融合高阶信息的社交网络重要节点识别算法

闫光辉，张萌，罗浩，李世魁，刘婷
（兰州交通大学电子与信息工程学院，甘肃 兰州 730070）

摘 要：识别重要节点是复杂网络研究的基础性问题。现有理论框架主要以“点–边”这种低阶结构为基本单元，往往忽略了多个节点之间可能存在的交互性、传递性等重要因素。为了更加精确地识别重要节点，对网络中以模体为基本单元的高阶结构进行了研究，首先，提出了节点高阶度的概念，进一步引入证据理论融合了节点的高阶结构和低阶结构信息，设计了一种融合节点高阶信息的半局部重要节点识别方法。在 3 个真实社交网络上的实验结果表明，相较于只关注低阶结构的已有方法，所提出的算法能够更加精确地识别网络中的重要节点。

关键词：重要节点；模体；高阶网络；证据理论；社交网络
中图分类号：TP181，TP391
文献标识码：A
doi: 10.11959/j.issn.1000−436x.2019198

Identifying vital nodes algorithm in social networks fusing higher-order information

YAN Guanghui, ZHANG Meng, LUO Hao, LI Shikui, LIU Ting
School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China

Abstract: Identifying vital nodes is a basic problem in complex network research. The existing theoretical framework, mainly considered from the lower-order structure of node-based and edge-based relations often ignores important factors such as interactivity and transitivity between multiple nodes. To identify vital nodes more accurately, the motif, the high-er-order structure of the network, was studied as the basic unit. Firstly, a notion of higher-order degree of nodes in a com-plex network was proposed. Then, the higher-order structure and lower-order structure of nodes were fused into evidence theory. A semi-local identifying vital nodes algorithm fusing higher-order information of nodes was designed. The results of experiments on three real social networks show that the proposed algorithm can identify vital nodes more accurately in the network than the existing methods which only focus on the low-order structure.

Key words: vital node, motif, higher-order network, evidence theory, social network

1 引言

信息技术的多元化发展，使人们日常交流、互动的形式趋于多样化，由此产生了海量的社交网络数据[1]。社交网络数据存在着大量人和人之间的交互信息，在这些数据中挖掘具有影响力的节点能够帮助人们在理解传播模式的基础上，更好地引导或阻止信息的传播。

节点重要性是节点影响力、地位或者其他因素的表现形式之一[2]。节点的重要性评价方法大体可以分为以下 4 类。1) 基于节点近邻的排序方法，这类方法主要关注节点的邻居信息。例如度中心性（DC, degree centrality）[3]，它是根据一个节点的邻居数目来判断该节点的影响力，这种方法能够简单直观地刻画节点的重要性。但是，由于没有考虑网络的全局结构，在多数情况下不够精确[4-6]。2) 基

收稿日期：2019−05−16；修回日期：2019−08−14
基金项目：国家自然科学基金资助项目（No.61662066, No.61163010）；甘肃省青年基金资助项目（No.1606RJYA222）
Foundation Items: The National Natural Science Foundation of China (No.61662066, No.61163010), The Natural Science Foundation for Young Scientists of Gansu Province (No.1606RJYA222)

于路径的排序方法，此类方法中，节点在传递过程中的重要程度决定着节点的重要性[4]。介数中心性（BC, betweenness centrality）[7]和接近中心性（CC, closeness centrality）[8]是基于路径的 2 种经典方法。BC 通过在 2 个节点之间所有的最短路径中计算某个节点的路径占总路径的比例，来刻画该节点在网络中的控制能力[4]。CC 通过一个节点与网络中其他节点的平均距离得到节点的重要性。CC 值越大，说明该节点在网络中对控制信息的流动越有意义[4]。虽然它们通过全局结构更好地识别了影响力更大的节点，但是计算复杂较高，很难应用于大规模网络。3) 基于特征向量的排序方法，主要代表有 PageRank 算法[9]和 LeaderRank 算法[10]。它们通过模拟用户上网浏览网页的过程，使节点的分值沿着访问路径增加，来识别网页重要性[4]。4) 基于节点移除和收缩的排序方法，其特点是网络的结构会随着重要节点排序的过程处于动态变化中，节点的重要性通过该节点被移除之后对网络的破坏程度来体现[4]。

尽管上述文献针对不同情况对网络中节点的重要性进行了度量，但是它们都是以节点和边为基本单元的研究方法。这些方法都忽略了在真实社交网络中存在于节点之间的高阶关系。大量研究表明，社交网络包含着丰富的子图结构，这些子图结构内部具有一定的传递性、交互性、平衡性等特性。人们一般将这种子图结构描述成网络模体或者图元结构[11-14]。同时，相较于以点和边为研究对象的低阶结构，这种以小子图结构为研究单元的网络结构被称为高阶网络结构[15]。自 2002 年模体的概念被提出后，人们的研究大多着眼于如何高效地统计网络中模体的数量；直到 2016 年，Benson 等[15-16]证明模体可以用于图聚类和社团发现，并提出了一系列的理论基础，高阶网络的研究成为当前研究的重点之一。

为了定量地度量社交网络中节点的重要性，本文将模体作为节点之间高阶关系的研究单元，采用高阶网络分析方法得到基于模体的加权邻接矩阵，进而给出高阶度定义，将节点度和高阶度这 2 个独立的信息源融合得到新的基本概率指派，并结合半局部中心性的思想，得到基于高阶结构的重要节点识别方法。最后，在 3 个真实社交网络上检验不同重要节点识别方法得到的节点影响力程度。

2 相关概念

2.1 复杂网络中的高阶结构

模体是高阶网络结构的一种表现形式[17]。将原始网络表示为高阶网络结构过程如下。首先，在网络中选定模体，此时，在指定阶数的模体集合中，选取原始网络中出现次数最多的连接形式。文献[17]指出，针对不同领域，网络中呈现的高阶连接结构各有偏重，结合社会学的相关理论选取三阶模体作为研究对象[18-19]。图 1 列举了三阶模体的所有存在形式。选定模体阶数后，再对所选阶数的模体逐个进行统计，最终选出原始网络中出现次数最多的模体连接形式。本文采用文献[20]提出的模体检测算法来统计网络中模体的数量。

