

探测和评估复杂网络影响力节点的路径多样性核度中心方法^①

杨 雄^{②*} 黄德才^{③*} 詹秀秀^{***} 张子柯^{***}

(* 浙江工业大学 计算机科学与技术学院、软件学院 杭州 310023)

(** 常州工学院 计算机信息工程学院 常州 213002)

(*** 杭州师范大学 阿里巴巴复杂科学研究中心 杭州 311121)

摘 要 针对当前复杂网络影响力节点探测和评估方法不能精确定位影响力节点、计算复杂等不足,在传统网络 K 核分解方法的基础上引入了路径多样性概念,从信息传播角度进行了研究,提出了一种基于路径多样性信息熵进行影响力节点探测与评估的新的核度中心方法,即路径多样性核度中心 (C_{ncd}) 方法。实验结果显示,相对于其他影响力节点探测与评估方法,如度中心法 (C_D)、介数中心法 (C_B)、接近中心法 (C_C)、 K 核中心法 (K_C) 及核度中心法 (C_{ncd}), C_{ncd} 方法能够更精确地对影响力节点进行定位,并且能更细粒度地对节点影响力进行有效排序。

关键词 节点影响力,度中心,介数中心,接近中心, K 核分解

0 引 言

近年来对以社会网络、计算机网络、生物网络、管理网络为代表的复杂网络科学的研究日益受到重视。挖掘和探测复杂网络系统中的影响力节点已成为研究复杂网络结构和动力学机制的基本方式,如何高效和准确地探测出网络影响力节点已经成为复杂网络研究领域的热点方向之一。网络影响力节点的挖掘和探测技术已经在包括传染病疫情控制^[1]、信息传播^[2]、市场广告投放^[3]、社会网络领袖^[4]、学术网络评级^[5]等领域产生了重要的理论和实际应用价值。然而当前网络影响力节点的探测和挖掘技术存在明显不足,例如高连接度方法和传统的网络 K 核分解方法不能很好地精确定位影响力节点,介数中心方法和接近中心方法虽能够较准确定位,但计算复杂,不适用于大规模复杂网络。针对这种情

况,本文提出了一种基于路径多样性的网络节点影响力探测和评估方法,该方法既考虑了节点核度,同时也考虑了路径多样性对复杂网络节点的影响力,能够更有效地对关键节点进行定位和更细粒度地对节点影响力进行有效的排序。本文将这种方法叫做路径多样性核度中心 (C_{ncd}) 方法。

1 相关研究

传统的网络影响力节点探测评估方法通常采用节点的连接度属性来衡量度中心 (degree centrality, C_D),方法较为简单,计算工作量较小,但忽视了网络的全局结构,片面地关注于节点的局部属性,因此准确度并不高。Freeman^[6] 和 Sabidussi^[7] 提出了非常著名且使用广泛的介数中心 (betweenness centrality, C_B) 方法和接近中心 (closeness centrality, C_C) 方法,这两种方法很好地解决了度中心方法不能处

① 国家自然科学基金(11305043),浙江省自然科学基金(LQ13F030015, LY14A050001),江苏省高校自然科学基金(13KJD520001)和常州市科技计划应用基础研究(CJ20159013)资助项目。
② 男,1980年生,博士生,讲师;研究方向:复杂网络,数据挖掘;E-mail: popobear801116@qq.com
③ 通讯作者,E-mail: hdc@zjut.edu.cn
(收稿日期:2015-08-18)

理的全局问题,具有较好的探测效果,但是由于网络规模的迅速扩大,这两种全局中心探测方法的计算将会非常耗时,因此并不适用于较大规模网络。2012 年 Chen^[8]提出了一种半局部的节点探测方法,该方法不仅考虑了节点的连接度而且还考虑了邻居的连接度属性,这样做的优势在于一定程度上考虑了网络的全局结构又借鉴了度中心计算简单的特点,解决了全局中心计算复杂性的问题。

最近,Kitsak^[9]提出了一种 K 核分解(K -core decomposition)方法,他发现很多具有高连接度和高介数中心的节点往往并不一定能够对网络中的信息传播起到决定作用,恰恰相反,实验发现能够对信息传播起到重要作用的节点通常位于网络核心位置。 K 核分解首先删除图 G 中所有连接度为 1 的节点及他们所有的边,若剩余节点中仍有度值为 1 的点继续将他们删除,将这些删除的节点划入核层为 1 的核中(本文用 kc 表示节点的核层);其次对剩余连接度为 2 的节点重复上面步骤,将它们划入核层为 2 的核中($kc=2$);最后该过程一直循环下去直到所有节点均被删除且分配了相对应的核层数 kc 。

然而 K 核中心(K -core centrality, C_{KC})法也有一定的局限性,比如这种方法经常会将具有不同传播能力的节点划入同一个核层,或者说同一核层中的不同节点往往对信息的传播具有不同的重要性。针对这个问题,2013 年 Basaras^[10]提出了一种融合介数中心和核层分解的综合方法,该方法较好地弥补了传统 K 核分解的不足之处。同年,Zeng^[11]针对传统 K 核中心方法只关注于节点核内连接度而不考虑核外连接度属性的缺点,提出了一种混合度分解的节点中心评估算法,Liu^[12]根据节点与网络最大核层中核心节点的距离来给出待评测节点的影响力,指出与网络核心距离越近的节点影响力越大的观点,但是该方法由于需要计算距离仍然会导致较大的计算复杂性。

值得注意的是,2014 年韩国庆北大学^[13]的研究小组首次提出了一种基于节点核度中心(neighborhood core centrality, C_{nc})的影响力评估方法,该工作给出的节点核度概念可以理解为该节点所有邻居所处的网络核层数之和,核度概念的提出一方面

