doi:10.3969/j.issn.1000-436x.2014.12.015

# 基于有向拓扑势的用户角色分析方法

段松青[1,2]，于兴隆[2]，吴斌[2]，王柏[2]
（1. 中国软件评测中心 云计算促进中心，北京 100048；2. 北京邮电大学 计算机学院，北京 100876）

**摘 要**：真实世界中存在大量有向、加权、动态的网络。针对有向加权网络的节点角色分析问题，提出了一种基于有向拓扑势的节点角色分析方法，该方法根据节点的行为模式及局部影响力将节点划分成 4 种角色。然后介绍了基于节点角色的动态网络演化分析方法，它能对角色行为进行动力学建模，展示了随时间连接模式的变化，并能检测较大影响的事件。实验结果表明，本方法能有效估计节点角色并检测动态网络的演化。

**关键词**：社会化网络；有向拓扑势；角色分析；动态网络演化
**中图分类号**：TP399 **文献标识码**：A **文章编号**：1000-436X(2014)12-0124-12

# User role analysis method based on directed topological potential

DUAN Song-qing[1,2], YU Xing-long[2], WU Bin[2], WANG Bai[2]
(1. Cloud Testing Center, China Softuare Testing Center, Beijing 100048, China;
2. School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China)

**Abstract:** The majority of real-world networks are directed, weighted and dynamic. Aiming at the problem of node role analysis in directed weighted network, a novel node role analysis method based on directed topological potential is proposed, which can divide nodes into four roles based on their behavior pattern and local influence. Then, a node role-based dynamic networks evolution analysis method is introduced, which can model the dynamics of behavioral roles representing the main connectivity patterns over time and detect the significant event. The experiment results indicate that proposed approaches can effectively estimate the node role and detect the dynamics of network evolution.

**Key words:** social network; directed topological potential; role analysis; dynamics of network evolution

## 1 引言

社会化网络由社会活动者和他们之间的关联组成[1]。实际研究中，学者常以节点代表活动者，以边代表活动者之间的关系。节点的角色可用于描述某个节点与其邻居，甚至与整个网络中所有节点的行为关系。角色分析着重于研究行动者之间的关系或找出行动者的集合[1]。识别用户角色具有重要的应用价值，例如，分析通信网络中客户角色有利于运营商定制和推广套餐业务[2]，而对恐怖组织成员进行角色分析有助于政府制定精确打击策略[3]。近些年，关于节点角色分析的研究较少，但 Scott 仍主张角色分析是社会网络分析的核心元素[4]。

在以往的节点分析研究中，Lorrain 和 White 认为具有同样角色的节点其所在的网络结构应该相同，即具有共同的邻居[5]。这种严格的定义随后又有各种形式的弱化，但仍难以分析实际场景中节点的角色。Oger 等人[6]指出，诸如互联网、新陈代谢网、航空运输网和蛋白质相互作用网等网络，不仅功能不同，而且不同角色节点之间的关联模式也不一样。

节点角色分析与重要性分析具有较强关联。社会化网络具有小世界、无标度、社区结构等特点，

**收稿日期**：2014-07-31；**修回日期**：2014-09-15
**基金项目**：国家重点基础研究发展计划（“973”计划）基金资助项目（2013CB329603）；国家自然科学基金资助项目（71231002,61375058）；北京市教育委员会共建项目专项基金资助项目；教育部-中国移动科研基金资助项目（MCM20123021）
**Foundation Items:** The National Basic Research Program of China (973 Program)(2013CB329603); The National Natural Science Foundation of China (71231002, 61375058); The Program of the Co-construction with Beijing Municipal Commission of Education of China; Ministry of Education-China Mobile Research Foundation (MCM20123021)

有效评估网络节点的重要性是网络化数据挖掘中的一个基本问题，也是复杂网络、系统科学、社会网络分析等领域中一个值得研究的方向。在万维网中，评价网页的重要性可以帮助用户找到更相关的网页；在医学神经网络中，发现重要的神经元对医学和生理研究有着重要的意义。如今已有大量从不同角度计算网络中个体节点重要性的方法，例如认为“重要性等价于显著性”的社会网络中心性分析方法，认为“破坏性等价于重要性”的系统科学分析方法，以及 PageRank、HITS 等网络链接分析方法。

在节点重要性排名的研究中，Hu 等人发现全局排名是一种由单一方式得到的弱偏序，而通过行为和局部作用所获得的所有节点的严格偏序可能更合理[7]。拓扑势考虑了节点的局部影响，并被应用于节点重要性排序、网络结构特征描述、社团发现及社团成员识别等领域[8~14]。文献[9]提出了一种度量无向网络中节点的重要程度的方法，其核心思想是“与重要节点相邻的节点可能也重要”，能较精细地反映无向网络拓扑结构中节点之间相互影响而产生重要性的差异。然而，在很多情况下，节点间关系或交互具有方向性和权重差异。例如，通话网络中，在一段时间内，呼叫记录由呼叫方、被叫方、累计呼叫次数等组成，故这种通话关系是有向加权的。如果只考虑网络中节点是否存在影响而忽略关系的指向和重要程度，势必失去大量可能有价值的信息。

为了解决有向加权网络中的节点角色分析问题，本文提出了一种新的分析方法，其贡献如下：将拓扑势的概念延伸到有向加权网络，形成 2 种新的节点重要性度量指标——入度拓扑势和出度拓扑势，所提基于影响范围的有向拓扑势算法具有较低的计算复杂度；基于二维有向拓扑势，提出一种新的节点角色分析算法；根据节点角色的变化，研究动态网络演化过程。

通过对社会网络实例的分析，本文算法能对节点的重要性和角色划分做出有效的评价。

## 2　基础知识

节点的拓扑势是基于认知物理学的数据场思想提出来[15]。根据有源场的思想，每个节点以自己为中心产生一个“拓扑场”，对场内所有节点（包括自身）产生一定的“拓扑势”；拓扑势的大小根据节点质量、节点间距离而定。故某节点拓扑势反映的是该节点受自身和近邻节点共同影响的程度。涂文燕等[10]基于数据场的思想，针对无向网络根据拓扑势对节点重要性进行衡量，较全面地阐述了无向拓扑势、无向拓扑势熵的定义，并分析影响因子与拓扑势熵的关联。

### 2.1　无向拓扑势

设网络拓扑为 $G=(V,E)$，$V$ 是节点的集合，$E$ 是边集合，节点 $i$ 处的拓扑势如式(1)所示。

$$\varphi(v_i)=\sum_{j=1}^{N}(m_j \mathrm{e}^{-(d_{ji}/\sigma)^2}) \tag{1}$$

其中，$N$ 为节点数目 $|V|$；$m_j$ 为节点 $j$ 的质量，具体可以映射为实际网络中的某些属性，如组织成员的社会地位等；$d_{ji}$ 是节点 $j$ 到节点 $i$ 的距离，即最短路径长度；$\sigma$ 为影响因子，指示节点影响的范围，可根据节点势熵对其进行优选。

### 2.2　无向拓扑势熵

根据信息熵的概念，如果所有节点的拓扑势相等，则节点位置差异性的不确定程度最大，具有最大的熵；反之，如果每个节点的拓扑势都不相同，则不确定性最小，具有最小的熵。对于给定的网络 $G=(V,E)$ 及拓扑势场，其拓扑势场的势熵可表示为

$$H=-\sum_{i=1}^{N}\frac{\varphi(v_i)}{Z}\log\left(\frac{\varphi(v_i)}{Z}\right) \tag{2}$$

其中，$Z=\sum_{i=1}^{N}\varphi(v_i)$ 为标准化因子。

### 2.3　影响因子和拓扑势熵的关系

影响因子的设置影响节点的拓扑势，从而影响到拓扑势熵的大小。从图 1 可以看出，随着 $\sigma$ 由小到大，势熵 $H$ 先减后增；当势熵取极小值时，对应的横坐标为最优 $\sigma$。

