物理学报 Acta Physica Sinica
Chinese Physical Society
Institute of Physics, CAS

基于有向渗流理论的关联微博转发网络信息传播研究

王小娟 宋梅 郭世泽 杨子龙

Information spreading in correlated microblog reposting network based on directed percolation theory

Wang Xiao-Juan Song Mei Guo Shi-Ze Yang Zi-Long

引用信息 Citation: Acta Physica Sinica, 64, 044502 (2015) DOI: 10.7498/aps.64.044502
在线阅读 View online: http://dx.doi.org/10.7498/aps.64.044502
当期内容 View table of contents: http://wulixb.iphy.ac.cn/CN/Y2015/V64/I4

基于有向渗流理论的关联微博转发网络信息传播研究*

王小娟[1)†] 宋梅[1)] 郭世泽[2)] 杨子龙[2)]

1) (北京邮电大学电子工程学院, 北京 100876)

2) (北方电子设备研究所, 北京 100083)

(2014 年 7 月 24 日收到; 2014 年 8 月 27 日收到修改稿)

微博网络的快速性、爆发性和时效性, 以及用户复杂的行为模式, 使得研究其信息传播模型及影响因素成为网络舆情的热点方向. 利用压缩映射定理, 分析不动点迭代过程的收敛条件, 得到有向网络信息传播过程的渗流阈值和巨出向分支的数值解法; 通过可变同配系数生成模型, 分析关联特征对信息传播的影响; 最后利用微博转发网络数据进行仿真对比实验. 结果表明: 虽然四类关联特征同时体现出同配、异配特征, 但信息传播结果更多受入度-入度相关性、入度-出度相关性影响; 通过删除少量节点的方法, 提取边同配比例, 验证大部分节点的四类关联特征呈现一致性.

关键词: 微博, 信息传播, 关联特征, 渗流阈值

PACS: 45.70.Vn, 42.65.Sf **DOI:** 10.7498/aps.64.044502

1 引 言

在线社会网络 (online social network, OSN) 是由大规模用户借助互联网形成的相对稳定的连接关系, 在一定程度上可以看作是现实关系 (如朋友、爱好、交往圈等) 在网络空间的映射. 数以百万的用户借助OSN 发布、共享和传递信息, 因此研究该平台下用户的行为[1]、信息的传播模型[2] 和影响因素分析[3] 成为网络舆情研究的热点方向.

微博 (microblog) 与电子邮件 (E-mail), 脸谱 (Facebook) 等传统 OSN 平台相比, 其短文本、快捷等特性使其信息交互更加频繁, 传播更加迅速. 粉丝用户的转发对信息在网络中的迅速蔓延的效果最为显著, 因此本文选定微博转发网络作为研究对象. 信息传播过程如下: 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, 该信息沿着多条诸如 $i \to j \to k \to \cdots$ 的路径在用户之间呈网状蔓延. 信息传播过程和有向渗流理论中的键渗流过程是存在映射关系的: 边占有概率[4] 等价于转发概率, 键渗流过程等价于信息沿着网络中的边以固定概率传播的过程 (当且仅当最终网络存在巨分支时, 信息才得以蔓延), 信息传播范围等价于巨分支大小, 传播难易程度等价于渗流阈值.

针对 OSN 传播特征的分析, 大多集中在无标度特征[5,6]、小世界特征[7]、聚类系数[8]、非拓扑信息等[9]. 然而, 在信息传播过程中, 节点的属性会直接影响其邻接节点[10,11], 节点“关联特征”主要衡量的是节点影响力的相似程度. 如在现实社会中, 人们的行为总是会有一定的倾向性, 有时人们倾向于和自己某些属性接近的交流, 例如年龄接近的人更容易成为伙伴. Newman[12] 针对无向网络研究发现: 当网络呈现同配特征 ($r > 0$) 时, 度相近的节点构成了联系紧密但规模较小的群体, 即降

* 国家自然科学基金 (批准号: 61171097, 61272491, 61309021) 资助的课题.

† 通信作者. E-mail: wj2718@163.com

低渗流阈值, 减小巨出向分支规模; 反之异配网络 ($r < 0$) 中虽然节点间连接广泛, 但结构也较为松散, 即提高渗流阈值的同时也增加了巨出向分支规模. 然而在有向网络中, 由于信息流动具有方向性, 因此同配系数就演变为四种: 入度- 入度、入度-出度、出度-出度、出度-入度四类相关性. 在OSN数据分析过程中, 我们发现并不是所有的网络在四类同配系数上都存在一致性. 其中微博转发网络就是一种典型的同异配混合网络. 不能仅仅考虑某一关联特征, 针对这种情况如何构建信息传播模型和分析四类关联特征的信息传播影响是本文的研究重点. 本文提取有向网络节点关联特征, 利用有向渗流理论分析关联特征对传播能力的影响. 主要工作如下: 构建传播模型, 利用不动点迭代的方法, 根据压缩映射原理, 给出考虑渗流关联特征的渗流阈值和巨分支大小的理论值, 并分析关联特征对传播的影响, 分析结果表明入度-入度相关性、入度-出度相关性对传播影响较为显著. 真实网络的实证分析表明: 虽然四类关联特征同时体现出同配、异配特征, 但信息传播结果更多受入度-入度相关性、入度-出度相关性影响. 进一步分析发现, 节点的出度的不均匀分布导致极少量的节点占用了大量的资源, 通过删除少量节点的方法, 提取边同配比例, 验证大部分节点的四类关联特征呈现一致性.