考虑了节点所处的网络位置,同时也考虑了节点外部邻域的信息。该方法解决了网络边缘连接度大的节点没有获得合理评估的缺陷,由于这种方法将注意力集中在节点的局部属性——连接度,因此具有较好的计算效率。

但是在很多场合中仅用核度的概念并不能很好地解决节点影响力的判断问题。如图 1 所示,在该网络中节点 1、2、3、22 均位于网络中心核层 3 中($kc=3$),而其他节点均位于核层 1 中($kc=1$)。当使用传统的 K 核分解时并不能判断出这 4 个节点之间的影响力有何差异,而使用文献[13]中提出的核度方法,因为节点 1 和 2 的核度等于 12,节点 3 和 22 的核度等于 9,则可以得知节点 1 和 2 的影响力大于节点 3 和 22。从图 1 中可以非常明确地看出,由于节点 1 和 2 外连着核层 1 中所有的节点,因此这两个节点的影响力将远远大于节点 3 和 22。但是核度方法在判断出节点 1 和 2 的影响力大于同核中的节点 3 和 22 后并不能对彼此进行更为细致的影响力判断。由于两者核度相同,节点 1 和 2 会被认为具有同等重要的地位。然而可以发现,虽然节点 1 和节点 2 各自外连着核层 1 中数量等同的 3 个邻居节点和 6 个 2 级邻居节点(节点 1 连着 4~6 和 10~15,节点 2 连着 7~9 和 16~21),但是由于节点 2 将信息传往核层为 1 的 2 级邻居节点的路径均重叠于节点 7,因此节点 7 和节点 2 之间路径的状

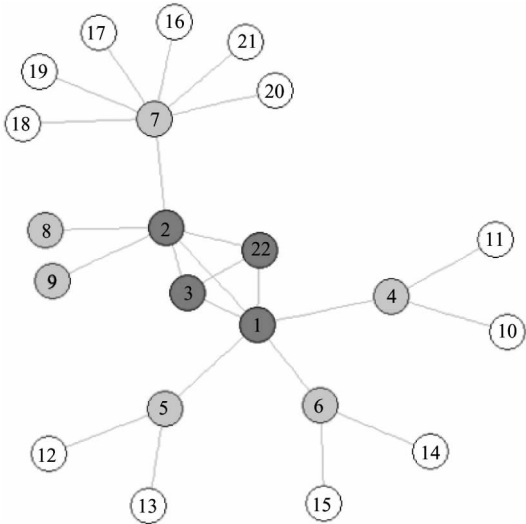


图 1 路径多样性的网络核度示意图

态对于节点 2 来说对整个信息的传递具有非常重要的作用,如果该路径崩溃则对于节点 2 来说能够传递的节点数只有 2 个(节点 8 和 9)。而对于节点 1 而言由于路径多样性、同样将信息传给核层为 1 的节点时,节点 4、5、6 和节点 1 之间同等情况任意单条路径的崩溃却仍然能够保证信息可以到达其余 6 个节点(2 个邻居节点、4 个 2 级邻居节点),即相对于节点 2 而言同等路径损坏条件下节点 1 更容易使得信息传播出去,因此具有更重要的传播影响力。

从上述讨论中可以看到,单纯依靠 K 核分解和核度计算并不总能够对网络节点影响力进行合理的评估,实际上路径的多样性也将对节点影响力评估产生深刻的影响。本文另辟蹊径,在考虑节点核度的同时引入路径多样性概念来研究复杂网络节点的影响力,相对于介数中心方法和接近中心方法需要全局属性、不适合大规模网络的缺点,本文方法具有聚焦于节点的局部属性、计算简单的优点。实验结果显示,路径多样性核度评估方法相对于其他方法更能够准确地评估节点影响力,同时对网络信息传播的高效性也有着重要的作用。

2 节点影响力评估方法描述

传统的网络节点影响力评估大都利用图论中割点、割边等概念对网络拓扑进行研究,而以社会网络为代表的复杂网络对节点影响力评估大都借助其结构属性和信息传播动力学模型进行度量。 $G = (V, E)$ 表示一个无权无向网络图,节点数 $n = |V|$,借助已有研究成果对任意节点 i 的影响力评估指标表示如下:度中心($C_D(i)$)、介数中心($C_B(i)$)^[6]、接近中心($C_C(i)$)^[7]、 K 核中心($C_{KC}(i)$)^[9]、核度中心($C_{nc}(i)$)^[13]。

2.1 路径信息熵的节点影响力计算

目前大多数研究工作都集中于利用局部信息如节点的度进行影响力评估,而本文采用核度中心评估方法并引入路径多样性的概念来提高节点影响力判断的准确性和高效性。

步骤 1 任意节点的路径多样性问题均可通过其邻居节点连接度的不均衡性进行有效表达,本文

采用信息熵^[14]方法对任意节点 i 的路径多样性进行定量描述。

定义 h_i 为节点 i 的路径多样性信息熵, $N(i)$ 为节点 i 的邻居集合, j 为节点 i 的邻居 ($j \in N(i)$), k_j 为节点 j 的连接度, p_j 为节点 j 的度占 i 所有邻居节点度之和的百分比,公式如下:

$$h_i = \sum_{j \in N(i)} -p_j \times \log(p_j)$$

$$p_j = \frac{k_j}{\sum_{w \in N(i)} k_w} \quad (1)$$

h_i 值越大表明节点 i 的邻居节点连接越均衡,也就是节点 i 的信息传播路径越多,在随机传播时感染能力越强。

步骤 2 对式(1)进行标准化处理得到公式

$$H_i = \frac{\sum_{j \in N(i)} -p_j \log(p_j)}{\sum_{j \in N(i)} -\frac{1}{k_i} \log(\frac{1}{k_i})} = \frac{\sum_{j \in N(i)} -p_j \log(p_j)}{\log(k_i)} \quad (2)$$