### 2.4　影响因子和影响范围的关系

根据高斯函数“$3\sigma$”规则，每个对象的作用范围是以该对象为中心，半径为 $\frac{3\sigma}{\sqrt{2}}$ 的邻域空间。涂文燕等[10]指出：1) 当 $0<\sigma<\frac{\sqrt{2}}{3}$ 时，节点间没有相互作用，每个节点的势值等于 1；2) 当 $\frac{\sqrt{2}}{3}<\sigma<\frac{2\sqrt{2}}{3}$ 时，每个节点只影响一跳邻居节点，任意节点的拓扑势与其度数近似相差一个比例常数，此时的拓扑势影响力等价于按照节点度排序的影响力算法；3) 当 $\frac{2\sqrt{2}}{3}\leqslant\sigma<\sqrt{2}$ 时，每个节点影

响 2 跳以内可达节点。根据以上讨论，当 $\frac{\sqrt{2}}{3}l \leqslant \sigma < \frac{\sqrt{2}}{3}(l+1)$ 时，每个节点的影响范围为 $l$ 跳邻居节点。

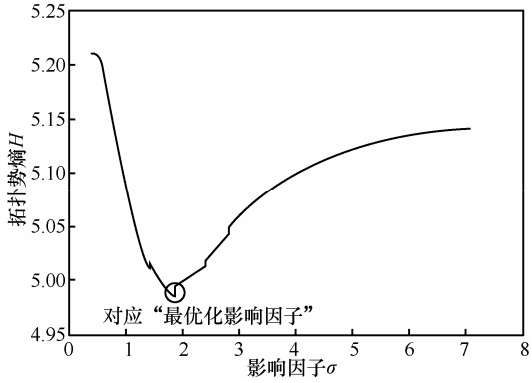



图 1 影响因子 $\sigma$ 和势熵 $H$ 的关系曲线

## 3 基于有向拓扑势的节点重要性评价

### 3.1 有向拓扑势

设有向加权网络 $G=(V,E,M,W)$，$M$ 是节点的属性集合，$W$ 是有向边的权重集合，节点 $j$ 在 $i$ 处的拓扑势值为式(3)，因此节点 $i$ 的入度拓扑势值 $\varphi_{\text{in}}(v_i)$ 和出度拓扑势值 $\varphi_{\text{out}}(v_i)$ 分别如式(4)、式(5)所示。

$$\varphi(v_j \to v_i) = m_j \mathrm{e}^{-(dw_{j\to i}/\sigma)^2} \tag{3}$$

$$\varphi_{\text{in}}(v_i) = \sum_{j=1}^{N}(m_j \mathrm{e}^{-(dw_{j\to i}/\sigma)^2}) \tag{4}$$

$$\varphi_{\text{out}}(v_i) = \sum_{j=1}^{N}(m_i \mathrm{e}^{-(dw_{i\to j}/\sigma)^2}) \tag{5}$$

$dw_{j\to i}$ 是“边权距”，即在边的权值影响下节点间距离。在实际网络中，如通话网络，边的权值一般表示 2 个节点之间的通话次数，边权重越大表明节点间的联系越密切，两者距离会相对变小。设节点 $j$ 到节点 $i$ 的最短路径依次通过边 $e_1,e_2,\cdots,e_h$，共有 $h$ 段，$d_r$ 为途经第 $r$ 段距离长度，$w_r$ 为对应的边的权重，则

$$dw_{j\to i} = \sum_{r=1}^{h}\frac{d_r}{w_r} \tag{6}$$

忽略节点本身的质量和每段距离的长度影响，假设节点质量均为 1，每段距离长度为 1。根据 $\sigma$ 确定的影响范围 $l$ 不再针对节点间跳数，而是针对边权距，即边权距小于等于 $l$ 的节点都在范围之内。

图 2 是一个简单的有向加权网络（记为“ExampleNet”），其对应的入度拓扑势、出度拓扑势随影响因子变化情况如图 3 所示。可见，当影响因子从 0.47 增大到 1.41 时，由节点在网络中位置的不同而产生的节点拓扑势值差异性就越来越明显；当取最优化 $\sigma_{\text{opt}}=2.36$ 时，区分度最高，因此应该以 $\sigma_{\text{opt}}$ 计算二维拓扑势值。

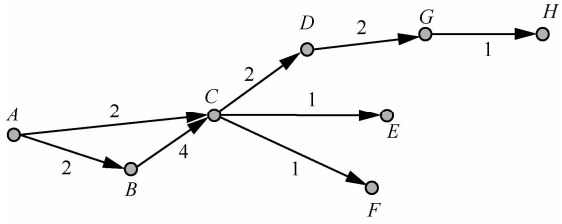



图 2 有向加权网络示例

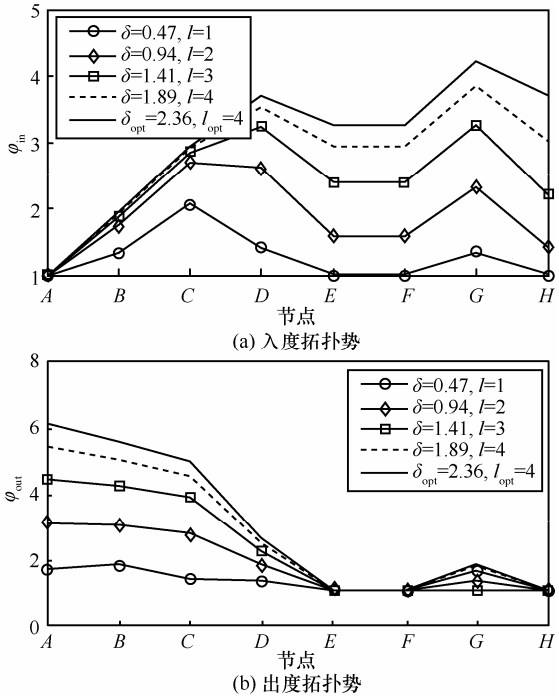



图 3 不同的影响因子计算出的入/出度拓扑势

### 3.2 基于影响范围的有向拓扑势算法

针对有向加权网络，以拓扑势理论分析节点重要性，主要过程是：1）获取该网络的最优影响因子 $\sigma$；2）使用最优影响因子 $\sigma_{\text{opt}}$ 来计算每个节点的出度、入度拓扑势值；3）依据节点出度、入度拓扑势值的大小给出节点重要性的排序结果。

根据节点拓扑势的公式，若要计算某个节点的拓扑势，需先获得该节点与其他所有节点的边权距；若要获得网络所有节点的拓扑势，则需要任意 2 个节点之间的边权距。边权距的计算本质是求解两点间的最短距离，求解网络所有节点之间最短路

径的常用算法是 Dijkstra 和 Floyd 算法，其复杂度均为 $O(N^3)$，显然不适于规模较大的网络。而在求解最优化影响因子时，要反复计算网络各节点的拓扑势熵，时间复杂度为 $O(N^3 s)$，$s$ 为迭代次数。

优化节点拓扑势的计算过程，从 2 方面考虑：1) 根据高斯函数的“$3\sigma$”规则，某节点的影响范围只是 $[0,\frac{3\sigma}{\sqrt{2}})$，在影响范围之外的节点不需要计算边权距，因为它们与该节点的拓扑势值约等于 0；2) 随着影响因子 $\sigma$ 从 0 不断增加，网络拓扑势熵 $H$ 从高到低、从低再高，需要求得 $H$ 最小时对应的最优化 $\sigma$，而 $\sigma$ 能直接换算成影响范围，因此在求解最优化 $\sigma$ 的过程中，可获得节点的影响范围 $l$ 的下限和上限。

根据上文的分析，定义节点在影响范围内的入度、出度邻域，它们是节点在计算拓扑势时需要考虑的节点集合。

**定义 1**　节点 $i$ 在影响范围 $l$ 内的入度邻域：有向加权图中，与节点 $i$ 的边权距小于等于 $\lceil l \rceil$，且能到达 $i$ 的节点集合，记为 $InAdj_i^{(l)}$。

**定义 2**　节点 $i$ 在影响范围 $l$ 内的出度邻域：有向加权图中，与节点 $i$ 的边权距小于等于 $\lceil l \rceil$，且从 $i$ 出发能到达的节点集合，记为 $OutAdj_i^{(l)}$。

当 $l=0$ 时，节点 $i$ 的入度邻域和出度邻域为自身；随着 $l$ 增大，节点的入度、出度邻域大小不变或增大。

本文提出一种基于影响范围的节点有向拓扑势计算方法，见算法 1。其中，第 1~10 步是预估 $l$，确保在 $(l-\Delta l, l)$ 会出现 $H$ 从低变高的转折点；第 11 步是快速探索影响范围 $(l-\Delta l, l)$ 之内合适的参数；第 12 步挑选最优化 $\sigma_{\text{opt}}$ 和 $l_{\text{opt}}$；第 13~20 步根据最优化参数得到指定节点的入度、出度拓扑势。需要计算多个节点入度、出度拓扑势时，前 12 步只需运算一次，再逐一计算每个节点的结果。

