DHC定理在有向含权网络上的推广及应用

范天龙[1]，朱燕燕[1]，吴蕾蕾[1]，任晓龙[2]，吕琳媛[1,3]

(1. 杭州师范大学阿里巴巴复杂科学研究中心 杭州 311121；2. 苏黎世联邦理工学院计算社会学 瑞士 苏黎世 CH-8092；
3. 电子科技大学基础与前沿研究院 成都 610054)

【摘要】节点影响力排序是网络科学研究领域的热点问题，对该问题的研究极具理论意义与应用价值。最近有研究将原本用于衡量科学家科研影响力的H指数，引入到复杂网络中刻画节点的影响力，并发现节点的度、H指数和核数的内在联系，称为DHC定理。本文在原有研究基础上提出了有向含权网络上的H指数，并证明了DHC定理在有向含权网络中仍然成立。在此基础上，本文比较了这些节点中心性指标在含权网络上进行节点排序的准确性和分辨力，并考察了权重因素对排序准确性的影响。最后本文用含权有向网络上的DHC定理深入分析了中国城市间微博转发网络，对中国城市的在线媒体影响力进行排名，并总结了信息在不同城市用户之间的传播模式。

关 键 词 复杂网络； DHC定理； 有向含权H指数； 节点排序； 微博
中图分类号 TP399 **文献标志码** A doi:10.3969/j.issn.1001-0548.2017.05.020

Generalization and Application of DHC Theorem on Directed and Weighted Networks

FAN Tian-long[1], ZHU Yan-yan[1], WU Lei-lei[1], REN Xiao-long[2], and LÜ Lin-yuan[1,3]

(1. Alibaba Research Center for Complexity Sciences, Hangzhou Normal University Hangzhou 311121;
2. Computational Social Science, ETH Zurich Zurich Switzerland CH-8092;
3. Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China Chengdu 610054)

Abstract Identifying influential nodes is a hotspot issue in the research area of network science, and is of great significance in both theory and application. In recent studies, the H-index which was used to estimate the scientific impact of scientists was applied to identify influential nodes on complex networks. A fundamental relation, called the DHC theorem, was found among node degree, H-index, and coreness. In this paper, we extend the DHC theorem to directed and weighted networks, and proof that the DHC theorem is also valid. Then on this basis, this paper compares the ranking accuracy and resolution of the centralities in real weighted networks, and explores the role of weight ranking accuracy. Finally, by using the directed and weighted DHC theorem, the paper deeply analyzes the China Microblog retweet network between cities, assesses the online media influence of Chinese cities, and summarizes the information propagation patterns between users in different cities.

Keywords complex networks; DHC theorem; directed and weighted H-index; ranking nodes; Weibo

无论是自然界还是人类社会，复杂系统中的大量实体及它们之间的关系都可以用网络进行刻画：网络中的节点代表实体，连边代表实体间的关系[1]。近年来飞速发展的网络科学理论及方法让我们能够更加深入地理解复杂系统的结构特征与演化机理，并对一些复杂系统中的传统问题提出基于网络科学视角的解决办法。常用网络科学的理论及方法进行分析的网络包括社交网络[2]、通信网络[3]、交通网络[4]、生物网络[5]，甚至经济网络[6]、文化史网络[7]、宇宙网络[8]等。同时，根据具体系统的不同特性，网络又可以分为简单网络、有向网络、含权网络、含时网络、多层网络、多关系网络等。

复杂网络具有异质性。正如帕累托定律[9]、齐普夫定律[10]、洛特卡定律[11]和赖普斯定律[12]等所阐述的那样，在以上提到的几乎所有类型的复杂网络中，不同节点对于整个网络的结构和功能影响差异巨大。那些能够在较大程度上影响网络的结构与功能的特殊节点被称为重要节点[13]。节点的影响力排序和重要节点挖掘对网络的优化、避灾、控制、同步等方面的研究意义重大。

收稿日期：2017－04－06；修回日期：2017－07－02
基金项目：国家自然科学基金(11622538, 61673150)；浙江省自然科学基金(LR16A050001).
作者简介：范天龙(1992－)，男，主要从事网络信息挖掘、社交网络分析等研究.

时至今日，人们已经提出了多种重要节点挖掘方法[13-16]。其中文献[13]将无权网络上具有代表性的算法分成4类：基于节点近邻的方法，如度中心性(degree centrality)[17-18]，半局部中心性(semi-local centrality)[19]，*k*-壳分解法(*k*-shell decomposition)[20]等；基于路径的方法，如接近中心性(closeness centrality)[21]，介数中心性(betweenness centrality)[22]等；基于特征向量的排序方法，如无向网络上的特征向量中心性(eigenvector centrality)[17, 23]，有向网络上的PageRank算法[24]、LeaderRank算法[25]等；基于节点移除和收缩的方法等[26-28]。虽然其中的一些方法可以推广到含权网络上，但相比于无权网络，含权网络上的重要节点挖掘方法研究较少。因此，本文将重点研究含权(有向)网络上的节点排序算法。