需要说明的是,对式(1)进行标准化处理是因为即使两个不同连接度的节点的路径多样性都具有相同的均衡性,通过式(1)计算得到的路径多样性信息熵也将会不同。如图 2(a)所示,节点 A 的路径多样性信息熵 $h_A = 2 \times (-\frac{3}{6}) \log(\frac{3}{6}) = \log 2$, 而图 2(b)中节点 B 的路径多样性信息熵 $h_B = 3 \times (-\frac{3}{9}) \log(\frac{3}{9}) = \log 3$, 可以看到无论 A 或 B 的任意邻居节点的度均为 3,因此对于 A 或 B 而言它们的任意邻居节点连接度都具有相同均衡性分布,也就是说节点 A 和节点 B 应该具有相同的路径多样性信息熵 ($H_A = H_B = 1$)。

步骤 3 在传统度中心评估方法基础上考虑邻居节点的度来提升影响力评估的准确性,文献[8,15]提出了节点 i 改进后的度中心评估指标如下式所示:

$$C'_D(i) = k_i + \lambda \frac{\sum_{j \in N(i)} k_j}{k_i} \quad (3)$$

其中 λ 为可调参数。鉴于该思想,我们在核度中心评估方法 C_{nc} 基础上考虑了节点邻域信息,并且

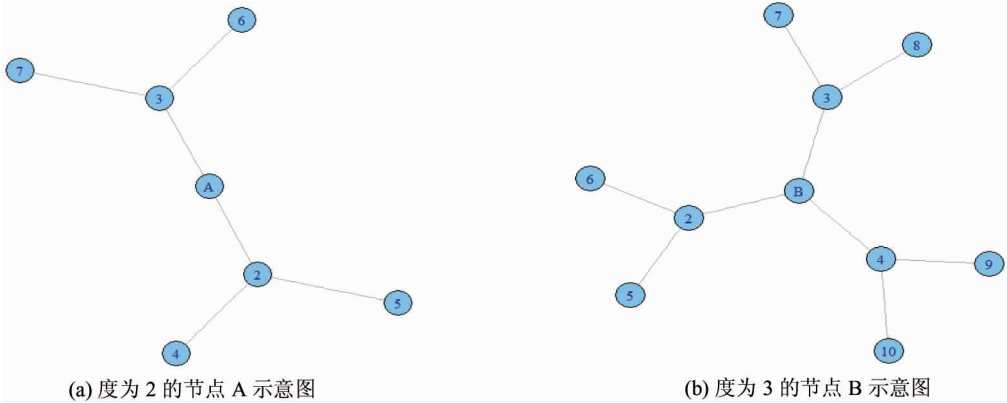


图 2 节点路径多样性信息熵示意图

引入路径多样性参数 H_i , 提出了节点 i 的一种路径多样性核度中心评估方法 $C_{ncd}(i)$ (neighborhood core diversity centrality) 和核度中心评估方法 (C_{nc})^[13] 和路径多样性核度中心评估方法公式如下:

$$C_{nc}(i) = \sum_{j \in N(i)} C_{KC}(j)$$
$$C_{ncd}(i) = \sum_{j \in N(i)} C_{KC}(j) + \lambda H_i \frac{\sum_{j \in N(i)} \sum_{w \in N(j)} C_{KC}(w)}{\sum_{j \in N(i)} C_{KC}(j)} \tag{4}$$

其中 $C_{KC}(i)$ 表示经过 K 核分解后节点 i 所处的核层。

步骤 4 遍历网络中各个节点, 计算它们的路径多样性核度中心评估值 $C_{ncd}(i)$, 并根据该值从大到小排序, 即为节点的影响力排序。

下面以图 1 的小规模样本网络为例, 分别利用各种影响力评估方法 ($C_D(i)$ 、 $C_B(i)$ 、 $C_C(i)$ 、 $C_{KC}(i)$ 、 $C_{nc}(i)$ 及 $C_{ncd}(i)$) 对节点影响力进行评估及排序。表 1 显示了不同影响力评估方法排序的结果 ($\lambda = 2$), 可以看到有些评估方法同一位次上具有多个节点, 因此采用一种优化方法进行细粒度排序划分就显得非常重要; 如利用核度中心方法 C_{nc} 评估在第一位的节点 1 和 2 在改进后的路径多样性核度中心方法 C_{ncd} 中进行了更细粒度的排序; 同时利用度中心法 C_D 排在第三位序的节点 3、4、5、6 和 22 也在路径多样性核度中心方法中进行了不同顺序的划分。表 2 显示了节点 1-7 和节点 22 这 8 个重要节点在不同影响力评估方法中的计算值。

表 1 不同影响力中心评估方法下的节点排序

排序	C_D	C_B	C_C	C_{KC}	C_{nc}	C_{ncd}
1	7	1	1, 2	1, 2, 3, 22	1, 2	1
2	1, 2	2	7	其余节点	3, 7, 22	2
3	3, 4, 5, 6, 22	7	3, 22		4, 5, 6	3, 22
4	其余节点	4, 5, 6	4, 5, 6		8, 9	7
5		其余节点	8, 9		其余节点	4, 5, 6