**算法 1**　基于影响范围的节点有向拓扑势计算方法

输入：$v_i \in V$，$G=(V,E,W)$，每次迭代影响范围的增量 $\Delta l$；

输出：$\varphi_{\text{in}}(v_i), \varphi_{\text{out}}(v_i)$

步骤：

1) $l=0$；
2) $H = minH = \log(N)$；
3) While $H \leqslant minH$ Do
4) $minH = H$；
5) $l = l + \Delta l$；
6) $\sigma = \sqrt{2}l/3$；
7) 计算并记录 $InAdj_i^l$ 和 $OutAdj_i^l$，$i=1,\cdots,N$；
8) 根据 $l$、$\sigma$ 计算所有节点的入度拓扑势；
9) 计算 $H$，并保存 $l$、$\sigma$ 和 $H$；
10) End While
11) 在 $(l-\Delta l, l)$ 之间用黄金分割法多次选择 $l'$，求对应的 $\sigma$ 和 $H$，并保存三者；
12) 选择所有记录中 $H$ 最小时，对应的 $\sigma$ 和 $l$，令 $\sigma_{\text{opt}}=\sigma$，$l_{\text{opt}}=l$；
13) $\varphi_{\text{in}}(v_i)=0$；
14) For $v_j$ in $InAdj_i^{(l_{\text{opt}})}$ Do
15) $\varphi_{\text{in}}(v_i) += \varphi(v_j \to v_i)$；
16) End For
17) $\varphi_{\text{out}}(v_i)=0$；
18) For $v_j$ in $outAdj_i^{(l_{\text{opt}})}$ Do
19) $\varphi_{\text{out}}(v_i) += \varphi(v_i \to v_j)$；
20) End For
21) Return $\varphi_{\text{in}}(v_i), \varphi_{\text{out}}(v_i)$

算法 1 计算某个节点入度、出度拓扑势，耗时 $O(k)$，其中 $k$ 为平均 $l$ 范围之内邻域的节点数；计算全网所有节点入度、出度拓扑势，耗时为 $O(Nk)$；计算最优化 $\sigma_{\text{opt}}$ 和 $l_{\text{opt}}$，耗时 $O(Nks)$。由于实际网络关系稀疏，导致 $k$ 远远小于 $N$，故算法 1 时间复杂度较低。

## 4　基于二维拓扑势的节点角色发现

### 4.1　角色的定性描述

根据 ExampleNet 中最优化 $\sigma_{\text{opt}}=2.36$ 时各节点的入度和出度拓扑势，可绘制节点二维拓扑势分布情况（如图 4 所示）。入度拓扑势代表节点受其他节点影响的程度，出度拓扑势用于度量对其他近邻节点的影响能力。

基于“拓扑势相似的节点属于相同角色”的假设，根据二维拓扑势分布图将节点的角色划分为 4 种。

角色 1：桥接节点，其入度拓扑势和出度拓扑势均较大，在网络中起到承上启下的左右（位于区域 I），如 $C$ 节点；

角色 2：贡献节点，具有相对较大的出度拓扑

势（位于区域 II），如 *A*、*B* 节点；

角色 3：接收节点，一般具有相对较大的入度拓扑势（位于区域 III），如 *G* 节点；

角色 4：普通节点，其出度和入度拓扑势均不显著（位于区域 IV）。

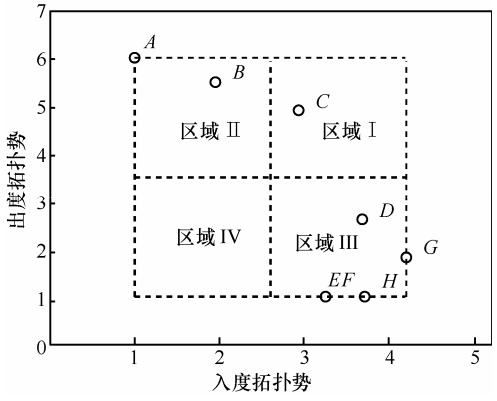



图 4 ExampleNet 的二维拓扑势分布

### 4.2 角色的定量描述

在图 4 中，每个区域代表了一种角色，而每个节点从属于该区域的程度是不同的，例如，同属于区域 II 的节点 *A* 和节点 *B*，节点 *A* 比节点 *B* 属于该区域的程度更大。因此将二维图区域的 4 个顶点定义为各个区域的“锚定点”(anchor)，用于表征该区域的特征属性。

$Anchor_{\mathrm{I}}(maxInTopo, maxOutTopo)$ 代表区域 I；

$Anchor_{\mathrm{II}}(1, maxOutTopo)$ 代表区域 II；

$Anchor_{\mathrm{III}}(maxInTopo, 1)$ 代表区域 III；

$Anchor_{\mathrm{IV}}(1,1)$ 代表区域 IV。

其中，*maxInTopo*、*maxOutTopo* 是网络节点二维拓扑势的入度、出度拓扑势的最大值。区域 I 的 (*maxInTopo*，*maxOutTopo*) 表明节点具有较大的出入度拓扑势，而区域 IV 的原点(1, 1) 代表了该区域中节点的出入度拓扑势均较小。节点 *i* 离锚定点 $Anchor_k$ 的距离( $d_{ik}$ )越近，则从属于该类角色的概率值( $p_{ik}$ )越大。

将节点 *i* 从属于 4 种角色的概率值组成“角色分布向量” $\boldsymbol{P}_i =< p_{i1}, p_{i2}, p_{i3}, p_{i4} >$，其中每个分量称为“角色从属概率”，记为式(7)，且满足 $\sum_{k=1}^{4} p_{ik}=1$。

$$p_{ik}=\frac{\sum_{a=1}^{4} d_{ia}-d_{ik}}{3\sum_{a=1}^{4} d_{ia}}, k\in\{1,2,3,4\} \tag{7}$$

在获得节点各种角色的从属概率后，比较它们的大小，选择概率最大的角色作为节点的角色。即当 $p_{ik}$ 大于等于任意 $p_{ia}$，且 $k\neq a$ 时，称节点 *i* 从属于角色 $role_k$，记为 $v_i \in role_k$。

### 4.3 节点角色发现算法

先计算节点 *i* 的角色分布向量 $\boldsymbol{P}_i$，其最大分量对应的角色就是节点 *i* 的角色。每个节点至少对应 1 个角色。根据分量的大小，可以判断隶属该角色的程度。综上，本文提出了节点角色发现算法（见算法 2）。