本文组织结构如下: 第2节介绍基于渗流理论的信息传播研究; 第3节将微博用户抽象为节点, 将转发关系抽象为有向边, 构建有向转发网络, 并分析了微博转发网络的节点关联特征在有向网络中的表达; 第4节, 利用不动点迭代的方法给出了考虑有向关联特征的渗流模型, 并分析了关联特征对渗流的影响; 第5节在真实微博网络中模拟信息传播过程, 验证了理论值的正确性, 并分析节点关联特征对传播能力的影响; 第6节总结全文.

2 基于渗流理论的信息传播研究

为简化模型, 假设信息传递概率是固定的[13], 即转发概率为常数 φ. 有向渗流理论主要用来分析部分边移除后的网络是否连通. 具体过程如下: 假设以占有概率 φ 随机占据网络中的边, 随着 φ 的增加, 相互连通的边组成的分支会逐渐变大, 最终合成一个巨分支 (giant component). 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点称为渗流阈值 φ_c(percolation threshold). 应用有向渗流理论研究信息传播, 需要解决以下两个问题: 一是在有向网络中如何评估信息传播; 二是如何分析同配系数的影响.

2.1 构建传播模型

利用渗流理论构建传播模型, 在无向网络中已有大量研究: Grabowski 和 Kosinski[14] 将渗流理论应用到真实在线社交网络; Callaway 等[15] 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; Schwartz 等[16] 在服从幂律分布的有向网络中, 讨论了出度-入度分布对渗流阈值的影响, 分析表明, 幂律分布参数 λ_{in}, λ_{out} 不同时, 网络性能有较大差异; Dorogovtsev 等[17] 将出度-入度分布的研究拓展到有向网络, 给出巨分支计算方法, 但在计算过程中并没有考虑关联特征; Newman 等[18] 在给定度分布的随机网络中, 分析巨分支大小, 通过与万维网、科学引文网等实证数据作对比发现, 仅靠度分布不能精确预测实际网络, 存在另外的指标来指示拓扑结构; Vázquez 和 Moreno[19] 在服从幂律分布的随机网络中, 进一步考虑节点关联特征对网络渗流阈值的影响; Goltsev 等[20] 将关联特征分析推广到一般无向网络.

2.2 分析特征影响

对于关联特征对信息传播的影响, 在无向网络中已有少量研究. Newman[12] 通过生成具有不同关联特征的无向网络, 对比这些网络的渗流结果, 研究表明网络同配特征降低了渗流阈值, 减小了最终巨分支规模. Goltsev 等[20] 利用分支矩阵 (branching matrix) 理论, 证明了无向网络中关联特征对于网络渗流阈值的影响, 但该方法并不适用于有向网络, 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.

综上所述, 目前利用渗流理论评估信息传播能力, 分析节点关联特征影响只在无向网络中有少量研究. 本文的主要工作是基于有向渗流理论, 分析微博网络节点关联特征对传播能力的影响, 并利用微博转发网络, E-mail 网络、Facebook 网络数据来进一步验证理论分析结果.

3 信息传播网络

为分析“关联特征”对传播能力的影响, 需要构建传播模型和提取节点关联特征(图1).

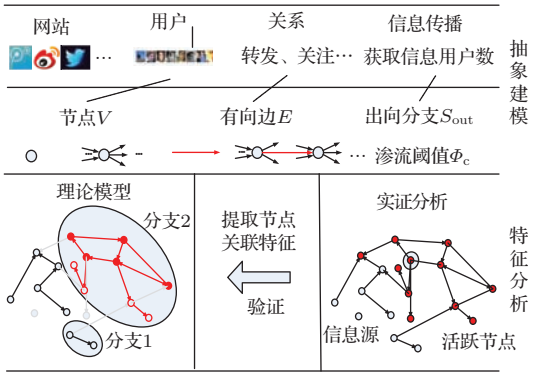

图1 (网刊彩色) 微博转发网络传播能力影响分析

3.1 传播模型构建

对用户及其行为进行抽象建模, 设为$G=(V, E)$, 其中, V是全部节点v_i的集合, 对应用户账号; E是边e_{ij}的集合, 对应用户之间的转发关系, 如果节点v_j转发了邻接节点v_i的信息, 则$e_{ij}=1$, 否则$e_{ij}=0$; 指向v_i的节点数为节点v_i的入度, 记为$k^i_{\rm in}$; v_i指向节点数为节点v_i的出度, 记为$k^i_{\rm out}$. 对于某条特定信息传播过程, 设信息源v_i发布信息, 如果对于v_j存在一条传播路径$v_i\to\cdots\to v_j$, 即v_j可以收到信息, 则称v_j为活跃节点(active node). 一般利用活跃节点数目来衡量传播能力, 活跃节点的数目越多, 对于该信息的传播能力越强, 在图论中一般利用出向分支(out component)来衡量.

定义1 出向分支$S_{\rm out}(v_i)$, 是指那些从节点v_i出发沿有向路径能够到达的所有节点的集合.

对于网络传播能力的衡量应考虑整体传播的平均情况. 在信息传播过程中, 每条可能的传播路径将以转发概率φ被保留, 某个节点是活跃节点的充分条件是与保留的边相连接. 如图1所示, 当信息源节点属于某强连通分支(红色实心节点)时, 活跃节点的集合可以用沿着保留边(实线边)到达的出向分支表示. 因此, 选择最大的出向分支, 即巨出向分支(out-giant component)作为传播能力的指标, 定义如下.

