Engineering Institute, Hefei 230037, China

Abstract: Microblog is an important social network with rapid propagation speed and great influence. The network influence is determined by users' node characteristic. Nodes' degree and propagation model in microblog forwarding network were investigated. Firstly, microblog forwarding network was constructed through distinguishing information flow direction. Secondly, the mean and variance of out-degree and in-degree were discussed. The difference between out-degree and in-degree was clarified. Finally, the simulation shows that the direction characteristic of edge has significantly influence on information propagation. The propagation becomes harder and propagation range diminishes while percolation threshold rises in directed graph under the same probability.

Key words: forwarding network, microblog, degree characteristic, information propagation

收稿日期: 2015-07-15; 修回日期: 2015-12-17

基金项目: 北京市教育委员会人文社会科学研究计划面上项目 (No. SM201411232005); 北京市属高校教师专项培训 2014 年中青年骨干教师一般国外访问学者研修培训项目 (No. 067145301400)

Foundation Items: Humanistic and Social Science Research Plan Project of Beijing Municipal Education Commission (No. SM201411232005), Young and Middle-Aged Backbone Teachers Training Program for Visiting Scholars Abroad in Universities Owned by the Municipal Government of Beijing in 2014 (No. 067145301400)

1 引言

随着在线社会网络的普及,微博、微信等平台逐渐渗透人们的生活,其影响力已经从虚拟网络拓展到实际生活中^[1]。因此,在线社会网络越来越集中到人们日常行为模式等领域。微博短文本形式使得信息传播更加快速,作为一种全新的在线社交应用,近年来得到了快速发展。由于微博这些前所未有的新特征以及其中信息传播产生的巨大社会效应,它已经成为了一把“双刃剑”。一方面微博为广大网民提供了一个更加自由便捷的信息获取和发布平台,推动了我国信息化社会的发展;另一方面,微博中也包含了谣言、诽谤、偏激观点、扭曲价值观等一系列负面问题,并在一些别有用心个人和组织的推动下,产生了极坏的影响。因此,提取、分析微博中的信息传播特征,构建相应的传播模型有着十分重要的理论和现实价值。对于已知的网络,如何评价网络的动力学性能一直是研究者关注的重点内容,当前主要利用网络疾病传播模型对信息传播进行研究,但疾病传播模型往往假设传播时间是无限长的。在很多现实情况中,关注的往往是一个时间段内信息的传播情况。SIR(susceptible infective removal)模型及其变型是研究信息传播最为常用的模型,在该模型中,个体仅在有限时间内保持感染态,随后便能恢复,甚至如果这些邻居个体足够幸运,就有可能永远不会得这种疾病。根据传染病模型中的SIR模型定义,利用积分可以求得传播概率,在一个固定时间段 τ 内,这种概率为:

$$\varphi = 1 - e^{-\beta\tau} \quad (1)$$

其中, β 表示传播速度。假设这些事件是独立的,即疾病在网络中沿任意一条边的传播是相互独立的,将 φ 作为一个常数,那么整个疾病传播过程可以认为是以概率占据每条边或者以概率 $1-\varphi$ 移除每条边,整个过程就可以看作一个“键渗流”过程。

利用渗流理论对微博的信息传播过程进行研究,通过构建动力学过程分析传播的影响因素,已经取得一定的成果:参考文献[2]主要分析度值对传播范围的映射关系,基于微博数据构建信息传播的动力学过程,利用复杂网络理论研究其机理;参考文献[3]主要针对在线社会网络中的观点形成过程进行研究,将其抽象为扩散过程,提出了舆论传播模型;参考文献[4]借鉴传染病模型,主要分析了不同的传播主体对传播结果的影响,对应3个不同状态将用户类型进行抽象,结合信息传播方向的影响,分析度值分布

的幂律特性对传播行为的影响;参考文献[5]在真实的微博数据中进行了渗流过程的仿真;参考文献[6]在服从幂律分布的随机网络中,通过随机删除和目标删除网络节点或边,计算不同占有概率下巨分支的大小来分析网络性能。