**算法 2** 节点角色发现算法

输入：边集合 *EdgeSet*；

输出：所有节点的角色分布向量 $\boldsymbol{P}$，4 类角色所对应的节点及程度

步骤：

1) 根据边集合 *EdgeSet*，构建有向加权网 $G=(V,E,W)$；

2) 计算最优影响因子 $\sigma_{opt}$ 及对应的跳数 $l_{opt}$；

3) 计算所有节点的入度拓扑势和出度拓扑势；

4) 根据最大入度、出度拓扑势构建 4 类角色锚定点；

5) 计算所有节点到 4 类锚定点的距离；

6) 计算所有节点的角色分布向量 $\boldsymbol{P}$；

7) 获得 4 类角色所属节点及程度；

8) Return $\boldsymbol{P}$，4 类角色所对应的节点及程度。

## 5 动态网络演化

动态网络包含了多个时间片的网络信息。本节主要介绍用于检测时序图演化的指标。

### 5.1 全网角色比例向量

以网络 4 种角色对应的节点比例所组成全网角色比例向量，记为 $\boldsymbol{R}=< r_1, r_2, r_3, r_4 >$，其分量计算方法见式(8)，且满足 $\sum_{k=1}^{4} r_k=1$。

$$r_k=\frac{|\{v_i \mid v_i \in role_k\}|}{N}, k\in\{1,2,3,4\} \tag{8}$$

### 5.2 全网角色分布向量

将所有节点的角色分布向量求均值，得到全网角色分布向量 $\boldsymbol{P}=< p_1, p_2, p_3, p_4 >$，其分量称为“全网角色从属概率”，记为式(9)，且满足 $\sum_{k=1}^{4} p_k=1$。

$$p_k=\frac{\sum_{i=1}^{N} p_{ik}}{N}, k\in\{1,2,3,4\} \tag{9}$$

### 5.3　全网角色距离向量

令 $\boldsymbol{D}=\boldsymbol{P}^{(2)}-\boldsymbol{P}^{(1)}$，表示 2 个数据集的节点行为角色分布变化。其中 $\boldsymbol{P}^{(1)}$、$\boldsymbol{P}^{(2)}$ 为数据集 1、2 对应的全网角色分布向量。同理，以 $\boldsymbol{D}_i=\boldsymbol{P}_i^{(2)}-\boldsymbol{P}_i^{(1)}$ 描述单个节点行为角色的演化。

### 5.4　排名距离

全网角色比例向量、全网角色分布向量、全网角色距离向量是从网络全局角色分布的变化来刻画网络演化，但当网络中发生较大变化时，出入度拓扑势排名前 $n$ 个节点会发生较大的改变，因此本文定义了“排名距离”，用于描述 2 个数据集排名差异程度的指标，见式(10)。其中，$n$ 是指前 $n$ 个排名；$Rank^{(2)}(v_i)$ 指节点 $i$ 在数据集 2 的排名；$topNSet$ 是数据集 1 和 2 的前 $n$ 个节点并集。

$$D_n=\sum_{v_i\in topNSet}|Rank^{(2)}(v_i)-Rank^{(1)}(v_i)| \tag{10}$$

为了排除 $n$ 的影响，给出“排名变化率”的计算公式，见式（11）。默认每个数据集前 $n$ 个节点排名之和为 $n(n+1)/2$，有 2 个数据集故分母设为 $n(n+1)$。

$$R_n=\frac{D_n}{n(n+1)} \tag{11}$$

例如，网络中 3 个节点($A$，$B$，$C$)在 $t$−1 时刻入度拓扑势排名为 1、2、3，而在 $t$ 时刻，入度拓扑势的排名为 3、2、1，Top2 排名距离为 |3−1|+|2−2|+|1−3|=4（考虑 $A$、$B$、$C$），排名变化率为 4/6=2/3；若 $t$ 时刻，入度拓扑势的排名为 2、1、3，Top2 排名距离为|2−1|+|1−2|=2（只考虑 $A$、$B$），排名变化率为 2/6=1/3。

当 2 个数据集节点不完全一致，出现节点缺失、得不到排名的情况，则该节点默认排名为 $n$+1，即在所有已知节点排名之后。例如，$D$ 在($A$，$B$，$C$)数据集中 Top2 的排名为 3，Top3 排名为 4。

## 6　实验分析

为了评判入度、出度拓扑势在节点重要性方面的效果时，选取了 11 种经典的对比指标，包括：1）链接分析方法，含 HITS-authority（内容权威度）、HITS-hub（链接权威度）、page rank；2）中心性分析方法，含 weighted degree（加权度）、weighted indegree（加权入度）、weighted outdegree（加权出度）、betweenness centrality（介数中心性）、closeness centrality（接近中心性）、eigenvector centrality（特征向量中心性）、clustering coefficient（聚类系数）、eccentricity（离心率）。

### 6.1　Strike 数据集的实验

Strike 数据集描述了一场罢工运动，图 5 为罢工运动中工人之间的沟通网络，以工人为节点（24 个），工人之间的联系为有向边（38 条），边权值均为 1（相当于无权值）。数据背景中介绍，由于语言的不同，说西班牙语的工人与说英语的工人之间的沟通较少，其中 Bob 说一些西班牙语并且和 Norm 关系密切。

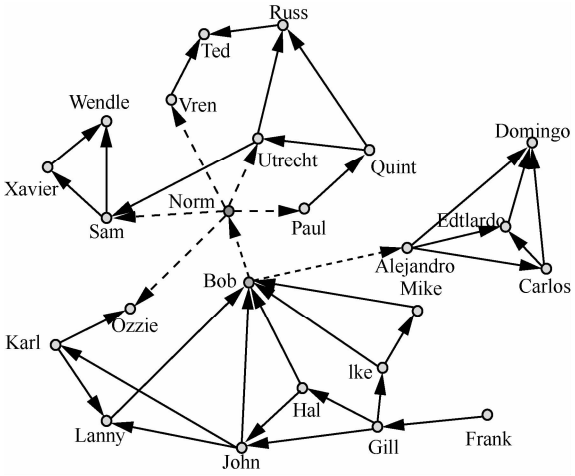



图 5　Strike 人员沟通网络

从表 1 和表 2 可看出：1）入度拓扑势和出度拓扑势排名前 5 的节点，被其余重要性度量指标所认可,例如，入度拓扑势排名第 1 的 Sam，也被另 7 种指标选为前 5 名；2）入度拓扑势与加权入度、内容权威度的排名结果最接近，但是前者精度更高，例如，加权入度认为 Sam、Ted、Utrecht 分值一样，并列第三，而入度拓扑势由于采用优化的 $\sigma_{\text{opt}}$，排名并列少，区分度高。同理，出度拓扑势与加权出度的排名结果相近，但前者精度高。从计算复杂度来看，入度、出度拓扑势比加权度、加权入度、加权出度、聚类系数、离心率高，比介数中心性低，与其他对比方法基本相当。

据图 6 划分的区域及对应的角色定义，可看出：1）Bob 和 Norm 属于“桥接节点”（图右上方），有较大的入度、出度拓扑势，在罢工运动中是核心的领导、联系人。特别是 Norm，其入度、出度拓扑势排名并不出众（分别是 8 和 5），但在图 6 可清晰看出其所属角色和重要程度。2）其余工人分别属于“贡献节点”（图左上方）和“接收节点”（图右下方），不存在“普通节点”（图左下方）。

**表 1**　　**Strike 入度拓扑势排名前 5 的节点信息**

| 节点 | 入度拓扑势 | 出度拓扑势 | 加权度 | 加权入度 | 加权出度 | 介数中心性 | 接近中心性 | Page Rank | 内容权威度 | 链接权威度 | 聚类系数 | 离心率 | 特征向量中心性 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sam | 1 | 15 | 4 | 3 | 5 | 4 | 14 | 9 | 3 | 2 | 11 | 14 | 3 |
| Bob | 2 | 1 | 1 | 1 | 5 | 1 | 10 | 1 | 1 | 1 | 18 | 10 | 10 |
| Domingo | 3 | 21 | 8 | 2 | 21 | 16 | 21 | 2 | 2 | 21 | 1 | 21 | 1 |
| Ted | 4 | 21 | 17 | 3 | 21 | 16 | 21 | 3 | 3 | 21 | 20 | 21 | 5 |
| Utrecht | 5 | 12 | 4 | 3 | 5 | 7 | 12 | 10 | 3 | 2 | 11 | 12 | 6 |

**表 2**　　**Strike 出度拓扑势排名前 5 的节点信息**

| 节点 | 入度拓扑势 | 出度拓扑势 | 加权度 | 加权入度 | 加权出度 | 介数中心性 | 接近中心性 | Page Rank | 内容权威度 | 链接权威度 | 聚类系数 | 离心率 | 特征向量中心性 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Bob | 2 | 1 | 1 | 1 | 5 | 1 | 10 | 1 | 1 | 1 | 18 | 10 | 10 |
| Hal | 21 | 2 | 8 | 12 | 5 | 10 | 7 | 22 | 12 | 8 | 7 | 4 | 21 |
| John | 19 | 3 | 3 | 3 | 2 | 6 | 8 | 14 | 3 | 2 | 10 | 4 | 19 |
| Gill | 23 | 4 | 4 | 12 | 2 | 5 | 3 | 17 | 12 | 19 | 17 | 2 | 23 |
| Norm | 8 | 5 | 2 | 12 | 1 | 2 | 13 | 5 | 12 | 8 | 19 | 12 | 11 |