定义2 巨出向分支$S_{\rm out}$: 当节点数目$n=||V||\to\infty$时, 如果网络的最大分支$S_{\max}$的节点数和n成比例, 也就是说$||S_{\max}||\propto n$, 则称最大分支$S_{\max}$为网络巨出向分支. 具体定义如下:

$$S_{\rm out}=\{v_j\in S_{cc},\ \text{iff}\ \exists\ v_i\in S_{cc}\wedge\ e_{ij}=1\},\quad(1)$$

其中S_{cc}是巨强连通分支, 定义如下:

$$S_{cc}=\{\forall v_i, v_j\in S_{cc},\ \text{iff}\ e_{ij}=1\wedge\ e_{ji}=1\}.\quad(2)$$

可以发现, $S_{\rm out}$越大, 信息传播范围越广. 随着φ的增大, $S_{\rm out}$也会增大, 但$S_{\rm out}$并不是关于φ严格单调递增的, 而是在某个特定φ值时发生渗流相变, 所以其临界值$\varphi_{\rm c}$也是网络信息传播能力的度量. $\varphi_{\rm c}$越小, 信息传播越容易. 综上所述, 本文将从两个方面衡量网络的传播能力: 巨出向分支指标表征传播范围, 渗流阈值指标表征传播难易程度.

3.2 节点关联特征

用户度值属性上的相似, 可以用一条边两端的度值(出度/入度)的相关系数来计算[12,21,22]:

$$r_{\alpha,\beta}=\frac{E[k^i_\alpha k^j_\beta]-\mu(k^i_\alpha)\mu(k^j_\beta)}{\sigma(k^i_\alpha)\sigma(k^j_\beta)},\quad(3)$$

其中, α, $\beta\in\{\text{in, out}\}$代表的是网络度值类型; $\mu(\cdot)$与$\sigma(\cdot)$分别表示期望与方差. r取值范围为$[-1, 1]$, 相关系数代表节点关联特征: 当$r=1$时, 网络表现出完全同配特征, 即所有的节点连接到度值相同的节点, 对应规则网络; 当$r=0$时, 网络是随机的, 即节点间没有关联特征, 则对应随机网络; 当$r=-1$时, 网络表现出完全异配特征, 即所有的节点连接到完全不同度值的节点, 则对应星型网络.

以入度-入度相关性为例, 令$\boldsymbol{K}_{\rm in}$, $\boldsymbol{K}_{\rm out}$分别表示节点的入度、出度向量, 由于

$$\mu(k^i_{\rm in})=\boldsymbol{K}^{\rm T}_{\rm in}\cdot\boldsymbol{K}_{\rm out}/L,\quad(4)$$

$$\mu(k^j_{\rm in})=\boldsymbol{K}^{\rm T}_{\rm in}\cdot\boldsymbol{K}_{\rm in}/L,\quad(5)$$

$$\sigma(k^i_{\rm in})=\sqrt{\boldsymbol{K}_{\rm out}\cdot(\boldsymbol{K}_{\rm in})^2-(\boldsymbol{K}_{\rm in}\cdot\boldsymbol{K}_{\rm in}/L)^2},\quad(6)$$

$$\sigma(k^i_{\rm in})=\sqrt{(\boldsymbol{K}_{\rm in})^3-(\boldsymbol{K}_{\rm in}\cdot\boldsymbol{K}_{\rm in}/L)^2}.\quad(7)$$

由于$\mu(k^i_{\rm in})$, $\mu(k^j_{\rm in})$和$\sigma(k^i_{\rm in})$, $\sigma(k^j_{\rm in})$只与网络度分布相关, 即在给定度序列情况下为固定值, 则关联特征的值主要取决于

$$E\left[k^i_{\rm in}k^j_{\rm in}\right]=\boldsymbol{K}^{\rm T}_{\rm in}\cdot\boldsymbol{E}\cdot\boldsymbol{K}_{\rm in}.\quad(8)$$

在有向网络中, 由于信息传递具有方向性, 出度代表信息流出的方向, 表征了节点的影响程度;

入度代表信息流入的方向, 表征节点的活跃程度, 即入度- 入度、入度-出度、出度-出度、出度-入度四类相关性分别代表不同意义. 如何揭示关联特征和渗流阈值、巨出向分支的关系成为下节的研究重点.

4 节点关联特征分析

4.1 考虑有向关联特征的渗流模型

在有向网络中, 巨分支的大小不仅和转发概率 φ 相关, 还和以下两类联合概率分布相关:

1) $p(k^i_{\rm in}, k^i_{\rm out})$ 为随机选择一个节点 v_i, 其入度为 $k^i_{\rm in}$, 出度为 $k^i_{\rm out}$ 的概率;

2) $q(k^i_{\rm in}, k^j_{\rm in})$ 为随机选择一条有向边 e_{ij}, 初始节点 v_i 入度为 $k^i_{\rm in}$, 末端节点 v_j 入度为 $k^j_{\rm in}$ 的概率.