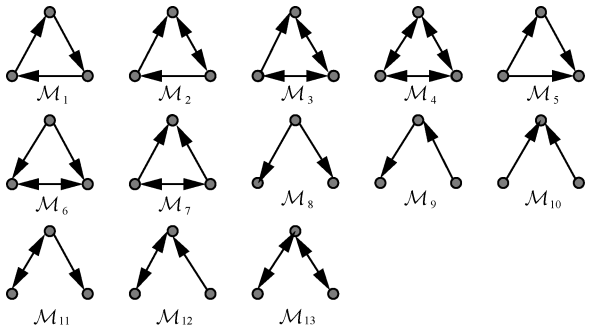


图 1 三阶网络模体的所有存在形式

定义 1 网络模体（network motif）。用一个元组 $(\boldsymbol{B},\mathcal{A})$ 表示由 k 个节点组成的模体。其中，$\boldsymbol{B}$ 是 $k\times k$ 的二进制矩阵，矩阵 $\boldsymbol{B}$ 中的元素代表节点之间是否有连边，$\mathcal{A}$ 是一个目标节点的集合，$\mathcal{A}\subset\{1,2,\cdots,k\}$。一般而言，选取的目标节点是 $\boldsymbol{B}$ 中所有的节点。图 2 表示由 3 个节点组成的模体 $\mathcal{M}_7$。

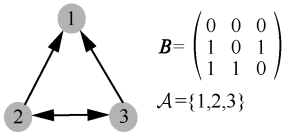


图 2 模体 $\mathcal{M}_7$ 的定义

定义 2 基于模体的邻接矩阵（motif-based adjacency matrix）。给定一个模体 $\mathcal{M}$，模体邻接矩阵可以定义为 $\boldsymbol{W}_{\mathcal{M}}=\{w_{ij}\}_{N\times N}$，矩阵元素 w_{ij} 为在 G 中节点对 (v_i,v_j) 在选定的模体中出现的次数，可定义为

$$w_{ij}=\begin{cases}\sum 1, e_{ij}\in\mathcal{M}且 i\neq j\\ 0, 其他\end{cases} \tag{1}$$

定义 3　高阶网络（higher-order network）。高阶网络可以表示为 $G=(V,E,\boldsymbol{W}_{\mathcal{M}})$，其中，$|V|=\{v_i\mid i=1,2,3,\cdots,n\}$ 表示点集，$|E|=\{e_{ij}\mid i,j=1,2,\cdots,m\}$ 表示弧集，e_{ij} 是一条由节点 v_i 指向节点 v_j 的有向边，$\boldsymbol{W}_{\mathcal{M}}$ 是基于模体 $\mathcal{M}$ 的加权邻接矩阵。

在社交网络中识别重要个体可以转化为识别用户之间关系构成的图中重要节点的问题。同时，文献[17]指出在不同的场景下，模体的主要存在形式各有偏重。本文以社交网络为研究对象，结合社交网络中三元必包的理论，选取三阶模体中 $\mathcal{M}_1\sim\mathcal{M}_7$ 这 7 种存在形式作为基本研究单元。

2.2　D-S 证据理论

D-S 证据理论本质是对概率论的一种推广，将概率论中的基本事件空间拓展到基本事件的幂集空间，以便更好地表达事件的不确定性，并在其上建立基本概率指派函数[21-22]。本节内容给出辨识框架以及它的幂集这些概念，并将其拓展到网络中的节点上[5]。

定义 4　辨识框架（frame of discernment）。$\Theta=\{\theta_1,\theta_2,\cdots,\theta_N\}$ 是由 N 个两两互斥的元素组成的有限完备集合，则称 Θ 为辨识框架。辨识框架是所考察判断的事物或对象的全体集合 Θ。Θ 的幂集 2^{Θ} 所构成的集合表示为

$$2^{\Theta}=\{\varnothing,\theta_1,\theta_2,\cdots,\theta_1\cup\theta_2,\cdots,\theta_1\cup\theta_2\cup\theta_3,\cdots,\Theta\} \tag{2}$$

本文将网络中节点是否重要作为考察判断的对象。因此，节点的辨识框架可以定义为 $\Theta=\{h,l\}$，其中 h和l 分别表示节点重要和不重要这 2 个互斥的元素。

定义 5　基本概率指派（BPA, basic probability assignment）。设 Θ 是辨识框架，Θ 的幂集 2^{Θ} 构成命题集合 2^{Θ}，$\forall A\subseteq\Theta$，若函数 $m:2^{\Theta}\to[0,1]$ 满足以下 2 个条件

$$\begin{cases}m(\varnothing)=0\\ \sum\limits_{A\subseteq\Theta}m(A)=1\end{cases} \tag{3}$$

则称 m 为基本概率指派，$m(A)$ 为命题 A 的基本概率数，即准确分配给 A 的信度。

定义 6　Dempster 组合规则。设 m_1 和 m_2 为两组基本概率指派，对应的焦元分别为 $A_1,A_2,\cdots,A_k$ 和 $B_1,B_2,\cdots,B_l$，用 m 表示 m_1 和 m_2 组合后的新证据。Dempster 组合规则表示为

$$\begin{cases}m(\varnothing)=0\\ m(A)=\dfrac{1}{1-k}\sum\limits_{A_i\cap B_i=A}m_1(A_i)m_2(B_i)\end{cases} \tag{4}$$

其中，$k=\sum\limits_{A_i\cap B_i=\varnothing}m_1(A_i)m_2(B_i)$ 称为冲突系数，用于衡量证据焦元间的冲突程度，k 越大，则冲突越大。当 k=1 时，组合规则无法使用。

2.3　半局部中心性

半局部中心性（SCL, semi-local centrality）是一种基于半局部信息的节点重要性排序方法，它不仅考虑了节点的直接邻居，还考虑了直接邻居的一阶、二阶邻居数，即最多涉及节点的四阶邻居信息[23]。节点 v_i 的计算过程为

$$\mathrm{SLC}(i)=\sum_{j\in\Gamma(i)}\sum_{k\in\Gamma(j)}N(k) \tag{5}$$

其中，$N(k)$ 是节点 v_k 的两层邻居度和，其值等于从 v_k 出发两步内可到达的邻居的数目；$\Gamma(i)$ 是节点 v_i 的一阶邻居节点集合，$\Gamma(j)$ 是节点 v_j 的一阶邻居节点集合。