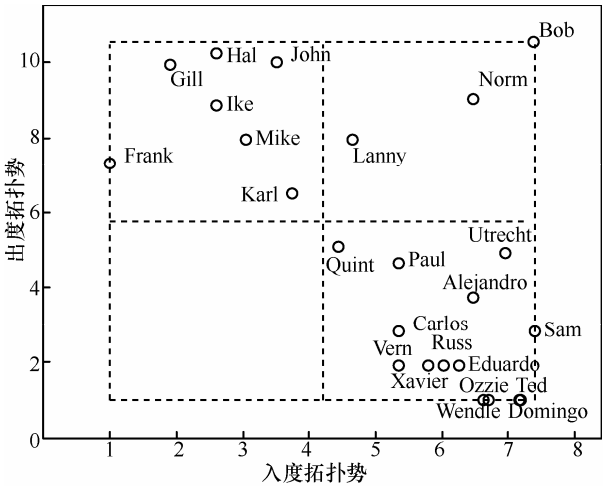



图 6　Strike 二维拓扑势分布

结合数据背景，进一步分析可得：1) 离 $Anchor_{\mathrm{I}}$ 越近的节点，在运动中发挥的作用越大；2) 离 $Anchor_{\mathrm{III}}$ 越近的节点，其入度拓扑势大、出度拓扑势小，可能属于听指挥的普通群众，例如 Domingo，只有别人联系他，他从不联系别人；3) 具有最小入度或出度拓扑势的点，必然位于网络边缘，例如 Frank、Domingo、Wendle；4) 由于网络直径为 6，选择的优化 $\sigma_{\mathrm{opt}}$=3.296，对应影响范围为 $l_{\mathrm{opt}}$=6，覆盖了整个网络，各节点的入度、出度拓扑势至少有一项数值较高，故未发现“普通节点”（图左下方），也就是说运动中各工人都比较积极，未出现被孤立的情况。

综上，从 Strike 数据集的实验中能得出结论：1）本文提出的入度、出度拓扑势指标能反应节点重要性；2）根据区域、角色、锚点的定义，可以从二维拓扑势分布图中识别各类用户角色，并判断其作用和所属程度。

### 6.2　北邮人数据集的实验

采自北邮人论坛 2004 年 6 月 1 日到 2011 年 12 月 29 日 6 个版块。以用户为节点，用户之间回帖关系为边，回帖次数为权值，构建 6 张有向加权网络，其基本特征如表 3 所示。

从图 7 可以直观看出各版块大部分用户都属于“普通节点”，少数人属于“桥接节点”。表 4 的“全网角色比例向量” $\boldsymbol{R}$ 显示“普通用户”比例远远高

**表 3**　　**北邮人回复关系网络的特征**

| 数据集 | 中文名 | 节点数 | 边数 | 平均度 | 平均加权度 | 网络直径 | 图密度 | 模块度 | 社团个数 | 弱联通分量数目 | 强连通分量数目 | 平均聚类系数 | 平均路径长 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Talking | 谈天说地 | 2 070 | 6 627 | 3.201 | 3.879 | 10 | 0.02 | 0.403 | 68 | 48 | 1 158 | 0.068 | 3.912 |
| Feeling | 情感天空 | 3 019 | 13 050 | 4.323 | 5.671 | 11 | 0.001 | 0.335 | 48 | 28 | 1 665 | 0.06 | 3.816 |
| Friends | 缘来如此 | 1 206 | 2 021 | 1.676 | 1.934 | 18 | 0.001 | 0.664 | 118 | 92 | 901 | 0.029 | 5.228 |
| Home | 安居乐业 | 1 663 | 6 811 | 4.096 | 5.425 | 9 | 0.002 | 0.371 | 29 | 18 | 792 | 0.095 | 3.684 |
| Joke | 笑口常开 | 3 606 | 13 179 | 3.655 | 4.925 | 14 | 0.001 | 0.346 | 96 | 69 | 2 283 | 0.082 | 3.892 |
| Picture | 贴图秀 | 2 407 | 4 151 | 1.725 | 2.486 | 14 | 0.001 | 0.6 | 264 | 162 | 1 910 | 0.031 | 4.648 |

于其他角色；而 “全网角色分布向量” ***P*** 虽然显示“普通节点”最多，但和其他角色相比，差异并不大。既然有了 ***R***，为什么还要定义 ***P*** 呢？下一节实验中将会展示，***P*** 能捕捉动态网络结构变化所引发的角色从属概率的变化，从而识别事件的发生。

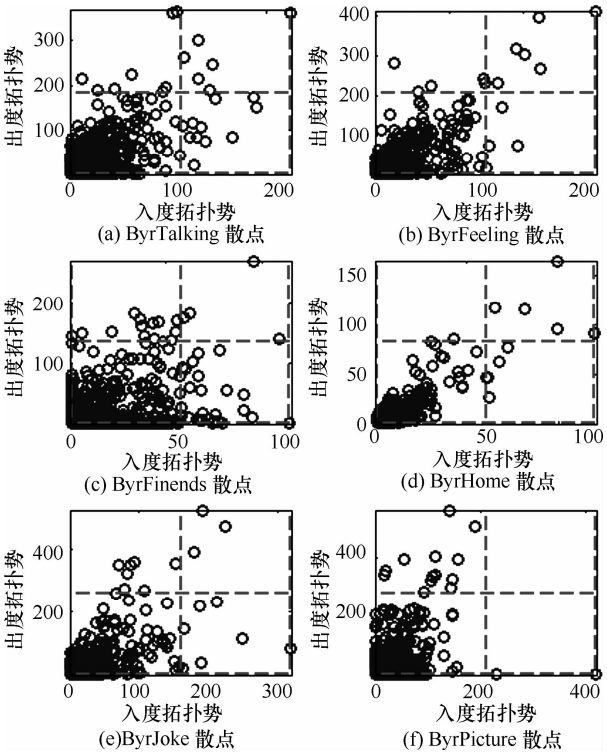



图 7　北邮人二维拓扑势分布

**表 4　　北邮人角色统计信息**

| 数据集 | 全网角色比例向量 | 全网角色分布向量 |
|---|---|---|
| Talking | <0.003,0.004,0.005,0.988> | <0.196,0.216,0.262,0.327> |
| Feeling | <0.002,0.002,0.001,0.995> | <0.196,0.221,0.253,0.330> |
| Friends | <0.003 3,0.011 6,0.021 6,0.964> | <0.194,0.203,0.280,0.323> |
| Home | <0.003,0.001,0.003,0.993> | <0.196,0.222,0.253,0.330> |
| Joke | <0.001,0.002,0.002,0.995> | <0.194,0.214,0.261,0.331> |
| Picture | <0,0.006,0.001,0.994> | <0.196,0.223,0.251,0.330> |

着重研究分析“桥接节点”、“贡献节点”、“接收节点”。选取各数据集、上述 3 种角色，所属概率最高的用户，统计各类信息，得到表 5。

1) “身份”，分为普通用户和管理员。在 6 个数据中发过帖的用户共有 16 549 名，普通用户占 98.36%，管理员仅占 1.63%；而本文识别出的最典型用户共 17 名，普通用户 9 名，管理员 8 名。现实中，管理员确实发挥着重要作用，如管理版务、贡献原创素材、调节纷争、引导舆论等，因此识别出近 50%的典型用户是管理员恰恰说明本文算法是合理的。

2) “总帖数”、“积分”和“活跃度”是用户在全论坛活跃程度的标志，而“该版块发帖数”和“该版块参与话题数”体现用户在所处版块的活跃程度。选中用户的数值几乎都高于平均水平，从侧面反映他们具有代表性。

3)“回复帖数”和“被回复帖数”，分别是用户评论他人和被他人评论的次数，与节点的边数、边权值有直接关联，因此一般而言“桥接节点”两项数值都高，“贡献节点”的回复帖数偏高，“接收节点”的被回复帖数偏高。