无向网络对于渗流阈值的求导依赖于邻接矩阵的对称性,因此原有的分析方法不能直接应用于有向网络,这是因为有向网络中的邻接矩阵通常是非对称的,无法得到正交特征向量。因此本文借鉴已有研究成果,利用数值方法分析有/无关联特征情况下的渗流过程,通过对比,得到有向网络下的传播能力分析。

2 基于微博信息的有向转发网络度值特征

2.1 基于微博信息的有向转发网络

在微博的信息传播过程中,用户 i 发布一条信息,其所有粉丝用户都会接到该信息,若其中一个用户 j 转发这条信息,则用户 j 的所有粉丝用户(如用户 k)都会继续接到信息,以此类推,该信息沿着多条 $i \rightarrow j \rightarrow k \rightarrow \dots$ 的路径在用户之间呈网状蔓延。实际中,微博用户数量庞大,信息传播过程复杂,如何衡量网络的传播能力是一个重要课题。本节通过采集微博用户转发行为的数据,抽象得到有向转发网络模型,分析信息传播和渗流过程的映射关系,设计传播能力的评估指标,为后续研究提供模型基础。

首先对微博用户及其行为进行抽象建模。微博转发网络抽象为 $G=(V,E)$,其中, G 是全部节点 v_i 的集合,对应微博的用户; E 是边 e_{ij} 的集合,对应用户之间的关系,这里的关系是“转发行为”,转发行为是信息传递的基础,如果节点 v_j 转发了邻接节点 v_i 的信息,则在传播过程中, $v_i \rightarrow v_j$ 将在下一次以更大的概率成为信息传递的路径,记为 e_{ij} ,否则 $e_{ij}=0$;指向 v_i 的节点数为节点 v_i 的入度,记为 k_{in} ; v_i 指向节点数为节点 v_i 的出度,记为 k_{out} 。

在信息传播过程中,只会有一部分个体以及个体间部分关系会参与到传播过程中来,所有的用户关系中只有部分关系会发挥作用,而这个过程同样能够用渗流过程表示。

在微博转发网络中,有一些随机均匀选择的节点被删除。微博转发过程是删除过程的“逆向”,一条边被保留意味着一条消息被转发,衡量的是保留概率,可以用概率参数 φ 表示,表示节点存在于网络中。因此, $\varphi=1$ 就表示网络中的所有节点都被保留(即没有节点被删除), $\varphi=0$ 表示没有节点被保留(即所有的节点都被删除)。在实际情况中,



网络中不可能只存在一个分支,初始时,所有节点被保留,但是随着节点被删除以后网络逐渐分割,如分裂为两个分支或者更多,将其中最大的一个分支称为巨分支。本文只考虑一个分支的情况,也就是分析巨分支的消亡过程,即从 $\varphi=1$ 到 $\varphi=0$ 的过渡。巨分支的形成和解体叫做渗流过渡。当网络中包含一个巨分支时,称这个网络是可渗流的,而出现渗流过渡的临界值就叫做渗流阈值。在很多方面,渗流过渡与泊松随机图中巨分支形成的相变很相似。在随机图中变化的不是被保留的节点,而是节点间连接的概率。然而,在这两种情况下,当网络中有足够多的点被删除时,巨分支就会分裂成一些小分支。

2.2 出度—入度分布

度(degree)是针对网络中的节点来说的,节点 i 的度 $k(i)$ 定义为与节点 i 相连的所有边的数目。度描述了个体的特性,而度分布描述了网络的整体特性。现实中的很多网络,如科学家合作网、捕食网络、国际互联网,都服从幂率分布,幂指数在 -3.4 和 -2.0 之间,而不是之前人们所认为的泊松分布。参考文献[7]中,研究了含权科学家合作网的幂律分布等统计特性;参考文献[8]中,将现实社交关系推广到合作网络中来衡量其特性;参考文献[1]中,系统地研究了大量在线社会网络的幂律分布特性。一般地,根据幂律分布可绘出一条向右偏斜得很厉害,拖着长长“尾巴”的累积分布曲线,它与钟形的泊松分布曲线有显著的不同。这种“长尾”分布表明,绝大多数个体的尺度很小,而只有少数个体的尺度相当大。

以往通常利用最小二乘法来拟合数据,但得出的结果经常会出现很大的误差,这主要是由以下两个原因引起的:

- 在“长尾”中出现的大的波动;
- 很难确定幂律分布成立的范围。

在分析度分布特性时可以从两方面着手,一是设定幂律分布成立的范围,二是和其他分布结合。这里假设度分布服从幂律分布和指数分布的双分布:

$$p(x) = \begin{cases} C(x/x_{\min})^{-\alpha}, & x \geq x_{\min} \\ C e^{-\alpha(x/x_{\min}-1)}, & x < x_{\min} \end{cases} \quad (2)$$

综上所述,需要在度分布特性分析中辨识两个参数,分别是 α 和 $x_{\min}$ 。

度分布参数辨识的流程如下。

(1)利用极大似然估计的方法,在给定 $x_{\min}$ 情况下,可求出:

$$\alpha = 1 + n \left[\sum_{i=1}^n \ln \frac{x_i}{x_{\min}} \right]^{-1} \quad (3)$$

(2)首先利用幂律分布的密度函数的 $\lg-\lg$ 图像或者利用 $x_{\min} - \hat{\alpha}$ 图像判定 $x_{\min}$ 的范围,然后从这个范围里选取对于实际数据最合适的 $x_{\min}$ 。选择 $x_{\min}$,令:

$$D = \max_{x \geq x_{\min}} |S(x) - P(x)| \quad (4)$$

取得最小值,其中 $S(x)$ 、 $P(x)$ 分别为拟合数据和原始数据的分布函数。

(3)对于拟合数据和原始数据,利用 K-S 统计量求出 p -value, p 的值越接近 1,则拟合数据与原始数据越吻合,通常当 $p \leq 0.1$ 时,认为利用幂律函数对原始数据进行的拟合失败。在求 p -value 的过程中,由于认为当 $x \leq x_{\min}$ 时,分布并不符合幂律分布,所以,利用幂律分布拟合出来的数据与原始数据相比是不完全的,忽略了 $x \leq x_{\min}$ 的数据,那么利用 K-S 统计量计算 p -value 比较拟合情况的结果也不精确。笔者希望可以使拟合数据在 $x \leq x_{\min}$ 时非常接近原始数据,而在 $x > x_{\min}$ 时,符合幂律分布。

(4)根据实际数据,再找出其他可能符合的分布,并得出拟合数据。利用似然比检验比较不同分布的拟合情况,从而确定符合原始数据的最佳分布。其中似然比检验用来评估两个模型中哪个模型更适合当前的数据分析。具体来说,一个相对复杂的模型与一个相对简单的模型比较,检验哪个模型能够显著地适合一个特定的数据集。引入统计量 LR 表示两种模型似然值的差异。

$$LR = 2(\ln L_1 - \ln L_2) \quad (5)$$

在这里,统计量 LR 近似服从卡方分布,自由度是两个模型的参数个数之差。其中, L_1 和 L_2 分别是两个分布的似然函数。对于微博转发网络,节点 v_i 的 k_{in} 和 k_{out} 是分析网络的基本属性,本节主要分析同一节点的 k_{in} 、 k_{out} 的统计特征和不同节点 k_{in} 、 k_{out} 之间的关系。拟合参数分别为出度为 $x_{\min}=5$, $\alpha=-2.0600$, 入度为 $x_{\min}=4.7$, $\alpha=-2.9200$ 。研究同一节点出度—入度值特征,可以看出:在双对数坐标中, k_{in} 、 k_{out} 与 $F(k)$ 不存在明显的线性关系(如图 1(a)所示),即不服从严格的幂律分布;节点的 k_{in} 、 k_{out} 的分布不同(如图 1(a)所示), k_{in} 下降速度快于 k_{out} ,即 k_{out} 具有更显著的长尾效应。图 1(b)横轴为 k_{in} 和 k_{out} 的度值,纵轴为对应的方差。可以看出,当 k_{in} 固定时, k_{out} 的方差明显大于 k_{out} 固定时 k_{in} 的方差。这是因为,出度表征了节点的客观影响力,更能体现节点间的差异。因此,在分析传播能力时,应考虑网络的出