可见，高阶网络分析方法可以更加具体地表达节点之间的交互关系，在算法复杂度上也有一定的优势。因此，需要考虑如何将高阶信息融合在证据理论中，以达到刻画节点重要性这个模糊概念的目的。

3　基于高阶结构的中心性算法

本节先定义了高阶网络中高阶度的概念，再规定 2 个基本的独立信息源，最后将这 2 个独立信息源用作证据理论的 BPA，融合得到一种高阶证据中心性度量方法。简单来说，该方法不仅考虑了节点的度，还将节点参与网络的程度，即节点的高阶度作为考虑因素。随后，结合半局部中心性的思想对算法进一步优化。相比之前的方法，本文提出的算法不仅考虑了多个节点之间存在的高阶结构，并且结合了半局部中心性的思想。

3.1　高阶度

定义 7　高阶度（higher-order degree）。节点 v_i 在选定的模体中出现的次数，即在模体邻接矩阵中节点 v_i 在第 i 行或第 i 列最大的元素值，记为 HD_i，即

$$\mathrm{HD}_i = \max\left\{w_{ij}\right\} \tag{6}$$

在高阶网络中，模体邻接矩阵刻画的是节点对$\left(v_i, v_j\right)$的局部连接密度，其权值越高表示以节点对$\left(v_i, v_j\right)$为边的模体结构越多，说明该节点对的抗攻击性越差，重要性也就越高。因此，高阶度反映的是节点所在边参与网络的程度。

3.2 定义 2 个基本概率指派函数

本文将节点的度和高阶度看作 2 种重要性指标，由不同的独立信息源可以得到关于度和高阶度的 2 个基本概率指派函数。此时，节点v_i的 2 个基本概率分布分别为

$$\begin{aligned} &m_{d_i} : m_{d_i}(h), m_{d_i}(l), m_{d_i}(\theta) \\ &m_{\mathrm{HD}_i} : m_{\mathrm{HD}_i}(h), m_{\mathrm{HD}_i}(l), m_{\mathrm{HD}_i}(\theta) \end{aligned} \tag{7}$$

其中，$m_{d_i}(\theta) = 1 - \left(m_{d_i}(h) + m_{d_i}(l)\right)$，$m_{\mathrm{HD}_i}(\theta) = 1 - \left(m_{\mathrm{HD}_i}(h) + m_{\mathrm{HD}_i}(l)\right)$表示在这 2 种指标下，不确定节点是否重要的信任程度。

文献[24]用网络中的度分布，对基于度的基本概率指派函数进行改进，得到了一种更符合真实情况的概率指派函数，避免了证据理论计算所得到结果是均匀分布，而真实网络中的节点呈幂律分布这一冲突。因此，基于度和高阶度的 BPA 通过式(8)~式(11)计算得到。

$$m_{d_i}(h) = \lambda_i \frac{\left|k_i - k_m\right|}{\sigma} \tag{8}$$

$$m_{d_i}(l) = (1 - \lambda_i) \frac{\left|k_i - k_M\right|}{\sigma} \tag{9}$$

$$m_{\mathrm{HD}_i}(h) = \frac{\left|\mathrm{HD}_i - \mathrm{HD}_m\right|}{\delta} \tag{10}$$

$$m_{\mathrm{HD}_i}(l) = \frac{\left|\mathrm{HD}_i - \mathrm{HD}_M\right|}{\delta} \tag{11}$$

其中，σ和δ分别通过式(12)和式(13)计算得到。

$$\sigma = k_M + \mu - (k_m - \mu) \tag{12}$$

$$\delta = \mathrm{HD}_M + \epsilon - (\mathrm{HD}_m - \epsilon) \tag{13}$$

此时，$0 < \mu, \epsilon \leqslant 1$。文献[5]指出$\mu$和$\epsilon$的取值对节点的排序没有影响。

引入 Dempster 证据合成规则，将节点的度和高阶度融合在一起形成一个新的指标 $M(i)$。

$$M(i) = \left(m_i(h), m_i(l), m_i(\theta)\right) \tag{14}$$

通过式(4)分别计算节点重要的基本概率指派$m_i(h)$、节点不重要的基本概率指派$m_i(l)$和不确定节点是否重要的基本概率指派$m_i(\theta)$。通常情况下，为了计算方便将$m_i(\theta)$的值平分给$m_i(h)$和$m_i(l)$，得到$M_i(h)$和$M_i(l)$。此时，$M_i(h)$和$M_i(l)$分别代表节点重要和节点不重要的信任程度。对于一个节点来说，$M_i(h)$越大，同时$M_i(l)$越小，节点的重要性就越大[5]。

3.3 高阶证据中心性

在上述定义的基础上，提出节点的高阶证据中心性（HEC, higher-order evidence centrality）方法。HEC 通过节点的重要程度和不重要程度的差值来表示

$$\mathrm{HEC}_i^{'} = M_i(h) - M_i(l) = m_i(h) - m_i(l) \tag{15}$$

为了确保数值为正，对式(15)进行归一化处理，即

$$\mathrm{HEC}_i = \frac{\left|\min\left(\mathrm{HEC}^{'}\right)\right| + \mathrm{HEC}_i^{'}}{\sum_{i=1}^{N}\left\{\left|\min\left(\mathrm{HEC}^{'}\right)\right| + \mathrm{HEC}_i^{'}\right\}} \tag{16}$$

例 1 图 3 所示是一个简单的有向无权图。根据上述定义计算图 3 中每个节点的高阶证据中心性值。选取图 3 中模体存在形式最多的一种，即$\mathcal{M}_7$。表 1 为例 1 中所有节点计算高阶证据中心性所需要的相关指标及高阶证据中心性值。本文取μ和ϵ的值为 0.15。