表 2 不同影响力中心评估方法下的节点数值

节点	C_D	C_B	C_C	C_{KC}	C_{nc}	C_{ncd}
1	6	135	0.0238	3	12	19.316
2	6	123	0.0238	3	12	18.572
3	3	0	0.0196	3	9	16.042
22	3	0	0.0196	3	9	16.042
4	3	39	0.0172	1	5	8.749
5	3	39	0.0172	1	5	8.749
6	3	39	0.0172	1	5	8.749
7	7	105	0.02	1	9	12.266

2.2 节点传播信息的定义

影响力传播节点对信息传播具有促进作用, 这里借鉴经典传染病 SI (susceptible infection) 模型定义信息传播过程。任意节点的状态只有 2 个: 感染 (接收) 状态和未感染 (未接收) 状态。

设 t 表示信息在网络中传播的轮次, $V_1(i, t)$ 表

示在 t 时刻被源点 V_i 感染的节点集合, $V_1(i, 0) = \{V_i\}$ 表示 $t=0$ 初始时刻感染节点只有 V_i , $V_0(i, 0) = \{V_1, V_2, V_3, \dots, V_{i-1}, V_{i+1}, V_{i+2}, \dots, V_n\}$ 表示初始时刻未感染节点集合。规定节点一旦感染将从集合 V_0 中删除并被放入集合 V_1 且 V_1 中节点不可重新回到集合 V_0 。当第 t 轮次(步)传播时, 遍历 $V_1(i, t-1)$ 集合中每个节点及它们的未被感染邻居节点, 令 $\forall V_p \in V_1(i, t-1)$; $V_q = \{N(V_p) \in V_0(i, t-1)\}$, 针对每个 V_p 随机选择 l 个 V_q 中的节点将他们放入 $V_1(i, t)$ 集合, 同时从 $V_0(i, t-1)$ 删除这 l 个节点生成新的 $V_0(i, t)$ 集合。令 $f(V_i, t)$ 表示节点 V_i 在 t 时刻的传播影响力, 也就是信源在经过 t 轮次(步)后能将信息感染传播给网络节点的总数, $f(V_i, t) = |V_1(i, t)|$ 。

2.3 路径信息熵对节点影响力评估的意义

考虑实际情况中节点每次传播信息时并非和所有邻居进行通信^[16], 可以设定每次与邻居通信的范围非线性正比于该节点的度(如社会网络中人们与知心朋友的交流并非随着人际关系拓宽呈现简单的线性关系)。令节点向邻居传播信息的概率为 β , 且 t 时刻已经被感染的节点 i 属于 V_0 集合的邻居个数为 d_i , 节点 i 每次非线性随机与 $\sqrt{d_i}$ 个邻居进行通信, 则该节点在 t 时刻可传播感染节点数为 $l = \beta \times \sqrt{d_i}$ (实验中取整)。

令图 1 深灰色节点表示 $t=0$ 时刻的信源节点, 从表 2 可知节点 1 和 2 具有相同的核度中心指标 ($C_{nc}(1) = C_{nc}(2) = 12$), 即核度中心方法认为这两个节点具有同等的传播影响力, 但是在 $t=2$ 时刻可以发现节点 1 和节点 2 各自对节点传播感染规模却完全不同(包括邻居和 2 级邻居)。拥有较大路径信息熵的节点 1 可以传播的节点个数 $f(V_1, 2) = \sqrt{3}\beta + (\sqrt{6}\beta^2 + \sqrt{2}\beta)$, 而路径信息熵较小的节点 2 可以传播的节点个数 $f(V_2, 2) = \sqrt{3}\beta + (\frac{\beta}{\sqrt{3}} \times \sqrt{6}\beta + \sqrt{2}\beta) = \sqrt{3}\beta + (\sqrt{2}\beta^2 + \sqrt{2}\beta)$ 。很明显, 相对于 2.1 节无法有效区分节点传播影响力的其他评估方法, 基于路径多样性信息熵的方法能够更细致地对节点影响力进行有效区分和排序。

3 实验结果分析

3.1 实验网络拓扑数据

实验过程中使用 2 个真实网络进行结果分析, 分别为 Gleiser^[17] 的爵士音乐家网络 (Jazz) 和 Duch^[18] 的细胞代谢网络 (C. elegans metabolic network), 两个网络基本拓扑特性如表 3 所示。

表 3 两个实验网络的拓扑数据

网络	节点数	边数	最大度数	簇系数	同配数
Jazz	198	2742	100	0.52	0.02
C. elegans	453	2025	237	0.125	-0.226

从表 3 中可以看到 Jazz 网络的节点个数比 C. elegans 网络少了将近一半但边数反而更多, 这个现象说明该网络节点之间的联系较为紧密(簇系数更大), 同配系数相对 C. elegans 网络更大说明 Jazz 网络中同类节点易于互连, 这个现象也符合社交网络的实际规律。

3.2 实验结果分析

从两个角度对实验结果进行分析:(a) 采用 2.2 节所述传播模型比较评估结果与节点实际传播效果之间的关系, 判断不同方法的有效性;(b) 使用 SI 模型和独立级联模型 IC^[19] (independent cascade model) 比较不同方法选取差异节点的信息传播效果。实验平台为 Intel Core i3-2348M、4G 的 Win7 系统, 编程环境为 R 3.1-win。