4) “最大出度跳数”和“最大入度跳数”，决定了拓扑势影响范围的边界，有助于判断节点位置。例如，典型的“接收节点”最大出度跳数极小，说明它们靠近网络边缘，出度拓扑势较小；而最大入度跳数大，故入度拓扑势偏大。

分析各角色对应节点的具体情况，以“Talking”的“桥接节点”11 574 为例。该版块只有 30.90%的帖子指明了回复对象，而用户 11 574 发帖的回复率为 168.27%；该用户回复了 84 个用户的帖子，同时被 132 人回复。可见该用户具有很高的互动性，符合“桥接节点”的特性。

表 5 中，有 2 处看似不合理的地方：1)“Friends”的用户 20 503，回复帖数和被回复帖数极低，却被选为典型的“贡献节点”；2)“Friends”的用户 31 732，在该版块仅发了 2 个帖，却被选为典型的“接收节点”。原因在于“Friends”版块回复关系稀疏，加权度低，造成整体的入度、出度拓扑势都低；2 位用户所处网络位置好，邻域节点多。

### 6.3　VAST2008 数据集的实验

该数据集是 VAST2008 可视化分析竞赛中使用的电话网络数据。Caralano/Vidro 社会网络包含了 2006 年 6 月 10 天内的 400 个独立的电话号码之间的通话数据。其中，在第 8, 9 天发生了一个事件，使通信网络发生了变化[16]。本节实验侧重于研究动态网络演化过程，特别是本文方法能否捕捉到第 8, 9 天的事件。

首先，以用户为节点，通话关系为有向边，通话次数为边权值，构建每天的有向通话网络，基本特征见表 6。再计算每天每个节点的入度、出度拓扑势，可以清晰看见第 1,2 天和第 8,9 天时全网拓扑势分布明显不同（见图 8）。如果构建网络时默认边权值为 1，得到图 9，未发现第 8,9 天

**表 5** **北邮人典型用户的统计信息**

| 角色 | 数据集 | 节点 | 身份 | 总帖数 | 积分 | 生命值 | 该版块发帖数 | 该版块参与话题数 | 回复帖数 | 被回复帖数 | 最大出度跳数 | 最大入度跳数 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 桥接节点 | Talking | 11 574 | 用户 | 8 967 | 2 709 | 365 | 681 | 575 | 104 | 175 | 5 | 5 |
| | Feeling | 3 672 | 用户 | 3 897 | 2 254 | 365 | 648 | 229 | 468 | 544 | 2 | 6 |
| | Friends | 836 | 管理员 | 13 544 | 1 271 | 135 | 206 | 155 | 133 | 41 | 3 | 10 |
| | Home | 523 | 管理员 | 46 703 | 5 604 | 666 | 419 | 151 | 407 | 260 | 5 | 1 |
| | Joke | 590 | 管理员 | 89 785 | 6 677 | 365 | 532 | 256 | 480 | 410 | 5 | 7 |
| 贡献节点 | Talking | 598 | 管理员 | 24 108 | 4 293 | 365 | 129 | 100 | 107 | 53 | 2 | 6 |
| | Feeling | 14 458 | 用户 | 968 | 2 516 | 365 | 432 | 427 | 401 | 32 | 6 | 7 |
| | Friends | 20 503 | 用户 | 1 102 | 1 872 | 365 | 12 | 8 | 5 | 3 | 10 | 12 |
| | Home | 565 | 管理员 | 2 718 | 1 274 | 365 | 136 | 109 | 128 | 46 | 5 | 1 |
| | Joke | 838 | 管理员 | 18 375 | 4 265 | 365 | 470 | 425 | 275 | 54 | 4 | 7 |
| | Picture | 836 | 管理员 | 13 544 | 1 271 | 135 | 643 | 578 | 478 | 51 | 4 | 7 |
| 接收节点 | Talking | 18 903 | 用户 | 3 521 | 2 337 | 365 | 95 | 51 | 43 | 63 | 1 | 6 |
| | Feeling | 590 | 管理员 | 89 785 | 6 677 | 365 | 271 | 182 | 253 | 176 | 1 | 6 |
| | Friends | 31 732 | 用户 | 22 | 16 | 134 | 2 | 1 | 1 | 59 | 1 | 10 |
| | Home | 18 997 | 用户 | 7 442 | 6 597 | 666 | 151 | 129 | 30 | 85 | 0 | 1 |
| | Joke | 5 011 | 用户 | 40 321 | 4 356 | 365 | 939 | 410 | 594 | 941 | 1 | 6 |
| | Picture | 302 | 用户 | 12 471 | 1 381 | 666 | 29 | 19 | 10 | 316 | 1 | 7 |

**表 6** **VAST2008 通话关系网络的特征**

| 数据集 | 节点数 | 边数 | 平均度 | 平均加权度 | 网络直径 | 图密度 | 模块度 | 社团个数 | 平均聚类系数 | 平均路径长 |
|---|---|---|---|---|---|---|---|---|---|---|
| 第 1 天 | 370 | 633 | 1.711 | 2.668 | 24 | 0.005 | 0.74 | 22 | 0.016 | 8.957 |
| 第 2 天 | 373 | 617 | 1.654 | 2.584 | 29 | 0.004 | 0.768 | 24 | 0.007 | 11.646 |
| 第 3 天 | 374 | 627 | 1.676 | 2.548 | 35 | 0.004 | 0.748 | 22 | 0.005 | 10.151 |
| 第 4 天 | 374 | 625 | 1.671 | 2.709 | 32 | 0.004 | 0.756 | 22 | 0.007 | 10.302 |
| 第 5 天 | 373 | 627 | 1.681 | 2.657 | 22 | 0.005 | 0.754 | 22 | 0.006 | 7.567 |
| 第 6 天 | 373 | 629 | 1.686 | 2.582 | 33 | 0.005 | 0.747 | 19 | 0.008 | 9.383 |
| 第 7 天 | 367 | 603 | 1.643 | 2.55 | 22 | 0.004 | 0.755 | 27 | 0.006 | 8.063 |
| 第 8 天 | 365 | 622 | 1.704 | 2.753 | 30 | 0.005 | 0.748 | 20 | 0.01 | 9.433 |
| 第 9 天 | 374 | 632 | 1.69 | 2.626 | 24 | 0.005 | 0.751 | 24 | 0.012 | 9.616 |
| 第 10 天 | 384 | 662 | 1.724 | 2.708 | 33 | 0.005 | 0.748 | 21 | 0.004 | 10.808 |

拓扑势分布异常，从侧面说明边权值的确是一个不可忽视的维度。

计算 10 天的全网角色比例向量，得到图 10。可见角色 2 一直占有较高的比例，角色 3 一直较少，角色 1 和角色 4 呈此消彼长的态势；可看出第 1,2 天角色比例发生重大变化，第 8,9 天变化不明显。计算全网角色分布向量，得到图 11。可见角色 2 和 4，角色 1 和 3 分布比例及变化趋势较为一致，且第 1, 2 天、8, 9 天变化明显。

计算全网角色距离向量，得到图 12，幅度越大说明角色分布的变化越大，可见角色 2 和 4，角色 1 和 3 变化幅度较为一致，且第 1, 2 天、第 8, 9 天变化明显。

计算相邻 2 天，前 5、10、15、20、25 名的入度、出度排名变化率，得到图 13。首先观察入度拓扑势排名变化（如图 13(a)），Top5 曲线相对平滑，特别是在第 1、2 天时很低，说明入度拓扑势前 5 名浮动较小，而第 5 名之后变化较大；第 8, 9 天前 5 名几乎彻底变更，即被呼叫次数最多的用户改变了，并且在第 9, 10 天未出现改变，证明第 8, 9