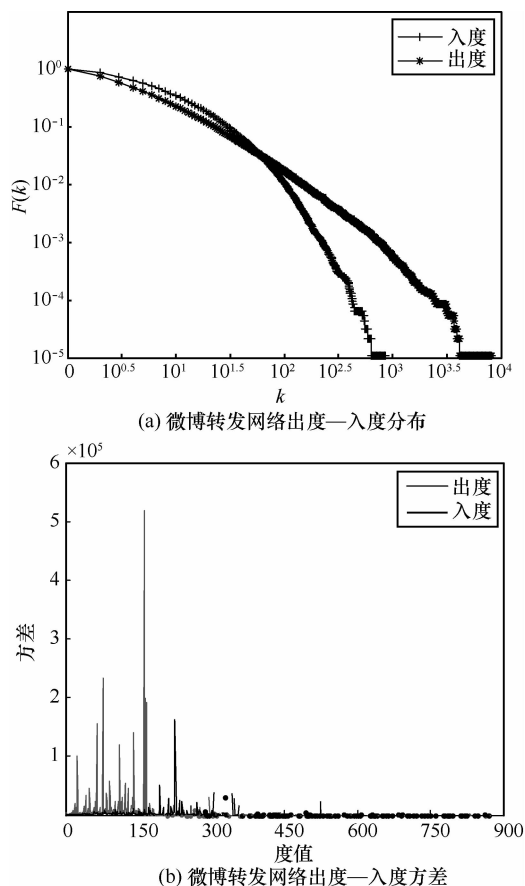


图1 微博转发网络节点度值特征

度—入度序列,保留出度—入度值特征,而不是简单地用幂律分布表示。

3 考虑度分布特性的微博信息传播

Newman 给出了无向网络渗流后巨分支规模的解,并得到渗流阈值,本文考虑有向网络情况。与无向的情况不同,有向渗流考虑的是巨出向分支而不是巨分支。假如信息是从某一个节点开始向外传播,那么最终所有获得信息的节点的集合实际上就是渗流之后的网络中初始节点的出向分支,认为最终网络中最大的出向分支规模就代表了最终接受信息的人数,但实际上只有当信息是从最大出向分支的初始节点开始传播的时候,接收信息的人数才等于最大出向分支的节点数。研究的转发网络与现实当中的转发关系相对应,如果一个用户发出的信息具有影响力,那么他在网络中出现分支的规模一定很大,那么相应地,渗流后对应的节点在最终的的网络中依然会有一个很大的出向分支,所以渗流之后的最大出向分支规模就表示了信息传播的最大可能影响范围,具有一定的

代表性。

节点是否属于某个出向分支是由其入边决定的,如果一个节点有一个入边来自于巨出向分支,那么它肯定属于巨出向分支,反之,则它所有入边的另一端点都不属于巨出向分支。假定 u_{out} 是节点不经过特定一条与其相连的入边与巨出分支相连接的平均概率,那么可以得到:

$$u_{\text{out}} = 1 - \varphi + \varphi \sum_k q_{\text{in}}(k) u_{\text{out}}^k \quad (6)$$

对于任意一个节点,一个邻接节点通过出边指向它, $q_{\text{in}}(k)$ 表示邻接节点入度为 k 的概率。利用联合分布 $p(a, b)$ 表示一个节点入度为 a 且出度为 b 的概率,那么通过出边连接的节点的邻接节点入度为 k , 出度为 i 的概率为:

$$\frac{p(k, i)}{m} = \frac{ip(k, i)}{\langle k_{\text{in}} \rangle} \quad (7)$$

那么,无论邻接节点出度取何值,其入度为 k 的概率为:

$$q_{\text{in}}(k) = \sum_i \frac{ip(k, i)}{\langle k_{\text{in}} \rangle} \quad (8)$$

可以利用图解法,将式(6)左右两边分别构造方程,分别为方程 $y = u_{\text{out}}$ 与方程:

$$y = 1 - \varphi + \varphi \sum_k q_{\text{in}}(k) u_{\text{out}}^k \quad (9)$$

由于 $q_{\text{in}}(k)$ 为条件概率,其和为 1, 所以 $u_{\text{out}} = 1$ 时为恒等式。在 $u_{\text{out}} = 1$ 处相切得到渗流阈值:

$$\varphi_c = \frac{1}{\sum_k q_{\text{in}}(k)} = \frac{\langle k_{\text{in}} \rangle}{\sum_k \sum_i ikp(k, i)} \quad (10)$$