对于一个已形成的巨出向分支中的节点 v_i, $p(k^i_{\rm in}, k^i_{\rm out})$ 反映了节点 v_i 自身出入度相关性; $q(k^i_{\rm in}, k^j_{\rm in})$ 反映了一条边两端节点的关联性, 而转发概率 φ 则是任意一条边的保留概率. 根据邻接矩阵可以计算得到 $p(k^i_{\rm in}, k^i_{\rm out})$ 和 $q(k^i_{\rm in}, k^j_{\rm in})$. 令 $u(k^j_{\rm in})$ 为随机选择一条边 e_{ij}, 末端节点 v_j 不经过 e_{ij} 连到巨出向分支 $S_{\rm out}$ 的概率. $u(k^j_{\rm in})$ 的计算可以分为两种情况: 情况 1 为 e_{ij} 被删除, 概率为 $1-\varphi$; 情况 2 为 e_{ij} 被保留, 但初始节点 v_i 不属于 $S_{\rm out}$, 概率为

$$\varphi \sum\nolimits_{k^i_{\rm in}} q\left(k^i_{\rm in}|k^j_{\rm in}\right) u(k^i_{\rm in})^{k^i_{\rm in}},$$

其中

$$\begin{aligned} & q\left(k^i_{\rm in}|k^j_{\rm in}\right) \\ =& \frac{q(k^i_{\rm in}, k^j_{\rm in})}{\sum_{k^i_{\rm in}} q(k^i_{\rm in}, k^j_{\rm in})} \\ =& \frac{\sum_{k^i_{\rm out}} k^i_{\rm out} q(k^i_{\rm in}, k^i_{\rm out}, k^j_{\rm in})}{\sum_{k^i_{\rm in}} \sum_{k^i_{\rm out}} k^i_{\rm out} q(k^i_{\rm in}, k^i_{\rm out}, k^j_{\rm in})}, \end{aligned} \tag{9}$$

表示末端节点 v_j 入度为 $k^j_{\rm in}$ 时, 随机选定一条边 e_{ij}, 初始节点 v_i 入度为 $k^i_{\rm in}$ 的条件概率. 可以得到

$$\begin{aligned} & u(k^j_{\rm in}) \\ =& 1-\varphi+\varphi \sum_{k^i_{\rm in}=0} q\left(k^i_{\rm in}|k^j_{\rm in}\right) u(k^i_{\rm in})^{k^i_{\rm in}}. \end{aligned} \tag{10}$$

由于无法得到方程组 (10) 的解析解, 令

$$\begin{aligned} \boldsymbol{U} &= \left(u\left(1\right), u\left(2\right), u\left(3\right), \cdots\right)^{\rm T}, \\ \boldsymbol{F} &= \left(f_1\left(\boldsymbol{U}\right), f_2\left(\boldsymbol{U}\right), f_3\left(\boldsymbol{U}\right), \cdots\right)^{\rm T}, \end{aligned}$$

适当选取初始向量 $\boldsymbol{U}^0$, 构成迭代公式

$$\boldsymbol{U}^{(k+1)} = \boldsymbol{F}(\boldsymbol{U}^{(k)}) \ \ (k=1,2,3,\cdots),$$

若存在 $\boldsymbol{U}*$ 满足 $\boldsymbol{U}* = \boldsymbol{F}(\boldsymbol{U}*)$, 则 $\boldsymbol{U}*$ 为方程组 (10) 的数值解, 该方法称为不动点迭代法.

下面证明不动点迭代法在区间上的收敛性. 任取 $\boldsymbol{U}_1$, $\boldsymbol{U}_2 \in [0, 1)$, 有

$$\boldsymbol{F}\left(\boldsymbol{U}_1\right) - \boldsymbol{F}\left(\boldsymbol{U}_2\right) = \varphi \boldsymbol{B} \cdot \left(\boldsymbol{U}_1 - \boldsymbol{U}_2\right),$$

其中矩阵 $\boldsymbol{B}$ 的第 i, j 项

$$\begin{aligned} b_{ij} &= k^j_{\rm in} q\left(k^j_{\rm in}|k^i_{\rm in}\right) \\ &= k^j_{\rm in} \sum_{k^j_{\rm out}} k^j_{\rm out} q\left(k^j_{\rm in}, k^j_{\rm out}|k^i_{\rm in}\right), \end{aligned}$$

同时可以得到

$$\begin{aligned} & \varphi \left\| \boldsymbol{B} \cdot \left(\boldsymbol{U}_1 - \boldsymbol{U}_2\right) \right\| \\ =& \varphi \sqrt{\left(\boldsymbol{U}_1 - \boldsymbol{U}_2\right)^{\rm T} \cdot \boldsymbol{B}^{\rm T} \cdot \boldsymbol{B} \cdot \left(\boldsymbol{U}_1 - \boldsymbol{U}_2\right)} \\ \leqslant& \varphi \lambda_{\max} \left\| \left(\boldsymbol{U}_1 - \boldsymbol{U}_2\right) \right\|, \end{aligned} \tag{11}$$

其中 $\lambda_{\max}$ 为矩阵 $\boldsymbol{B}$ 的主特征值, 而当 $\boldsymbol{U}_1$, $\boldsymbol{U}_2$ 中所有项趋近于 1 时, 矩阵 $\boldsymbol{B}$ 的主特征值取得最大值 $\lambda'_{\max}$, 此时矩阵 $\boldsymbol{B}*$ 的元素 $b'_{ij} = k^j_{\rm in} q\left(k^j_{\rm in}|k^i_{\rm in}\right)$, 当 $\varphi \leqslant 1/\lambda'_{\max}$ 时存在 $L \in (0,1)$ 使得

$$\varphi \left\| \boldsymbol{B} \cdot \left(\boldsymbol{U}_1 - \boldsymbol{U}_2\right) \right\| < L \left\| \left(\boldsymbol{U}_1 - \boldsymbol{U}_2\right) \right\|.$$