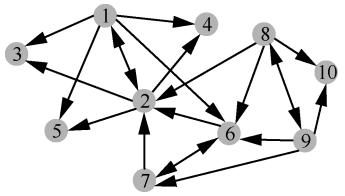


图 3 有向无权图的例子

表 1 高阶证据中心性计算

节点	k_i	HD_i	$m_i(h)$	$m_i(l)$	HEC(i)
1	5	3	0.661 0	0.260 8	0.273 6
2	4	4	0.900 6	0.018 3	0.370 5
3	0	1	0	0.960 5	0
4	0	1	0	0.960 5	0
5	0	1	0	0.960 5	0
6	2	2	0.211 3	0.736 2	0.087 6
7	2	1	0.003 6	0.953 8	0.002 1
8	4	2	0.314 1	0.612 3	0.133 1
9	4	2	0.314 1	0.612 3	0.133 1
10	0	1	0	0.960 5	0

3.4 高阶证据半局部中心性

通过上述计算，得到了网络中每一个节点的高阶证据中心性值。为了更精确地表达节点在网络中的传递性，结合半局部中心性的思想，计算网络中每一个节点的直接邻居和直接邻居的两层邻居 HEC(i) 的和。因此，高阶证据半局部中心性（HESC, higher-order evidence semi-local centrality）为

$$\mathrm{HESC}(i)=\sum_{j\in\Gamma(i)}\sum_{k\in\Gamma(j)}N(k) \tag{17}$$

其中，令 $Q(j)=\sum_{k\in\Gamma(j)}N(k)$，则高阶证据半局部中心性表示为 $\mathrm{HESC}(i)=\sum_{j\in\Gamma(i)}Q(j)$，此时，$N(k)$ 表示节点 v_k 的两层邻居 HEC(i) 的和，$\Gamma(i)$ 表示节点 v_i 的一阶邻居节点集合。

基于高阶结构的 HESC 重要节点识别算法如算法 1 所示。

算法 1　HESC 重要节点识别算法

输入　有向网络 $G=(V,E)$

输出　原始网络中每个节点的 HESC 值

1) 统计 G 中的模体，选取模体数量最多的形式 $\mathcal{M}$；

2) 得到基于 $\mathcal{M}$ 的加权邻接矩阵 $\boldsymbol{W}_{\mathcal{M}}$；

3) for node i in V do:

4) 根据式(6)得到 HD_i；

5) 根据式(7)~式(16)计算 HEC(i)；

6) 根据式(17)计算 HESC(i)；

7) end for

8) return V 中每个节点的 HESC 值

对于给定的 G，网络中节点的个数为 n，边的个数为 m，文献[20]提出统计原始网络中模体个数的算法，该算法的复杂度为 $O(m^{1.5})$。半局部中心性由于计算 $N(k)$ 需要在 2 个步骤中遍历节点 v_k 的邻居，因此，时间复杂度为 $O(n<k>^2)$，即同时考虑了节点直接邻居的一阶和二阶邻居数，其中 $<k>$ 为网络的平均度。证据理论只是在节点本身进行计算，时间复杂度为 $O(n)$。所以，HEC 的时间复杂度为 $O(m^{1.5}+n)$，HESC 的时间复杂度为 $O(m^{1.5}+n<k>^2)$。表 2 给出了本文用到的各算法的时间复杂度。从表 2 可以看出，BC 和 CC 的时间复杂度过高，不适合大规模网络。

表 2　各算法时间复杂度

算法	时间复杂度
DC	$O(n)$
BC	$O(n^3)$
CC	$O(n^3)$
HEC	$O(m^{1.5}+n)$
HESC	$O(m^{1.5}+n<k>^2)$

4 实验

本文采用 SIR 模型对上述提出的基于高阶结构的中心性算法进行评价。文献[4]指出 SIR 模型是一种基于传播动力学的评价方法，被广泛应用到评价各种节点重要性挖掘方法中。在 SIR 模型中，一般通过节点的传播范围和达到稳态时所用的时间作为节点重要性的评判标准。

4.1 SIR 传播模型

SIR 传播模型中的节点在任意时刻都有 3 种可能的存在状态：易感染态（S, susceptible）、感染态（I, infected）和恢复态（R, recover）。在 t 迭代步，这 3 类人在人群中的比例分别用 $S(t)$、$I(t)$ 和 $R(t)$ 表示[25]。$S(t)$ 代表在一个网络中处于 S 状态的节点占比；$I(t)$ 代表处于 I 状态且能向 S 状态节点传播疾病的节点占比，每一个被感染节点都可以通过一定的概率随机地向它的邻居节点传染疾病；$R(t)$ 表示感染过疾病但已经恢复且具有免疫能力的节点占比[25]。在复杂网络 SIR 模型中，假设被感染的节点周围所有的邻居节点都有机会被感染。在第 t 个时间迭代步，感染态和恢复态的节点在整个网络中所占的比例 $F(t)$ 作为传播范围衡量指标。$F(t)$ 随迭代次数 t 的增大而增大，最后趋于稳定。

4.2 数据集描述

Wiki-vote：维基百科为了从用户中选举出词条管理员，通过公开投票的方式选举。该数据集构建的是维基百科用户之间的投票关系网络。网络中的节点代表的是维基百科的用户，有向连边是由投票人指向被选举人。

Advogato：该数据集构建的是 Advogato 这个在线社交平台上用户信任关系的有向网络。网络中的节点是 Advogato 中的用户，有向边为用户之间的信任关系。

soc-Epinions1：该数据集描述的是消费者在评论网站上的信任关系，是否信任对方由网站上

的成员自行决定。由此形成的信任关系网络，可以有针对性地向用户展示消费者的评论。网络中节点代表的是消费者，有向边是消费者之间的信任关系。

表 3 展示了上述提到的 3 个数据集中节点、边数和模体总数的具体情况。

表 3　　数据集的统计量

数据集	节点数	边数	$\mathcal{M}_1 \sim \mathcal{M}_7$ 的总数量
Wiki-vote	7 100	103 700	608 400
Advogato	6 500	51 100	18 300
soc-Epinions1	75 900	508 800	1 600 000