参考文献[9]利用不同的方法,对有向网络中巨强连通分支求得了同样的渗流阈值,实际上巨出向分支、巨入向分支与巨强连通分支出现的条件是相同的。此时,出分支的规模为:

$$S_{\text{out}} = 1 - \sum_k p_{\text{in}}(k) u_{\text{out}}^k \quad (11)$$

通过上面的推导,得到了考虑节点关联性的有向渗流模型解法。接下来,将分别验证有向性和关联性对网络信息传播性能的影响以及所提出方法的正确性。

4 仿真实验

本文将研究对象限定为新浪“名人堂”用户。“名人堂”用户是新浪根据名人堂的相关规定筛选出的认证用户,是在信息传播过程中有着重要价值的真实用户。通过接口

“statuses/user_timeline”爬取微博信息,若 M_0 为原创初始微博信息,信息 M_1 转发自 M_0 ,则接口所返回的 M_1 数据中包含了“retweeted_status”字段,其中记录了包含初始发布用户在内的 M_0 的全部内容,但若此时另一条消息 M_2 转发自 M_1 , M_2 所返回的数据“retweeted_status”字段中包含的依然是 M_0 而非 M_1 的内容。因此针对转发长度大于 1 的转发消息,需要通过分析信息中的转发标志获知其直接转发信息来源。新浪微博中以前缀//@user_name(其中 user_name 表示用户名)来标识直接转发信息来源。分析 2012 年 9 月 23 日至 10 月 23 日这一个月内的名人堂用户转发信息,利用转发关系来定义边,利用账号来定义节点。其中出度是转发的数目,入度是被转发的数目。最终得到了一个由 92 933 个节点、1 083 584 条边构成的转发网络。

首先,分别利用无向转发和有向转发网络进行仿真,找出不同传播概率下子网络的最大出向分支,而这个最大出向分支的规模也就代表了最终的传播范围,通过这种方式可以粗略地模拟真实的传播情况,对比两种仿真结果,观察有向性对于信息传播的影响。同时利用 Newman 提出的无向渗流解法和本文提出的有向渗流解法预测出不同传播概率下的巨出向分支规模以及渗流阈值,与真实有向转发网络仿真进行对比,验证提出的有向渗流解法的正确性。图 2 中 4 条曲线分别表示以下内容。

- 有向仿真:在有向微博转发网络利用渗流模型进行仿真得到的结果。
- 有向渗流解法:利用前面提到的有向渗流模型解法求解得到的结果。
- 无向仿真:在无向微博转发网络利用渗流模型进行仿真得到的结果。
- 无向渗流解法:利用前面提到的无向渗流模型解法求解得到的结果。

观察图 2 可以发现,有向性对信息传播性能有着极大的影响,使得相同传播概率下信息传播范围明显变小,同时也可以发现,相比于无向渗流解法,提出的有向渗流解法与无向仿真结果较为吻合,但仍有明显差异。接下来观察渗流阈值,如图 3 所示。

图 3 中 4 条曲线与图 2 相同,观察图 3 可以发现仿真结果并没有一个十分明显的阈值,这是由于有限规模效应(finite size effect)造成的,但通过观察曲线的变化趋势,还是可以得到一个大致临界区域,大致对应阈值,边的有