(a) 排序结果与实际传播影响力的关系

依次遍历网络中每个节点 V_i 作为单独的初始传播源, 在指定时间($t=5$) 计算该节点的实际传播能力 $f(V_i, t=5)$ 。图 3 显示了不同评估方法在 2 个网络中与节点实际传播影响力 f 之间的关系。从图 3(a) ~ (d), (g) ~ (j) 可以看到度中心方法和接近中心方法的临近节点之间传播影响力差异较大, 因此散点图分布发散且不随评估值呈现单调递增趋势; 介数中心方法和 K 核中心方法不能对相同评估值(横坐标值)的节点实际影响力进行有效区分, 而且可以看到随着介数中心值的增加, 一些节点的传

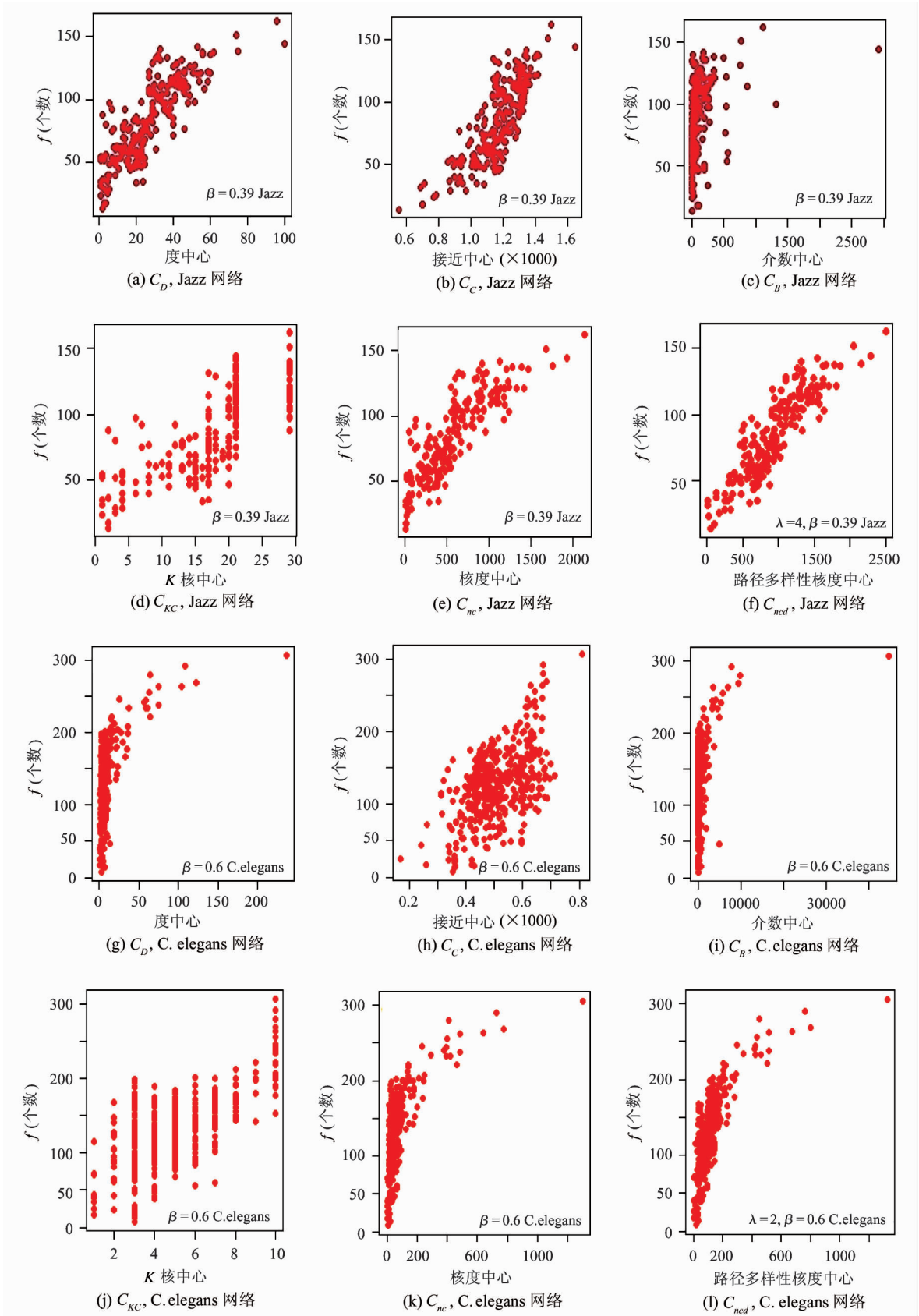


图 3 不同影响力评估方法与节点实际传播能力关系图

播影响力也并非呈现递增趋势,可见传统的度中心、接近中心、介数中心、K 核中心方法均不能对节点影响力进行细粒度区分。图 3(e) 和 (k) 显示了核度中心方法的效果,相对于前述方法,虽然总体呈现单调递增趋势但是仍然有临近节点间传播影响力差异较大的现象,散点图依然较为发散。从图 3(f) 和 (l) 可以看到,本文提出的路径多样性核度中心 (C_{ncd}) 方法呈现单调递增走势,且临近节点间的传播影响力差异较小、散点图较为收拢,因此可以有效对节点影响力进行细粒度排序。

图 4 为 Jazz 和 C. elegans 网络节点的路径多样性核度中心方法与其他几种方法在传播影响力方面

的热度图,横轴分别为节点的度中心、介数中心、K 核中心、核度中心;纵轴为节点的路径多样性核度中心;颜色坐标表示节点的实际传播影响力 f 。从图 4 (a)、(b)、(d)、(e) 可以看到度中心、介数中心、K 核中心与本文方法相关性并不强,例如 jazz 网络度中心和介数中心最大的节点并非实际传播影响力最大的节点,而且两个网络中绝大多数节点都重叠在度中心和介数中心(横坐标)较小的区域,无法有效区分这些节点的影响力;而在 C. elegans 网络的 K 核中心方法中大量不同传播影响力的节点被赋予相同核层,同样无法通过 K 核中心方法进行细粒度区分;从图 4(c)、(f) 可以发现本文方法与核度中心方