同时，对上述 3 个数据集上 $\mathcal{M}_1 \sim \mathcal{M}_7$ 各个模体的数量进行统计。表 4 列举了 $\mathcal{M}_1 \sim \mathcal{M}_7$ 的具体数量。可以看出，本文选取的数据集中模体出现次数最多的形式恰好都是 $\mathcal{M}_5$。因为 Wiki-Vote 数据集描述的是用户之间的投票关系，$\mathcal{M}_5$ 可以解释为用户更愿意给自己投过票的人所选举的对象再投出一票。Advogato 和 soc-Epinions1 数据集是用户之间的信任关系，$\mathcal{M}_5$ 可以解释为当一个用户选择信任另一个用户时，后者所信任的用户会作为前者信任的参考对象。因此，在这 3 个数据集上选取 $\mathcal{M}_5$ 对原始网络进行处理，得到基于模体 $\mathcal{M}_5$ 的加权邻接矩阵。

表 4　　各个数据集中 $\mathcal{M}_1 \sim \mathcal{M}_7$ 的数量

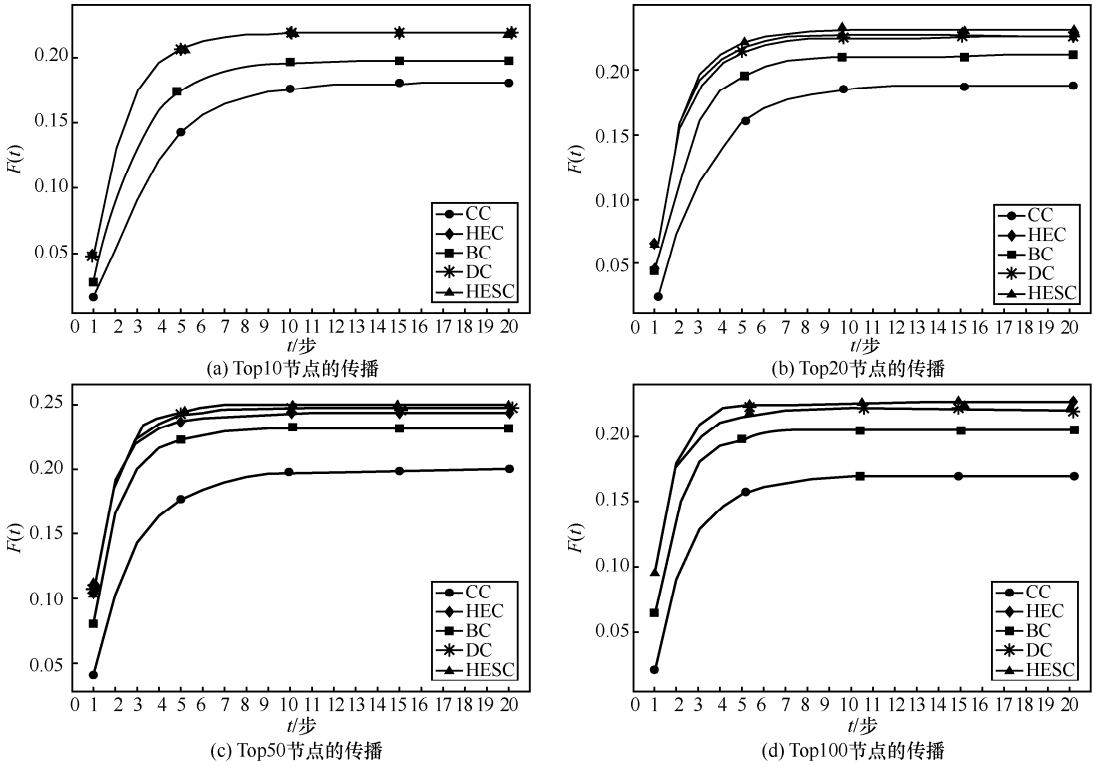
$\mathcal{M}$	Wiki-vote	Advogato	soc-Epinions1
$\mathcal{M}_1$	6 795	63	7 656
$\mathcal{M}_2$	17 667	2 230	84 384
$\mathcal{M}_3$	15 275	3 162	328 076
$\mathcal{M}_4$	2 119	1 992	160 097
$\mathcal{M}_5$	462 715	4 262	531 325
$\mathcal{M}_6$	45 559	3 019	281 093
$\mathcal{M}_7$	58 259	3 564	231 850

4.3　实验方法介绍

为了比较各排序算法之间的差异性和传播能力，本文选取了经典排序算法 DC、BC 和 CC 与本文提出的算法进行比较。首先得到各个算法的节点排序，选取各个算法在不同网络上节点排序的 Top10、Top20、Top50 和 Top100，然后将这些 Top*N* 节点作为初始传染源在 SIR 模型上进行传播，最后比较传播范围及达到稳态时所用的时间。