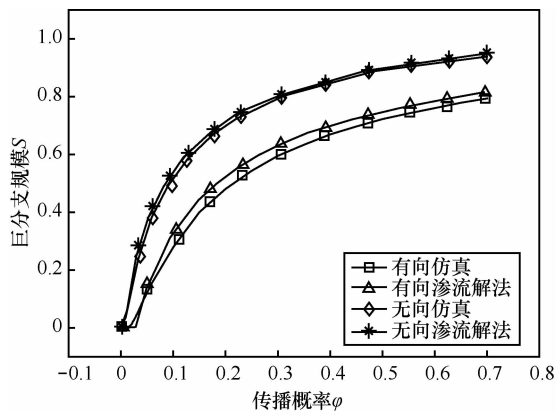


图 2 渗流有向性全局对比

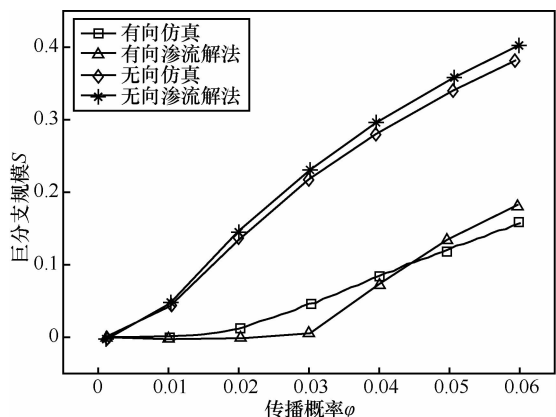


图 3 渗流有向性阈值对比

向性使得渗流阈值明显变大,即使得信息爆发变得更加困难,同时还可以发现,相比于无向渗流解法,提出的有向渗流解法给出的渗流阈值接近实际仿真阈值,但依然有偏差。

通过上面可以看出:

- 边的有向性对信息传播有着显著的影响,在有向条件下,渗流阈值增加,同样概率下传播范围变小,信息传播更为困难;
- 提出的有向渗流解法较好地贴合了有向性特征,但与有向转发网络信息传播仿真结果相比,依然存在一定差距。

5 结束语

本文通过区分信息传播的方向性,利用渗流理论研究微博转发网络中节点度值特征。区分信息传播的方向有助于更加细致地分析转发概率对传播性能的影响,有向渗流的关联更能够准确反映微博中的转发场景,为进一步分析影响渗流阈值、传播范围的因素提供了理论支撑。

参考文献:

- [1] 胡海波, 王科, 徐玲, 等. 基于复杂网络理论的在线社会网络分析[J]. 复杂系统与复杂性科学, 2008, 5(2): 1-14.
HU H B, WANG K, XU L, et al. Analysis of online social networks based on complex network theory[J]. Complex Systems and Complexity Science, 2008, 5(2): 1-14.
- [2] 张彦超, 刘云, 张海峰, 等. 基于在线社交网络的信息传播模型[J]. 物理学报, 2011, 60(5): 66-72.
ZHANG Y C, LIU Y, ZHANG H F, et al. The research of information dissemination model on online social network [J]. Acta Physica Sinica, 2011, 60(5): 66-72.
- [3] 熊熙, 胡勇. 基于社交网络的观点传播动力学研究 [J]. 物理学报, 2012, 61(15): 104-110.
XIONG X, HU Y. Research on the dynamics of opinion spread based on social network services [J]. Acta Physica Sinica, 2012, 61(15): 104-110.
- [4] LIU D C, CHEN X. Rumor propagation in online social networks like Twitter- a simulation study[C]//Proceedings of the Third International Conference on Multimedia Information Networking and Security, November 4-6, 2011, Washington DC, USA. New Jersey: IEEE Press, 2011: 278-282.
- [5] GRABOWSKI A, KOSINSKI R A. Epidemic spreading in a hierarchical social network [J]. Physical Review E, 2004, 70(3): 1027-1035.
- [6] CALLAWAY D S, NEWMAN M E, STROGATZ S H, et al. Network robustness and fragility: Percolation on random graphs[J]. Physical Review Letters, 2000, 85(25): 5468-5471.
- [7] LI M, FAN Y, CHEN J, et al. Weighted networks of scientific communication: the measurement and topological role of weight [J]. Physica A Statistical Mechanics & Its Applications, 2005, 350(2): 643-656.
- [8] CHANG H, SU B B, LIU C P, et al. Community, hierarchy and interweavement in collaboration networks [J]. International Journal of Modern Physics C, 2011, 19(10):1537-1554.
- [9] SCHWARTZ N, COHEN R, BEN-AVRAHAM D, et al. Percolation in directed scale-free networks [J]. Physical Review E, 2002, 66(1): 015104.

[作者简介]



李晋(1977-),女,北京信息科技大学讲师,主要研究方向为网络与新媒体传播。



杨子龙(1977-),男,解放军电子工程学院助理工程师,主要研究方向为复杂网络与信息传播。