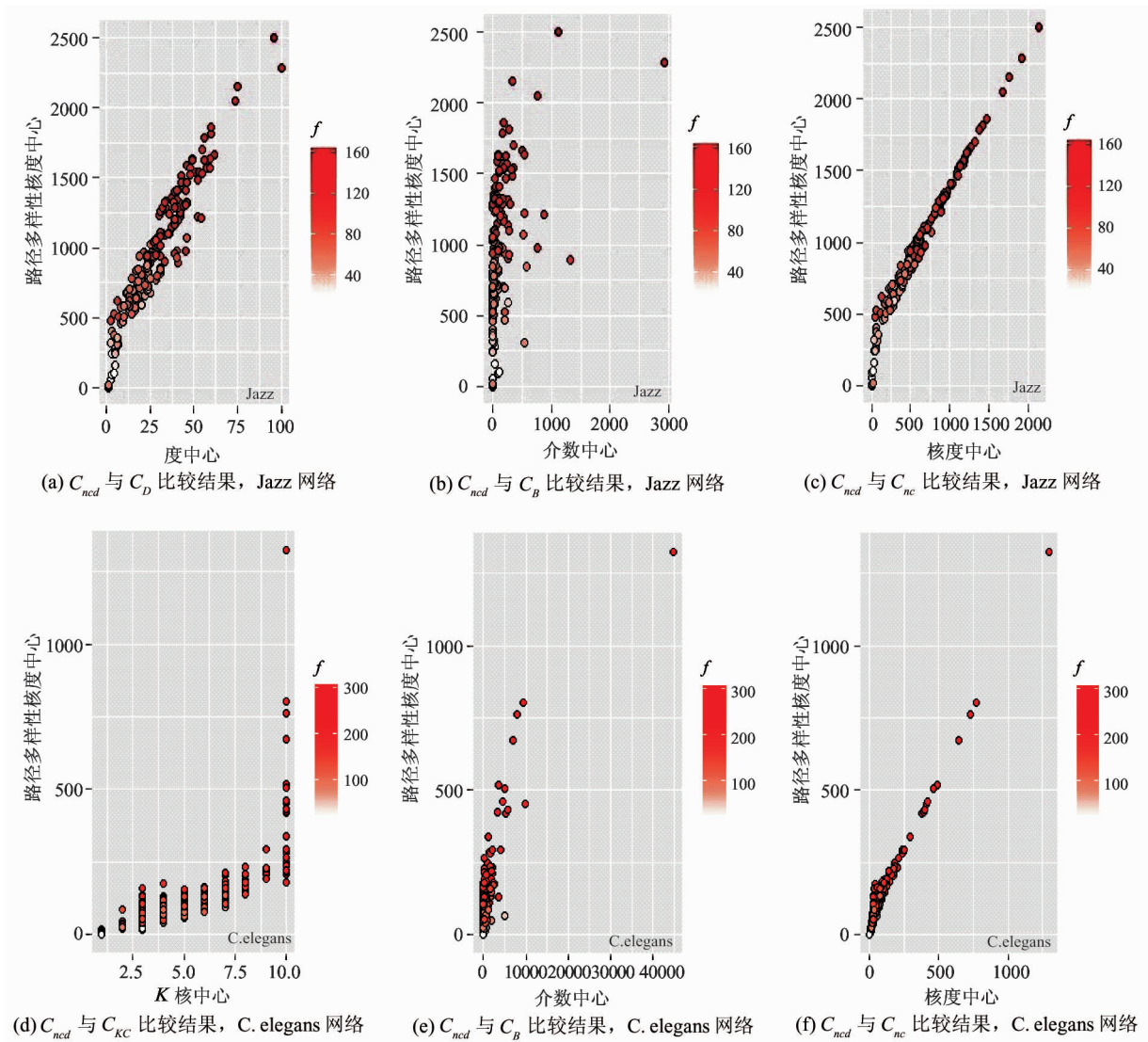


图 4 不同评估方法的传播影响力热度比较图

法相关性较强,随着核度中心值的增加、节点路径多样性核度中心值也会相应增加并且传播影响力也越大,但是在核度中心值(横坐标)较小的区域依旧无法对重叠节点的传播影响力进行有效区分。但从图 4(a)~(f)可以看到数据点在本文方法表示的纵坐标上分布较为均匀且随着纵坐标值的增加,节点传播影响力也越大(颜色越深),因此本文方法能够对网络节点传播影响力进行更细粒度地探测。

为了对不同评估方法的有效性进行分析,这里还采用 Kendall 系数^[20](τ)衡量不同评估排序方法与节点实际影响力(f)排序之间的相关性。 τ 取值最大为 1 表示两种排序方法完全正相关,取值最小为 -1 表示两种排序方法完全负相关, τ 值越大说明评估方法排序的结果越接近于节点实际影响力排序、该评估方法越准确。表 4 显示了不同评估方法与实际传播影响力排序之间的 Kendall 系数,可以看到路径多样性核度中心方法 C_{ncd} 的性能均优于其他方法,并且随着网络节点规模和感染率 β 的增加,评估方法的准确性有所下降,这个现象可以解释为随着感染率 β 的增加,在固定时间有越来越多的节点均达到饱和传播感染状态,从而无法辨别彼此间的实际传播能力,导致排名顺序精确度的下降。