根据压缩映射原理可知, $\boldsymbol{U}*$ 解存在, 即收敛且渗流阈值为

$$\varphi_{\rm c} = 1/\lambda'_{\max}. \tag{12}$$

利用 $\boldsymbol{U}$ 可以得到巨出向分支的大小:

$$S_{\rm out} = 1 - \sum_{k^i_{\rm in}} \sum_{k^i_{\rm out}} p\left(k^i_{\rm in}, k^i_{\rm out}\right) u(k^i_{\rm in})^{k^i_{\rm in}}. \tag{13}$$

4.2 有向关联特征对渗流的影响

通过前面的分析可以发现, 渗流得到的巨分支同时由节点自身的出入度分布以及边两端节点间的关联性决定. 节点自身的度分布与用户自身属性相关, 而节点间关联特征则由用户的行为决定. 对于无向网络, Grabowski 和 Kosinski [14] 证明, 如果节点之间关于度同配, 那么其渗流阈值会减小, 即网络的鲁棒性增强, 反之亦然. 对于有向网络, 首先在不考虑节点关联特征时, 对于矩阵 $\boldsymbol{B}*$ 有

$b'_{ij}=k_j^{\text{in}}q\left(k_j^{\text{in}}\right)$. 此时矩阵 $\boldsymbol{B}*$ 只有一个特征值, 由于

$$\sum_{k_{\text{in}}^j}k_{\text{in}}^j q\left(k_{\text{in}}^j\right)=\sum_{k_{\text{in}}^j,k_{\text{out}}^j}k_{\text{in}}^j k_{\text{out}}^j p\left(k_{\text{in}}^j,k_{\text{out}}^j\right),$$

对于矩阵 $\boldsymbol{B}*$ 的最大特征值 $\lambda_{\max}$, 其上限有

$$\begin{aligned}\lambda_{\max}&\leqslant\sum\nolimits_i b'_{ij}\\&=\max_j\sum_{k_{\text{in}}^i,k_{\text{out}}^i}k_{\text{in}}^i k_{\text{out}}^i q\left(k_{\text{in}}^i,k_{\text{out}}^i|k_{\text{in}}^j\right),\end{aligned}\tag{14}$$

对于任意向量 $\boldsymbol{X}$, 可以得到其下限

$$\lambda_{\max}\geqslant\sqrt{\boldsymbol{X}^{\text{T}}\boldsymbol{B}^{\text{T}}\boldsymbol{B}\boldsymbol{X}/(\boldsymbol{X}^{\text{T}}\boldsymbol{X})}.\tag{15}$$

假设边两端节点的入度完全同配, 即

$$\begin{cases}b'_{ii}=i_{\text{in}},\\b'_{ij}=0\quad(i\neq j),\end{cases}$$

那么, 矩阵 $\boldsymbol{B}$ 为对角阵, 其对角元素均为矩阵的特征值, 那么其最大特征值 $\lambda'_{\max}$ 为 $\max k_{\text{in}}^i$, 显然

$$\max i_{\text{in}}\geqslant\max\sum_{i_{\text{in}}}k_{\text{in}}^i q\left(k_{\text{in}}^i\right),\tag{16}$$

那么说明按照入度完全同配时, 网络的渗流阈值变小. 对于一般情况, 观测方程组 (10) 可以发现, 幂小的 $u(k_{\text{in}}^j)$ 对应较大的值, 幂大的 $u(k_{\text{in}}^j)$ 对应了较小的值, 使得

$$\sum_{k_{\text{in}}^i}\sum_{k_{\text{out}}^i}p\left(k_{\text{in}}^i,k_{\text{out}}^i\right)u(k_{\text{in}}^i)^{k_{\text{in}}^i}$$

变小, 因此, 相较于无关联情况, 边两端节点间入度同配使得 S_{out} 值变小. 可以发现, 入度同配的网络通过不同度值节点连接的倾向性, 使得网络的渗流阈值降低, 而同样保留概率下的巨分支规模变小. 也就是说在迭代过程中, 入度-出度相关性、入度-入度相关性对渗流阈值的影响较为明显.

下面用数值分析的方法验证该分析结果. 从 4.1 节的理论分析结果可以看出, 渗流阈值和巨分支都与矩阵 $\boldsymbol{B}$ 的最大特征值相关, 也就是说我们的目的是找到同配系数和最大特征值的关系. 从 (8) 式可以看出, 同配系数和一条边两端的度值的乘积相关, 如图 2 所示, 对于 $v_1\to u_1$ 和 $v_2\to u_2$, 如果互换之后度值的乘积比原来的值大, 则边互换可以提高网络的同配系数; 否则不互换. 利用这种贪婪算法思想, 可以控制入度-入度、入度-出度、出度-出度、出度-入度四类不同的参数.