4.4　实验结果分析

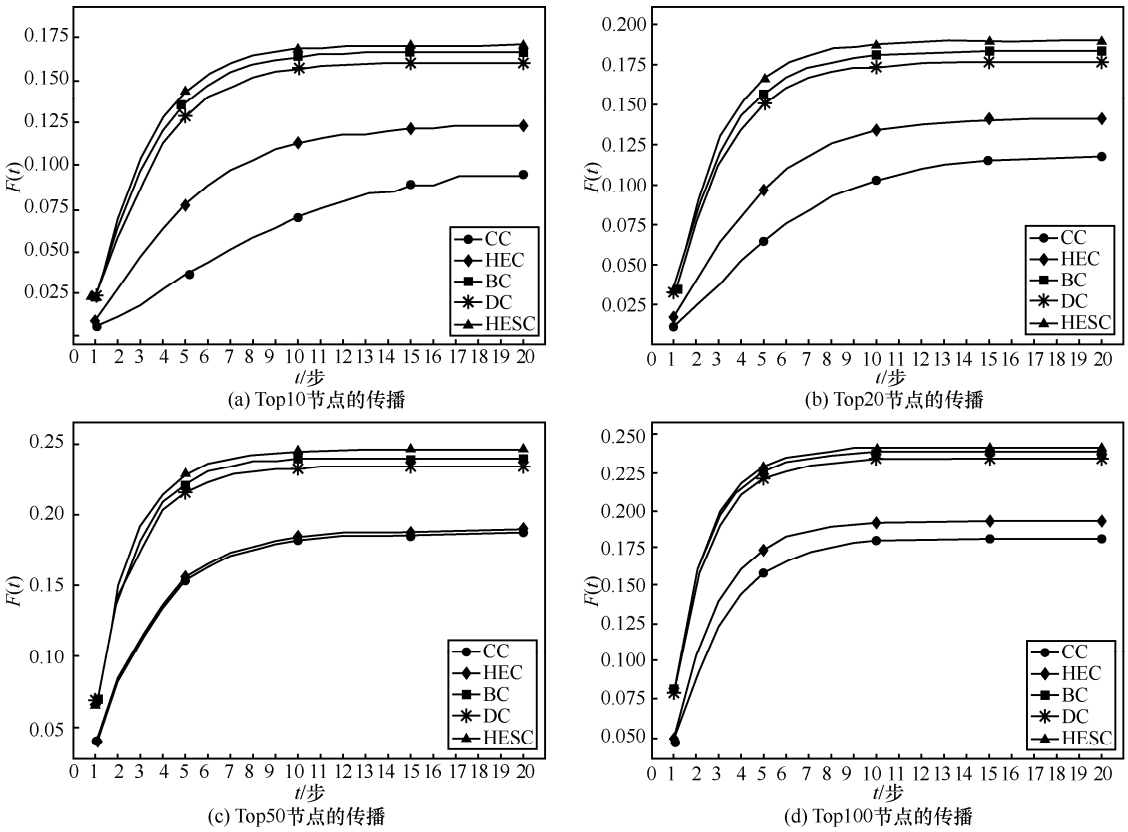
将 3 个数据集上各方法排序中的 Top10、Top20、Top50 和 Top100 作为 SIR 模型中的传染源进行传播，并比较各方法传播的差异性，如图 4~图 6 所示。



(a) Top10节点的传播　(b) Top20节点的传播

(c) Top50节点的传播　(d) Top100节点的传播

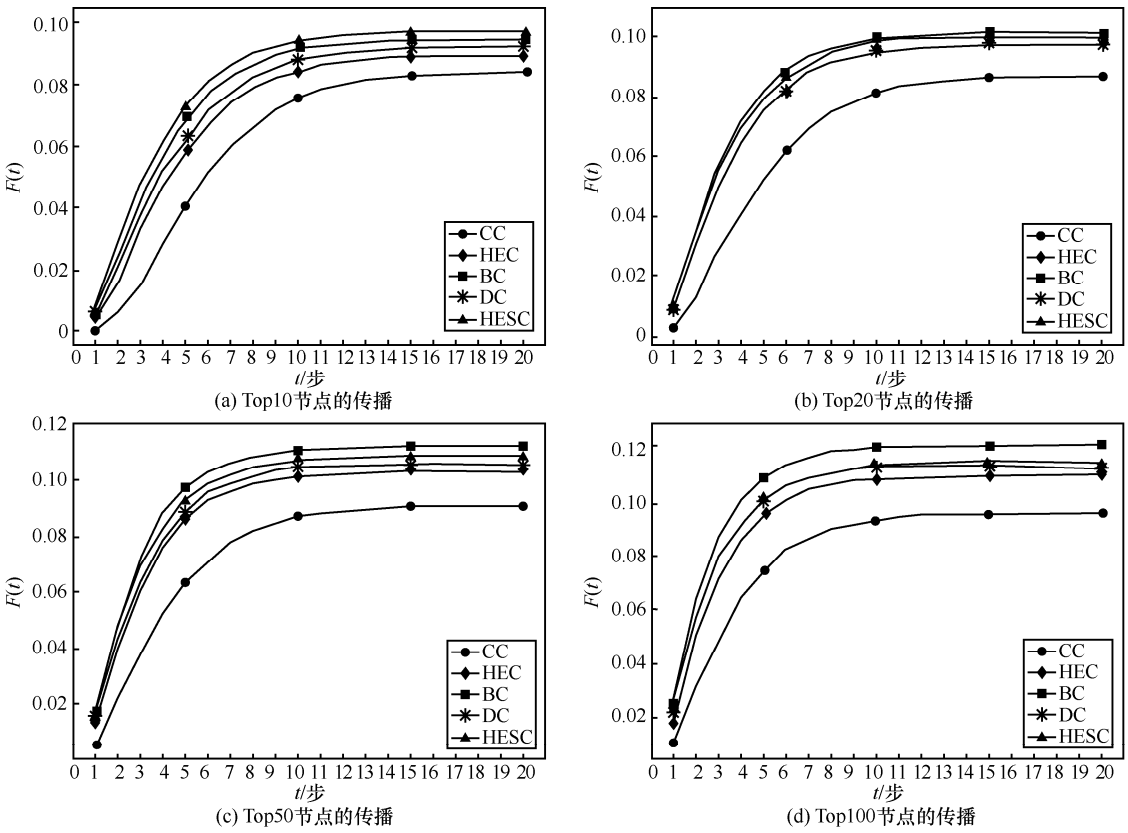
图 4　Wiki-vote 数据集 Top*N* 节点的传播



(a) Top10节点的传播　(b) Top20节点的传播

(c) Top50节点的传播　(d) Top100节点的传播

图 5　Advogato 数据集 Top*N* 节点的传播



(a) Top10节点的传播　(b) Top20节点的传播

(c) Top50节点的传播　(d) Top100节点的传播

图 6　soc-Epinions1 数据集 Top*N* 节点的传播

图 4 是 Wiki-vote 数据集上各种方法在 SIR 模型上的传播。Top10 节点作为传染源进行传播时，本文提出的中心性方法 HEC 和 HESC 在传播范围和传播速率上基本与 DC 相似，如图 4(a)所示。但是，从图 4(b)~图 4(d)可以看出，随着传染源节点的增加，HESC 不管是在传播速率还是传播范围上都好于传统的度量方法，并且网络达到稳态时所用的迭代步最少。