表 4 不同评估方法与实际影响力排名之间的 Kendall 系数

网络	β	$\tau(f, C_D)$	$\tau(f, C_B)$	$\tau(f, C_{nc})$	$\tau(f, C_{ncd})$
Jazz	0.39	0.863	0.786	0.902	0.914
C. elegans	0.6	0.691	0.728	0.761	0.796

(b)评估方法之间的差异分析

为了分析本文提出的路径多样性核度中心方法(C_{ncd})与其他方法对节点影响力评估的差异,分别将本文方法与度中心(C_D)、介数中心(C_B)、核度中心(C_{nc})方法进行了比较(由于 K 核中心(C_{KC})方法将大量不同影响力节点赋予相同值,准确度较差,因此直接排除)。实验将 C_{ncd} 方法分别与其他三种方法进行两两比较,选取非同时出现在两种方法排名前 20 位的节点作为各自的传播源节点(前 20 位中 C_{ncd} 与 C_D 有 3 个不同节点、 C_{ncd} 与 C_B 有 6 个不

同节点、 C_{ncd} 与 C_{nc} 有 1 个不同节点),因为仅仅出现在一种评估方法的节点一定程度反映了该方法对影响力评估的侧重点。为了比较传播效果,实验分别采用独立级联(IC)模型和 SI 模型进行信息传播,实验中将 IC 模型的成功激活概率设置为 0.2,SI 模型传播率 $\beta=0.6, \lambda=2$ 。限于篇幅,这里只给出了规模较大的 C. elegans 网络的实验结果。

从图 5(a)(b)(c)可以看到,相比于其他三种方法,仅被 C_{ncd} 方法发现的节点在独立级联(IC)模型的传播效果均优于其他方法,而图 5(d)(e)(f)中仅被 C_{ncd} 发现的节点在 SI 模型的传播速度均快于其他方法,达到同样传播效果所需的时间更少。综上可知,由于(C_{ncd})方法引入了核度概念,对处于网络边缘但连接度大的节点也给予了正确的定位,同时考虑了传播路径的多样性,因此能够更有效地对节点影响力进行评估。

4 结 论

本文针对当前网络节点影响力评估方法的不足,提出了一种基于路径多样性的节点核度中心评估方法,即路径多样性核度中心(C_{ncd})方法。该方法不同于已有网络 K 核分解粗略定位核层最大的节点作为影响力节点,而是同时考虑了节点在网络中的拓扑位置和局部邻域信息,通过引入路径多样性概念来提高影响力节点的定位效果。实验结果证明相对于度中心(C_D)、介数中心(C_B)、接近中心(C_C)、 K 核中心(C_{KC})、核度中心(C_{nc})等评估方法,本文提出的 C_{ncd} 方法能够更有效地对关键节点进行定位和更细粒度地对节点影响力进行有效排序。

和当前绝大多数研究工作一样,本文主要针对单个节点的影响力进行评估排序,实验过程中也发现挖掘多个最具传播影响力的节点并非简单地定位一组排名靠前的节点,网络结构和社区特性均会对其产生重要的影响,因此如何挖掘网络(特别是带有社区结构的网络)中可使影响力最大化的一组关键节点值得未来进一步研究。