如图 3 所示, 通过控制入度-入度相关性, 可以发现随着迭代次数的增加, 网络的入度-入度相关性逐渐增加到 1, 而其他三类相关性变化不大. 通过生成不同同配系数的网络来研究同配系数对渗流阈值的影响. 从图 4 可以看出: 入度-入度相关性对网络信息传播具有显著影响, 入度-出度相关性影响较弱, 出度-出度相关性、出度-入度相关性几乎没有影响.

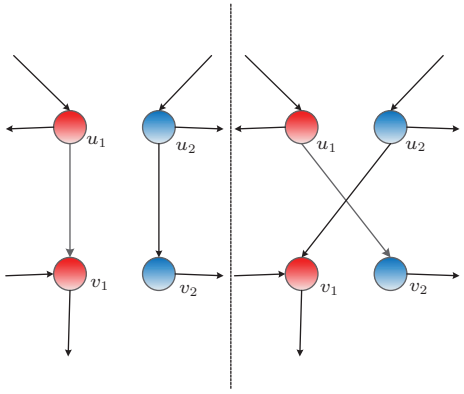

图 2 可变同配系数的生成模型

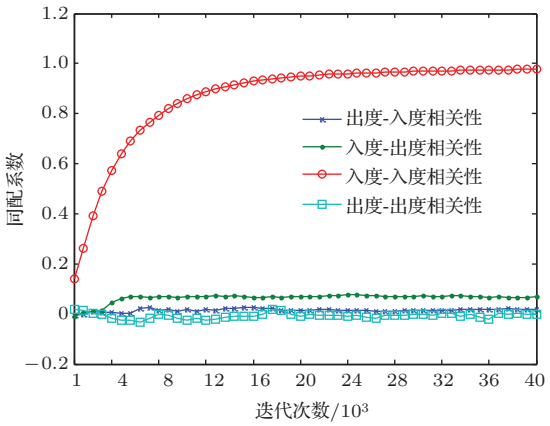

图 3 (网刊彩色) 基于贪婪算法的可变同配系数网络生成算法, 其中节点数为 500, 边数为 1500

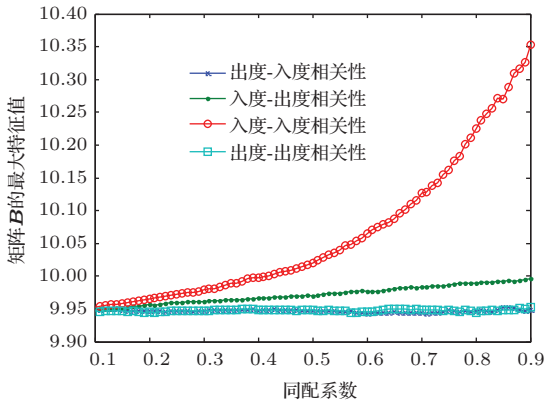

图 4 (网刊彩色) 同配系数对渗流阈值的影响

5 实证分析

5.1 不同类型信息同配特征描述

本节选定三个数据库提取节点关联特征, 并计算相关系数 $\boldsymbol{R} = [r_{\text{in,in}}, r_{\text{in,out}}, r_{\text{out,in}}, r_{\text{out,out}}]$, 其中数据信息如下.

1) E-mail: 数据来源于欧洲某研究机构, 节点为发送/接收邮件的用户, 边是发送关系, 由发送邮件用户指向接收邮件用户, 包含265214个节点和420045 条边, 采集时间为2003.10—2005.3. $\boldsymbol{R} = [-0.1829, -0.2104, -0.1459, -0.1625]$, 是典型异配网络, 四类相关系数 $r_{\alpha,\beta}$ 都小于0, 这是由于E-mail作为工作交流工具, 更多的是由上级到下级的传达或是由下级到上级的汇报, 而同一级别的人往往交流很少.

2) Facebook: 数据来源于WOSN 2009发布的OSN数据集, 节点为New Orleans所有的Facebook用户, 边是留言信息, 由留言用户指向被留言用户. 包含63891个节点和876993条边. $\boldsymbol{R} = [0.4548, 0.4596, 0.4418, 0.4069]$, 是典型同配网络, 四类相关系数 $r_{\alpha,\beta}$ 都大于0, 这是由于Facebook作为日常生活的社交平台, 聚集在一起的用户往往具有相似背景、兴趣、爱好.

3) 微博: 数据来源于新浪微博“名人堂”, 节点为微博账号, 边是转发关系, 由被转发用户指向转发用户. 包含92933个节点和1083584条边. 采集时间为2012.09.23—2012.10.23. $\boldsymbol{R} = [0.1254, 0.0556, -0.0305, -0.0084]$, $r_{\text{in,in}}$, $r_{\text{in,out}}$ 都大于0, 表现出同配特征; 而 $r_{\text{out,in}}$, $r_{\text{out,out}}$ 都小于0, 表现出异配特征. 也就是说, 微博网络在节点关联特征上, 与典型同配(或异配)网络不同, 四个相关系数并不一致, 即同时具有同配、异配特征.