Advogato 数据集上将各种方法得到的节点排序在 SIR 模型上传播，可以看出，本文提出的 HESC 不管是在传播范围还是传播速率都优于经典的方法，如图 5 所示。同时还可以看出，本文提出的 HESC 所得到的排序与 BC 的结果最为相似。而 HEC 的方法传播速率和传播范围却相对较差，可能是因为在此数据集中各个模体之间的数量比较接近（如表 4 所示），此时以网络中个数最多的模体作为分析对象体现不出该方法的优势。

图 6 展示的是 soc-Epinions1 数据集上各种方法通过不同 Top*N* 节点进行传播的差异性。可以看出，Top10 节点作为传染源进行传播时，HESC 的传播范围和传播效率都比其他方法好，如图 6(a)所示。但是，随着传染源节点的增多，HESC 的优势逐渐减退，BC 的传播范围和传播速率优势突显，如图 6(b)~图 6(d)所示。

对上述不同数据集实验结果分析，虽然本文方法 HESC 能够筛选出更重要的节点，但是针对不同的网络存在一定的差异性。在 soc-Epinions1 网络中只有当选取 Top10 节点作为传染源时，传播能力才会好于其他排序结果，其原因可能是该网络中强连通子图连接的节点比较多，而 BC 是基于路径的排序方法。所以，当选取的传染源节点较多时，BC 的优势就显现出来。

图 7 展示的是静态攻击下从网络中移除 Top*N* 节点时，网络中最大连通子图相对大小的变化。横坐标表示从网络中移除 Top*N* 节点，纵坐标表示网络中最大连通子图的相对大小。最大连通子图的相对大小是强连通子图连接的节点个数占总节点数的比例。可以看出，当移除 Top*N* 节点时，BC 得到的节点排序相较于 HESC 对网络的破坏程度更大，所以才会出现图 6 中随着传染源节点的增加，BC 的传播能力好于 HESC 的现象。

表 5 列举的是在 soc-Epinions1 网络上，各种方法得到的 Top10 节点。本文通过 Top1 节点在原始网络中的自我中心网络展示其在网络中的重要程度，如图 8 所示。自我中心网络是指以一个节点为中心，由其直接相连的节点和这些节点之间的连边组成的网络结构[26]。图 8 中处于中间位置直径较大的节点是各种方法得到的 Top1 节点，即 v_{18}、v_{44} 和 v_{645}，周围直径较小的节点是 Top1 节点的出度所相连的节点。显然，本文提出的中心性方法得到的节点连接了更多的节点，即在网络中所处的位置更为关键。同时，在 soc-Epinions1 数据集上，本文提出的方法与 BC 方法更为接近的现象也得以说明。

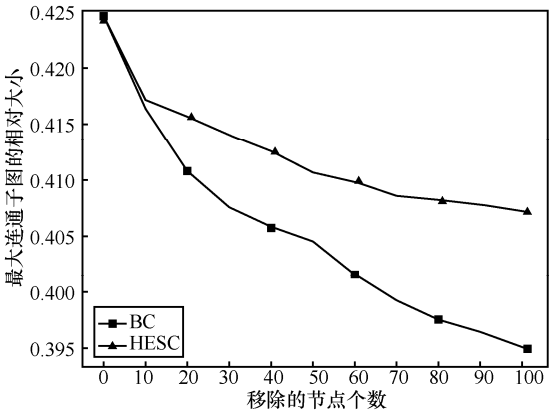


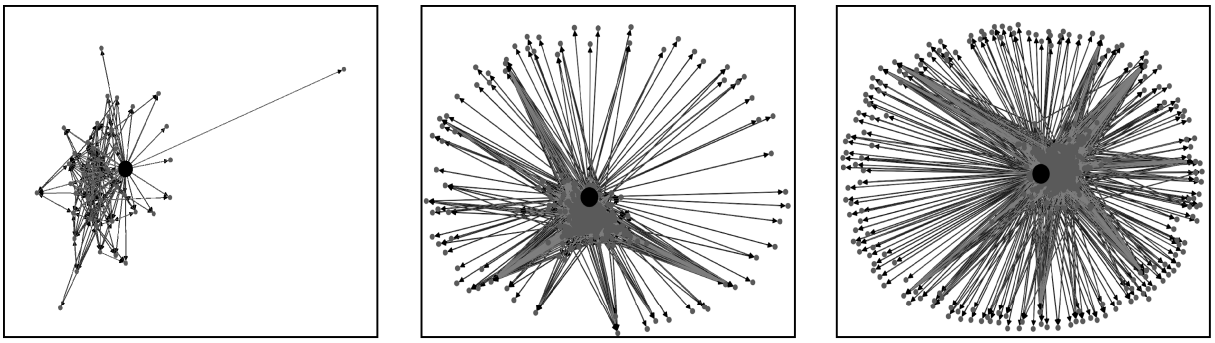
图 7 静态攻击下的最大连通子图相对大小

表 5 soc-Epinions1 数据集 Top10 节点

Top*N*	DC	CC	BC	HEC	HESC
Top1	18	18	44	18	645
Top2	645	737	763	645	634
Top3	634	136	634	634	44
Top4	763	790	2 066	143	71 399
Top5	143	143	645	790	763
Top6	737	1 719	1 409	44	637
Top7	44	118	546	136	34
Top8	790	128	2 969	1 719	145
Top9	34	40	737	737	1 059
Top10	136	1 179	2 118	1 179	824

5 结束语

高阶网络作为一种中尺度的网络结构，相较于从宏观和微观层面入手的研究，高阶网络考虑了节点之间的交互性、传递性等因素。因此，可以更精确地描述网络内部连接的特定模式，同时简化了网络结构。本文主要基于高阶结构提出了重要节点识别算法。该方法以 D-S 证据理论为理论基础，融合了高阶网络分析方法，同时，考虑了网络的度分布、



(a) v_{18}的自我中心网络　　(b) v_{44}的自我中心网络　　(c) v_{645}的自我中心网络

图 8　soc-Epinions1 的各种方法 Top1 节点的自我中心网络

半局部中心性。在 3 个真实社交网络数据集上，用 SIR 模型评价该方法得到的节点重要程度，并与传统的中心性度量方法（DC、BC 和 CC）进行对比。通过实验结果分析，本文所提的重要节点识别算法与其他典型的算法相比较，可以更加精确地识别网络中的重要节点。

本文的工作虽然在有向网络上结合了高阶网络分析方法，得到了节点重要性排序，但仍然缺乏对有向网络上节点重要性因素的综合考虑。因此，进一步的工作可以考虑如何将交互行为在时间上的差异性作为影响节点重要程度的因素，以及度量网络拓扑结构随时间的变化与节点在网络中影响力的关系。

参考文献：

[1] 张静, 唐杰. 社会影响力分析综述[J]. 中国科学: 信息科学, 2017, 47(8): 967-979.
ZHANG J, TANG J. Survey of social influence analysis and modeling[J]. SCIENTIA SINICA Informationis, 2017, 47(8): 967-979.