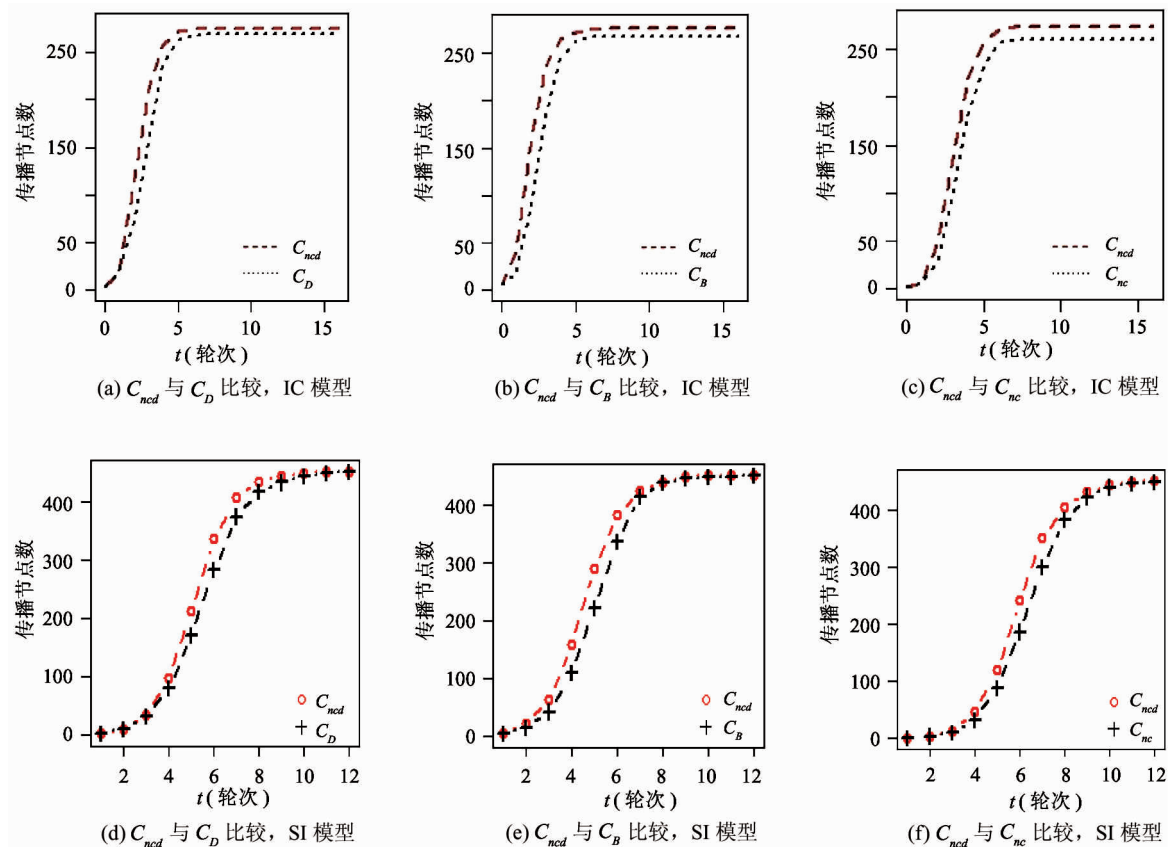


图5 不同时出现在各评估方法前 20 名的节点传播效果差异图

参考文献

- [1] Pastor-Satorras R, Vespignani A. Epidemic spreading in scale-free networks. *Physical Review Letters*, 2001, 86 (14) :3200-3203
- [2] Lü L Y, Chen D B, Zhou T. The small world yields the most effective information spreading. *New Journal of Physical*, 2011, 13(12) :123005
- [3] Yang J, Yao C, Ma W, et al. A study of the spreading scheme for viral marketing based on a complex network model. *Physica A Statistical Mechanics & Its Application*, 2010, 389(4) : 859-870
- [4] Lü L Y, Zhang Y C, Yeung C H, et al. Leaders in social networks, the delicious case. *PLoS One*, 2011, 6(6) : e21202
- [5] Zhou Y B, Lü L Y, Li M. Quantifying the influence of scientists and their publications; distinguishing between prestige and popularity. *New Journal of Physical*, 2012, 14(3) :033033
- [6] Freeman L C. Centrality in social networks conceptual clarification. *Social Networks*, 1979, 1(3) :215-239
- [7] Sabidussi G. The centrality index of a graph. *Psychometrika*, 1966, 31(4) : 581-603
- [8] Chen D B, Lü L Y, Shang M S, et al. Identifying influential nodes in complex networks. *Journal of the American Statistical Association*, 2012, 391(4) :1777-1787
- [9] Kitsak M, Gallos L K, Havlin S, et al. Identification of influential spreaders in complex networks. *Nature Physical*, 2010, 6(11) : 888-893
- [10] Basaras P, Katsaros D, Tassioulas L. Detecting influential spreaders in complex, dynamic networks. *IEEE Computer*, 2013, 46(4) : 24-29
- [11] Zeng A, Zhang C J. Ranking spreaders by decomposing complex networks. *Physical Letters A*, 2013, 377(14) : 1031-1035
- [12] Liu J G, Ren Z M, Guo Q. Ranking the spreading influence in complex networks. *Physica A Statistical Mechanics & Its Application*, 2013, 392(18) : 4154-4159
- [13] Joonhyun B, Sangwook K. Identifying and ranking influential spreaders in complex networks by neighborhood coreness. *Physica A Statistical Mechanics & Its Application*, 2013, 392(18) : 4154-4159

- tion, 2014, 395(4): 549-559
- [14] Shannon C E. A mathematical theory of communication. *Bell System Technical Journal*, 1948, 27(3): 379-423
- [15] Chen D B, Gao H, Lü L Y, et al. Identifying influential nodes in large-scale directed networks: The role of clustering. *PLoS One*, 2013, 8(10): e77455
- [16] Zhang Z K, Zhang C X, Han X P, et al. Emergence of Blind Areas in Information Spreading. *PLoS One*, 2014, 9(4): e95785
- [17] Gleiser P, L Danon. Community structure in Jazz. *Advances in Complex Systems*, 2003, 6(4): 565-573
- [18] Duch J, A Arenas. Community identification using extremal optimization. *Physical Review E*, 2005, 72(2): 027104
- [19] Kempe D, Kleinberg J, Tardos E. Maximizing the spread of influence through a social network. In: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, USA, 2003. 137-146
- [20] Knight W R. A computer method for calculating Kendall's tau with ungrouped data. *Journal of the American Statistical Association*, 1966, 61(314): 436-439

A method for identifying and ranking influential spreading nodes in complex networks based on neighborhood core diversity centrality

Yang Xiong^{* **}, Huang Decai^{*}, Zhan Xiuxiu^{***}, Zhang Zike^{***}

(^{*} School of Computer science and Technology, School of Software, Zhejiang University of Technology, Hangzhou 310023)

(^{**} School of Computer & Information Engineering, Changzhou Institute of Technology, Changzhou 213002)

(^{***} Alibaba Research Center for Complexity Sciences, Hangzhou Normal University, Hangzhou 311121)

Abstract

The study aimed to find an improved technique for identifying and ranking influential nodes in complex networks. In consideration of current techniques' problems of lower node locating accuracy, higher computing complexity, etc., a method to desect and extimate influential spreading nodes based on the information entropy of path diversity, called the neighborhood core diversity centrality (C_{ncd}) method, was put forward by introducing the conception of path diversity into the traditional K -core decomposition method to conduct the study from the perspective of information propagation. The experimental results show that, compared with other methods like degree centrality (C_D), betweenness centrality (C_B), closeness centrality (C_C), K -core centrality (C_{KC}) and neighborhood core centrality (C_{nc}), the proposed C_{ncd} method can identify influential nodes more accurately and rank influential nodes more finely.

Key words: influence of nodes, degree centrality, betweenness centrality, closeness centrality, K -core decomposition