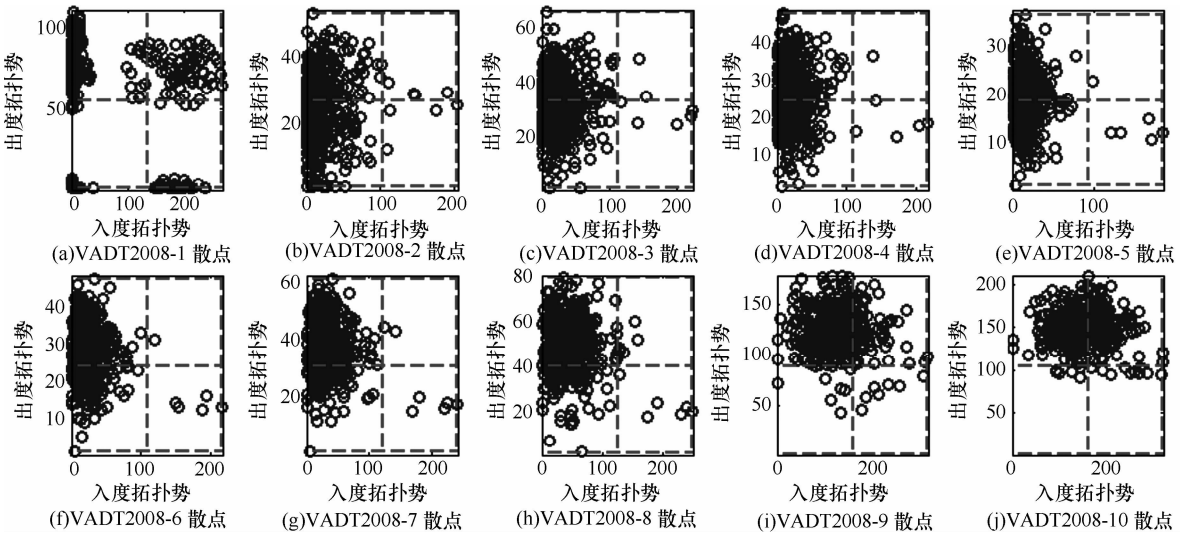



图 8　VAST2008（有权值）二维拓扑势分布

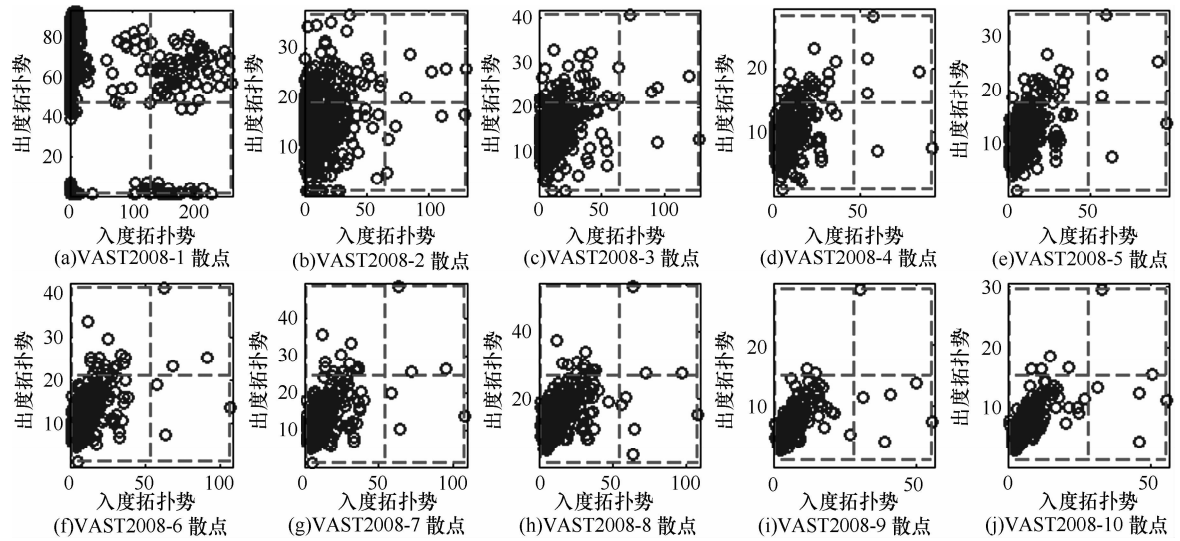



图 9　VAST2008（无权值）二维拓扑势分布

天出现了重大事件。再观察出度拓扑势排名变化（如图 13(b)所示），曲线也出现层次现象，根据高低可判断出度拓扑势前 10 名变动较大；在第 7, 8 天有个较小的上扬，并不能说明出现较大变动。

综上，各方法均检测出第 1, 2 天有重大变化；而拓扑势分布、全网角色分布向量、全网角色距离向量、拓扑势排名变化率等方法能识别第 8, 9 天的事件。能检测出事件发生的时间点是因为事件发生时，用户的交互频率、对象出现变化，影响了入度、出度拓扑势的取值，进而导致节点角色改变；根据角色在全局和局部 2 方面的比较，能发现动态网络中角色变化幅度大的时间点，即为事件发生的时刻。

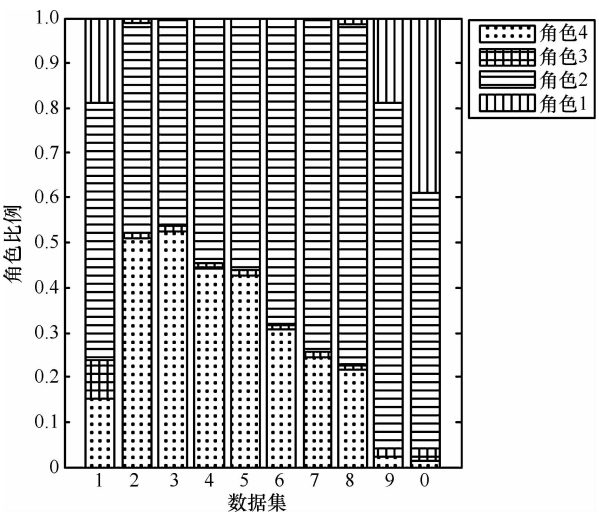



图 10　VAST2008 全网角色比例

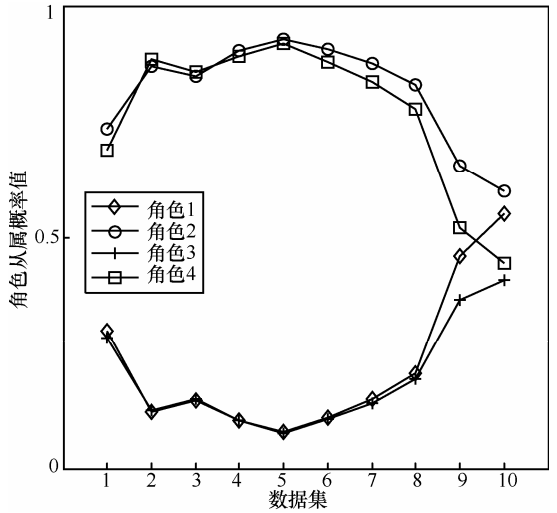



图 11　VAST2008 全网角色分布

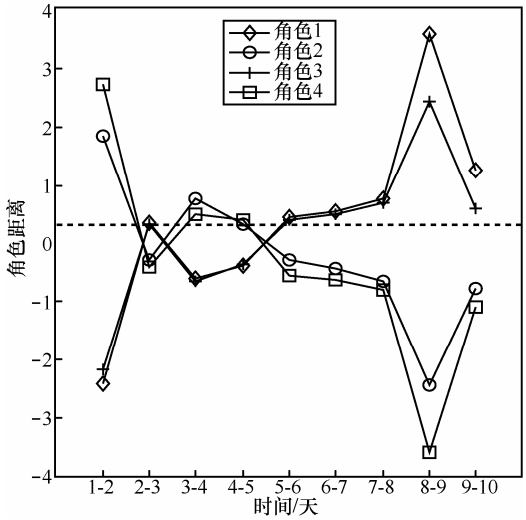
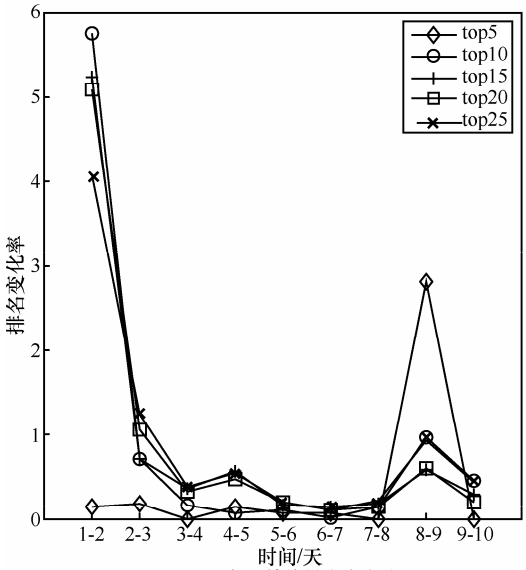
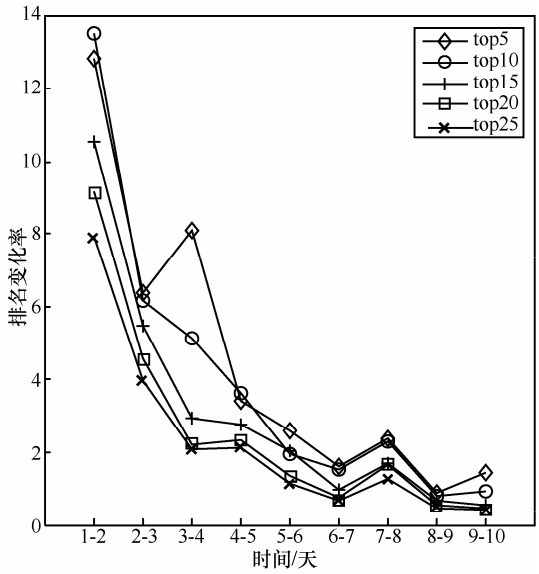