[2] 韩忠明, 陈炎, 刘雯, 等. 社会网络节点影响力分析研究[J]. 软件学报, 2017, 28(1): 84-104.
HAN Z M, CHEN Y, LIU W, et al. Research on node influence analysis in social networks[J]. Journal of Software, 2017, 28(1): 84-104.

[3] BONACICH P F. Factoring and weighting approaches to status scores and clique identification[J]. Journal of Mathematical Sociology, 1972, 2(1): 113-120.

[4] 任晓龙, 吕琳媛. 网络重要节点排序方法综述[J]. 科学通报, 2014, 59(13): 1175-1197.
REN X L, LYU L Y. Review of ranking nodes in complex networks. Chinese Science Bulletin, 2014, 59(13): 1175-1197.

[5] WEI D, DENG X, ZHANG X, et al. Identifying influential nodes in weighted networks based on evidence theory[J]. Physica A: Statistical Mechanics and its Applications. 2013, 392(10): 2564-2575.

[6] MICHAEL J M. Applied network analysis: a method logical introduction[M]. New York: Sage Publications, 1983.

[7] FREEMAN L C. A set of measures of centrality based on betweenness[J]. Sociometry, 1977, 40(1): 35-41.

[8] FREEMAN L C. Centrality in social networks conceptual clarification[J]. Social Networks, 1978, 1(3): 215-239.

[9] BRIN S, PAGE L. The anatomy of a large-scale hypertextual Web search engine[J]. Computer Networks and ISDN Systems, 1998: 107-117.

[10] LYU L Y,ZHANG Y C, YEUNG C H, et al. Leaders in social networks, the delicious case[J]. PLOS ONE, 2011, 6(6): 1-9.

[11] ALON U. Network motifs: theory and experimental approaches[J]. Nature Reviews Genetics, 2007, 8(6): 450-461.

[12] PRŽULJ N, CORNEIL D G, JURISICA I. Modeling interactome: scale-free or geometric?[J]. Bioinformatics, 2004, 20(18): 3508-3515.

[13] MILO R, SHEN-ORR S, ITZKOVITZ S, et al. Network motifs: simple building blocks of complex networks[J]. Science, 2002, 298(5594): 824-827.

[14] SHEN-ORR S, MILO R, MANGAN S, et al. Network motifs in the transcriptional regulation network of escherichia coli[J]. Nature Genetics, 2002, 31(1): 64-68.

[15] BENSON A R, GLEICH D F, LESKOVEC J. Higher-order organization of complex networks[J]. Science, 2016, 353(6295): 163-166.

[16] YIN H, BENSON A R, LESKOVEC J, et al. Local higher-order graph clustering[C]//The 23rd ACM SIGKDD International Conference. 2017: 555-564.

[17] BENSON A R. Tools for higher-order network analysis[D]. Palo Alto: Stanford University, 2018.

[18] DAVIS J, LEINHARDT S. The structure of positive interpersonal relations in small groups[J]. J Berger Sociological Theories in Progress, 1967: 54.

[19] WASSERMAN S. Social network analysis methods and applications[J]. Contemporary Sociology, 1995, 91(435): 219-220.

[20] SCHANK T, WAGNER D. Finding, counting and listing all triangles in large graphs, an experimental study[M]. Heidelberg: Springer Berlin Heidelberg, 2005.

[21] 蒋雯, 邓鑫洋. D-S 证据理论信息建模与应用[M]. 北京: 科学出版

社, 2018.

JIANG W, DENG X Y. Information modeling and application of D-S evidence theory[M]. Beijing: Science Press, 2018.

[22] SHAFER G. A mathematical theory of evidence by glenn shafer[J]. Journal of the American Statistical Association, 1978, 73(363): 677-678.

[23] CHEN D B, LYU L Y, SHANG M S, et al. Identifying influential nodes in complex networks[J]. Physica A: Statistical Mechanics and its Applications, 2012, 391(4): 1777-1787.

[24] GAO C, WEI D J, HU Y, et al. A modified evidential methodology of identifying influential nodes in weighted networks[J]. Physica A: Statistical Mechanics and its Applications, 2013, 392(21): 5490-5500.

[25] 周涛, 傅忠谦, 牛永伟, 等. 复杂网络上传播动力学研究综述[J]. 自然科学进展, 2005, 15(5): 513-518.

ZHOU T, FU Z Q, NIU Y W, et al. A survey of propagation dynamics in complex networks[J]. Progress in Natural Science, 2005, 15(5): 513-518.

[26] 王庆. 自我中心网络的结构建模与研究[D]. 北京: 北京邮电大学, 2017.

WANG Q. On ego network modeling and analysis[D]. Beijing: Beijing University of Posts and Telecommunications, 2017.

[作者简介]

闫光辉（1970− ），男，河南睢县人，博士，兰州交通大学教授、博士生导师，主要研究方向为数据库理论与系统、物联网工程与应用、数据挖掘、复杂网络分析等。

张萌（1995− ），女，山西芮城人，兰州交通大学硕士生，主要研究方向为社交网络分析、数据挖掘等。

罗浩（1988− ），男，山西原平人，兰州交通大学硕士生，主要研究方向为数据挖掘、多关系网络分析等。

李世魁（1994− ），男，甘肃民勤人，兰州交通大学硕士生，主要研究方向为数据挖掘、复杂网络分析等。

刘婷（1993− ），女，甘肃陇西人，兰州交通大学硕士生，主要研究方向为数据挖掘、信息安全等。