5.2 不同类型信息传播网络实证分析

本节在同一节点出度-入度不变的前提下, 将边随机重连, 使得节点间的度值没有关联. 重连规则如下: 将每条边 $v_i \to v_j$ 分为两个“短截线”, 出边截线“$v_i \to$”和入边截线“$\to v_j$”, 将所有的“短截线”进行随机重连, 因为只有出边截线与入边截线相匹配才能重连, 因此, 对比网络的节点出度-入度与原网络相比保持不变, 但节点关联特征已被随机化. 通过对比分析节点关联特征影响, 可以看出:

1) 如图5 (a) 所示, E-mail网络是典型的异配网络, 真实网络渗流阈值为0.0202, 对比网络渗流阈值为0.0083, 即相比于随机连接的情况, 渗流阈值较大, 传播更为困难, 巨分支规模也较大, 传播范围却增加, 也就是说, 在有向网络中, 完全异配情况下, Newman的结论依然成立;

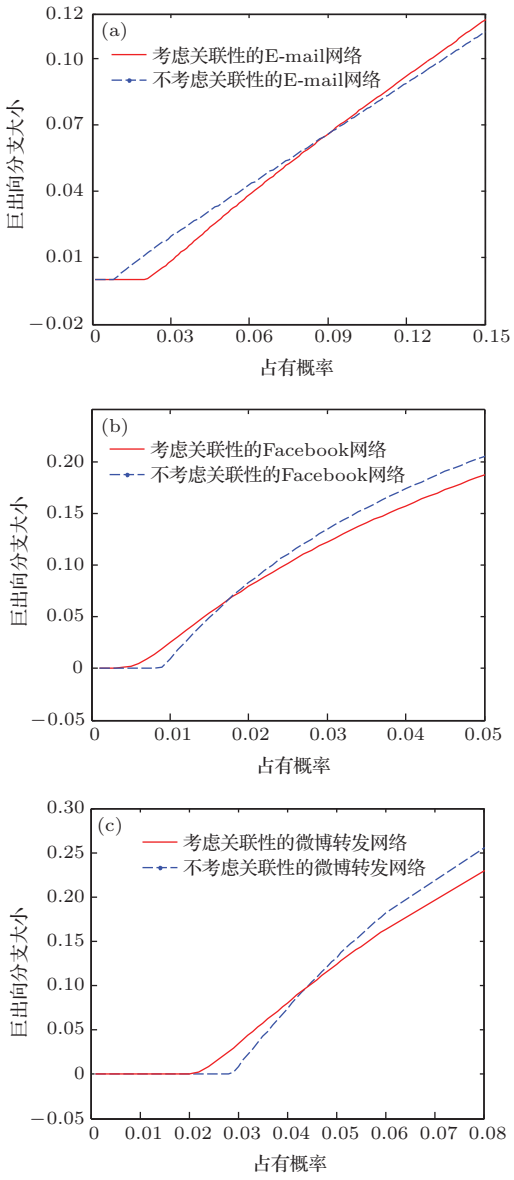

图5 (网刊彩色) 节点关联特征影响分析 (a) E-mail网络; (b) Facebook网络; (c) 微博转发网络

2) 如图5 (b) 所示, Facebook网络是典型同配网络, 真实网络渗流阈值为0.0012, 对比网络渗流阈值为0.009, 相比于随机连接的情况, 渗流阈值较小, 传播更为容易, 巨分支规模也较小, 但传播范围却减少, 也就是说, 在有向网络中, 完全同配情况下, Newman的结论依然成立;

3) 如图5 (c) 所示, 微博转发网络真实网络渗流阈值为0.0222, 对比网络渗流阈值为0.0288, 节点间关联的网络 (实线) 与节点间非关联的网络 (虚线) 相比 φ_c 变小, 即更容易形成 S_{out}, 随着 φ 的增加, 在同样的传播概率下, S_{out} 的规模更小.

综上所述, 真实的微博转发网络在传播能力上更多地表现出了同配特征. 这与4.2节的结论相符, $\boldsymbol{R} = [0.1254, 0.0556, -0.0305, -0.0084]$, $r_{\text{in,in}}$, $r_{\text{in,out}}$ 都大于0, 表现出同配特征, 与渗流结果一致.

5.3 节点关联特征实证分析

在微博转发网络中, 节点出度变化范围广泛, 分布不均匀, 意味着微博用户从影响力上来说, 极少量的用户占据了大量的"资源". 这些"核心人物"表现出较明显的"异配特征", 因为该节点的度值极大, 数量极少, 很难找到性质相同的节点. 为了消除这些极少节点带来的影响, 本文对节点间的关联特征做了更细致的度量 (表1):

第一步 计算出度小于1000的用户的关联特征, 即保留99.94%以上的节点以及90.95%以上的边, 可以发现, 节点间的四类相关系数分别为0.1254, 0.0956, 0.0260和0.0105, 即表现出一致的同配特征;

第二步 计算出度小于200的用户的关联特征, 由于出度分布的不均匀, 只有68.55%的边被保留, 但却有99.32%以上的节点被保留下来, 可以发现, 节点间的四类相关系数分别为0.1254, 0.1273, 0.0859和0.0538, 即相关性变大;

第三步 将初始节点的入度 (出度) 固定, 将终止节点的入度 (出度) 取均值, 分析相邻节点的趋势是否一致, 可以发现, 节点间的四类相关系数分别为0.7098, 0.5243, 0.4458和0.4329, 即大多数节点同配特征显著.

表1 不同节点度值的Pearson相关系数

类型	$k_{\text{out}} > 1000$		$k_{\text{out}} > 200$		相关系数 (取均值)
	相关系数	节点比例/%	相关系数	节点比例/%	
入度-入度	0.1254	100	0.1254	100	0.7098
入度-出度	0.0956	99.94	0.1273	99.32	0.5243
出度-入度	0.0260	99.94	0.0859	99.32	0.4458
出度-出度	0.0105	99.94	0.0538	99.32	0.4329

综上所述, 在真实微博转发网络中, 从出度-出度相关性、出度-入度相关性来看, 表现出异配特征, 但如果剔除极少量的"核心人物"的影响, 其他大部分节点的同配特征显著, 所以在网络的信息传播过程中, 四类相关系数从本质上来说还是一致的.