图 12　VAST2008 角色距离演化情况




(a) 入度拓扑势排名变化率




(b) 出度拓扑势排名变化率

图 13　VAST2008 入\出度拓扑势排名变化

## 7　结束语

本文针对有向加权网络中节点的重要性和角色发现的问题，引入了二维有向拓扑势，并提出节点角色发现算法；从角色变化的角度分析动态网络演化过程。后续工作将引入社团信息，分析各节点在社团内外的行为特点，进一步深化角色分析和动态网络演化的研究。

## 参考文献：

[1] FREEMAN L C. The development of social network analysis: A study in the sociology of science[M]. Vancouver: Empirical Press, 2004.

[2] ZHU T, WANG B, WU B. Role defining using behavior-based clustering in telecommunication network[J]. Expert Systems with Applications, 2011, 38(4): 3902-3908.

[3] LI B X, LI M J. A brief review of applications of social network analysis against terrorism[J]. Complex Systems and Complexity Science, 2012, 9(2): 85-93.

[4] SCOTT J. Social Network Analysis: A Handbook[M]. California: SAGE Publications, 1991.

[5] LORRAIN F, WHITE H. Structural equivalence of individuals in social networks[J]. Journal of Mathematical Socioology, 1971, 1(1): 49-80.

[6] GUIMERA R, SALES-PARDO M, AMARAL L A N. Classes of complex networks defined by role-to-role connectivity profiles[J]. Nature Physics, 2007, 3: 63-69.

[7] HU J, HAN Y N, HU J. Topological potential: modeling node importance with activity and local effect in complex networks[A]. Computer

Modeling and Simulation, Second International Conference[C]. 2010.411-415.

[8] HE N, GAN W Y, LI D Y. Evaluate nodes importance in the network using data field theory[A]. International Conference on Convergence Information Technology[C]. 2007.1225-1230.

[9] 肖俐平, 孟晖, 李德毅. 基于拓扑势的网络节点重要性排序及评价方法[J]. 武汉大学学报(信息科学版), 2008, 33(4): 379-383.
XIAO L P, MENG H, LI D Y. Approach to node ranking in a network based on topology potential[J]. Geometrics and Information Science of Wuhao University, 2008, 33(4):379-383.

[10] 淦文燕, 赫南, 李德毅等. 一种基于拓扑势的网络社区发现方法[J]. 软件学报, 2009, 20(8): 2241-2254.
GAN W Y, HE N, LI D Y, *et al*. Community discorery method in networks based on topology poetical[J]. Journal of Software, 2009,20(8): 2241-2254.

[11] 张健沛, 李泓波, 杨静等. 基于拓扑势的网络社区结点重要度排序算法[J]. 哈尔滨工程大学学报, 2012, 33(6): 745-752.
ZHANG J P, LI H B, YANG J, *et al*. An importance-sorting algorithm of network community nodes based on topological potential[J]. Journal of Harbin Engineering University, 2012,33(6):745-752.

[12] 张健沛, 李泓波, 杨静等. 基于归属不确定性的变规模网络重叠社区识别[J]. 电子学报, 2012, 40(12): 2512-2518.
ZHANG J P, LI H B, YNAG J, *et al*. Variable scale network overlapping community identification based on identity uncertainty[J]. Acta Electronica Sinica, 2012, 40(12):2512-2518.

[13] 李泓波, 张健沛, 杨静等. 基于社区节点重要性的社会网络压缩方法[J]. 北京大学学报 (自然科学版), 2013, 49(1):117-125.
LI H B, ZHANG J P, YANG J, *et al*. Social network compression based on the importance of the community nodes[J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2013, 49(1):117-125.

[14] 赵东杰, 王华, 李德毅等. 基于拓扑势熵的维基百科词条编辑演化研究[J]. 科技导报, 2012, 30(4): 71-74.
ZHAO D J, WANG H, LI D Y, *et al*. Article edit evolution in wikipedia based on topology potential entropy[J]. Sciena & Technology Review, 2012, 30(4):71-74.

[15] LI D. Artificial Intelligence with Uncertainty[M]. New York: CRC Press, 2007.

[16] YE Q, ZHU T, HU D Y, *et al*. Cell phone mini challenge award: social network accuracy-exploring temporal communication in mobile call graphs[A]. IEEE VAST 2008[C]. 2008.2007-2008.

**作者简介：**

**段松青**（1987-），男，湖南郴州人，北京邮电大学博士生，主要研究方向为云计算、文本分析、大数据。

**于兴隆**（1989-），男，河北邢台人，北京邮电大学硕士生，主要研究方向为云计算、复杂网络、数据挖掘。

**吴斌**（1969-），男，湖南长沙人，北京邮电大学教授、博士生导师，主要研究方向为云计算、复杂网络、智能信息处理。

**王柏**（1962-），女，吉林省吉林市人，北京邮电大学教授、博士生导师，主要研究方向为通信软件工程、智能信息处理。

（上接第 123 页）

[9] KANOULAS E, DU Y, XIA T, *et al*. Finding fastest paths on a road network with speed patterns[A]. Data Engineering, 2006, ICDE'06, Proceedings of the 22nd International Conference[C]. IEEE, 2006.10.

[10] DU Q, FABER V, GUNZBURGER M. Centroidal Voronoi tessellations: applications and algorithms[J]. SIAM Review, 1999, 41(4): 637-676.

[11] WANG J, CUI C, *et al*. A parallel algorithm for constructing Voronoi diagrams based on point-set adaptive grouping[J]. Concurrency Computat: Pract Exper, 2013, 26: 434-446.

[12] 龙其，叶晨，张亚英. 动态路网中基于实时路况信息的分布式路径生成算法[J]. 计算机科学, 2014, 41(9):259-262,278.
LONG Q, YE C, ZHANG Y Y. Distributed path generation algorithm based on real-time traffic information in dynamic road network[J]. Computer Science, 2014, 41(9):259-262,278.

[13] 张翼，唐国金，陈磊. 时相关车辆路径规划问题的改进 A*算法[J]. 控制工程, 2012, 19(5):750-756.
ZHANG Y, TANG G J, CHEN L. Improved A* algorithm for time - dependent vehicle routing problem[J]. Control Engineering of China, 2012, 19(5):750-756

**作者简介：**

**叶晨**（1980-），男，安徽天长人，同济大学博士生、讲师，主要研究方向为嵌入式计算、智能交通、无线网络等。

**杨振宇**（1988-），男，安徽淮北人，同济大学博士生，主要研究方向为智能交通、控制算法等。

**喻剑**（1975-），男，浙江义乌人，同济大学讲师，主要研究方向为物联网、RFID 及应用。

**龙其**（1987-），男，湖南长沙人，同济大学硕士生，主要研究方向为嵌入式计算、智能交通。