6 结 论

本文利用渗流理论研究了网络中节点间关联特征对于网络信息传播能力的影响. 一方面对于网络的度-度相关性进行了直观分析, 另一方面利用传播理论模型分析相关性的影响. 在影响分析过程中, 发现虽然节点四类关联特征同时体现出同配、异配特征, 但信息传播结果受入度-入度、入度-出度相关性影响较大, 所以渗流过程更接近同配网络, 且按照删除部分节点、边的方法分析了这一现象的原因. 更深入的分析, 将作为下一步工作的重点.

参考文献

[1] Centola D 2010 *Science* **329** 1194

[2] Zhang Y C, Liu Y, Zhang H F, Cheng H, Xiong F 2011 *Acta Phys. Sin.* **60** 050501 (in Chinese) [张彦超, 刘云, 张海峰, 程辉, 熊菲 2011 物理学报 **60** 050501]

[3] Centola D 2011 *Science* **334** 1269

[4] Miller J C 2007 *Phys. Rev. E* **76** 010101

[5] Java A, Song X, Finin T, Tseng B 2007 *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis* San Jose, USA, August 12–15, 2007 p56

[6] Kwak H, Lee C, Park H, Moon S 2010 *Proceedings of the 19th International Conference on World Wide Web* Raleigh USA, April 26–30, 2010 p591

[7] Backstrom L, Boldi P, Rosa M, Ugande J 2012 *Proceedings of the 3rd Annual ACM Web Science Conference* Evanston, USA, June 22–24, 2012 p33

[8] Xiong F, Liu Y, Si X M, Ding F 2010 *Acta Phys. Sin.* **59** 6889 (in Chinese) [熊菲, 刘云, 司夏萌, 丁飞 2010 物理学报 **59** 6889]

[9] Zou S R, Peng Y J, Liu A F, Xu X L, He D R 2011 *Chin. Phys. B* **20** 018902

[10] Watts D J, Dodds P S 2007 *J. Consum. Res.* **34** 441

[11] Crandall D, Cosley D, Huttenlocher D, Kleinberg J, Suri S 2008 *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* Las Vegas, USA, August 24–26, 2008 p160

[12] Newman M E J 2002 *Phys. Rev. Lett.* **89** 208701

[13] Kenah E, Robins J M 2007 *Phys. Rev. E* **76** 036113

[14] Grabowski A, Kosinski R A 2010 *Acta Phys. Pol. B* **41** 1135

[15] Callaway D S, Newman M E J, Strogatz S H, Watts D J 2000 *Phys. Rev. Lett.* **85** 5468

[16] Schwartz N, Cohen R, Ben-Avraham D, Barabási A L 2002 *Phys. Rev. E* **66** 015104

[17] Dorogovtsev S N, Mendes J F F, Samukhin A N 2001 *Phys. Rev. E* **64** 025101

[18] Newman M E J, Strogatz S H, Watts D J 2001 *Phys. Rev. E* **64** 026118

[19] Vázquez A, Moreno Y 2003 *Phys. Rev. E* **67** 015101

[20] Goltsev A V, Dorogovtsev S N, Mendes J F F 2008 *Phys. Rev. E* **78** 051105

[21] Foster J G, Foster D V, Grassberger P, Paczuski M 2010 *Proc. Nat. Acad. Sci.* **107** 10815

[22] Piraveenan M, Prokopenko M, Zomaya A 2012 *IEEEACM Trans. Computat. Biol. Bioinform.* **9** 66

Information spreading in correlated microblog reposting network based on directed percolation theory*

Wang Xiao-Juan[1)†] Song Mei[1)] Guo Shi-Ze[2)] Yang Zi-Long[2)]

1) (*Department of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China*)

2) (*The Institute of North Electronic Equipment, Beijing 100083, China*)

(Received 24 July 2014; revised manuscript received 27 August 2014)

Abstract

Due to the properties of rapidity, explosive, timeliness and complicated behavior for user, the research on information spreading progress and influence factors for microblog becomes a hot area of network public opinion. In this paper, firstly we use the contracting mapping principle to discuss the convergence conditions of the iterative algorithm. The numerical solution of the percolation threshold and the size of the largest out-component are proposed. Then the influence of assortativity is analyzed based on the generation model with varying parameter. The feasibility of the proposed algorithm is verified by collecting microblog reposting data. Experimental results demonstrate that four correlation characteristics are shown to have assortativity and disassortativity, but the results of message spreading are closer to that of the assortative network which is related to in-in and in-out degree correlation. It can be verified that the four types of correlation characteristics of a large part of nodes show their consistency for assortativity, through deleting a few nodes as well as extracting link scale for four degree correlations.

Keywords: microblog, information spreading, correlation characteristic, percolation threshold

PACS: 45.70.Vn, 42.65.Sf **DOI:** 10.7498/aps.64.044502

* Project supported by the National Natural Science Foundation of China (Grant Nos. 61171097, 61272491, 61309021).

† Corresponding author. E-mail: wj2718@163.com