2005年，文献[29]提出了用来评估科学家学术水平的H指数。一个科学家的H指数为 h，表示在其所有论文中，每篇被引用了至少 h 次的论文至多有 h 篇。H指数简洁且意义明确，在国际范围内获得了广泛使用。2009年，文献[30]将H指数拓展到网络中，用来评估网络中节点的重要性。一个节点的H指数等于 h 表示在这个节点的所有邻居中，度大于等于 h 的邻居至多有 h 个。近日，文献[31]发现过去被认为彼此独立的3个节点重要性指标度、H指数和核数存在一个内在的联系。文献[31]通过定义一个算子 $\mathcal{H}$，将度、H指数和核数连接起来，在数学上证明了它们分别对应 $\mathcal{H}$ 序列的初态、中间态和稳态。它们之间的这一关系被称为DHC定理。

相比无向无权网络，(有向)含权网络更加普遍。然而，H指数在(有向)含权网络中的扩展并非易事。本文首次将H指数及其与度和核数的关系拓展到(有向)含权网络中，从数学上证明了在(有向)含权网络中DHC定理依然成立。相比无权H指数，节点的含权H指数更加精细、更加准确地区分和刻画了节点的重要性。本文将这一方法应用于新浪微博数据的分析中。通过构建微博在不同城市之间的转发关系网络，考察微博平台上不同城市的信息传播影响力。分别计算了出向和入向上节点强度、不含权与含权的H指数、不含权与含权的核数，然后对不同指标得到的结果进行比较分析，评估不同指数在刻画有向含权网络上节点传播影响力方面的表现，并探究了权重因素对排序精度的影响。实验结果表明，有向含权的H指数在分辨粒度更加精细、准确率更高的同时，其收敛时间甚至会降低，在实际应用中更具优势。

1　节点的度、H指数和核数之间的关系

本章将先简单回顾无向无权网络中节点的度、H指数和核数之间的关系，而接下来的两小节，本文尝试将度、H指数和核数之间的关系推广到无向含权网络与有向含权网络中。这种推广并不那么显而易见，而是需要一定的特殊处理。

1.1　无向无权网络

文献[31]给出了无向无权网络上节点的度、H指数和核数之间的关系证明，此处做简单回顾。定义作用在有限的整数序列 $(x_1,x_2,\cdots,x_n)$ 上的算子 $\mathcal{H}$，它将返回一个整数 $y=\mathcal{H}(x_1,x_2,\cdots,x_n)>0$，$y$ 表示这个集合中最多存在 y 个不小于 y 的整数。例如，对于一个发表了 n 篇文章的学者来说，$x_1,x_2,\cdots,x_n$ 就是每一篇文章的他引次数，$\mathcal{H}(x_1,x_2,\cdots,x_n)$ 就是该学者的H指数。在一个无向简单网络 $G(V,E)$ 中，V 是节点集合，E 是连边集合，任意节点 i 的度被记为 k_i，它的邻居的度依次被记为 $k_{j_1},k_{j_2},\cdots,k_{j_{k_i}}$，则定义节点 i 的H指数为：

$$h_i=\mathcal{H}(k_{j_1},k_{j_2},\cdots,k_{j_{k_i}}) \tag{1}$$

同时可迭代定义节点 i 的 $n(n>0)$ 阶H指数为 $h_i^{(n)}$：

$$h_i^{(n)}=\mathcal{H}(h_{j_1}^{(n-1)},h_{j_2}^{(n-1)},\cdots,h_{j_{k_i}}^{(n-1)}) \tag{2}$$

式中，节点的度等于其0阶H指数 $h_i^{(0)}=k_i$，节点的H指数等于其1阶H指数，即 $h_i^{(1)}=h_i$。

定理 1　(DHC定理)对于无向无权网络 $G(V,E)$ 上的任一节点 $i\in V$，它的H指数 $h_i^{(0)},h_i^{(1)},h_i^{(2)},\cdots$ 最终收敛到节点 i 的核数 c_i，即：

$$c_i=\lim_{n\to\infty}h_i^n \tag{3}$$

这个定理说明度、H指数和核数，可以通过一个简单的算子 $\mathcal{H}$ 连接起来，而度、H指数和核数是这一算子运算的初态、中间态和稳态。给定一个网络 $G(V,E)$，当任给一个节点 i，都有 $h_i^{(n)}=h_i^{(n+1)}$ 成立时，则称节点 i 的核数 c_i 等于 $h_i^{(n)}$。网络的收敛时间 n_∞ 定义为 $\mathcal{H}$ 算子从度迭代计算到核数所需要的最小的迭代次数，即 n_∞ 是使得 $h_i^{(n_\infty)}=h_i^{\infty}=c_i,\forall i\in V$ 成立的的最小整数。

值得注意的是，定理1给出了一种不同于 k-壳分解的计算节点核数与网络层数的新方法。

定理 2　给定一个无向简单网络 $G(V,E)$，对于任意节点 $j\in V$，定义 $g_j=k_j$，在每一次异步更新迭代过程中，随机选择一个 i 节点更新其 g 值：

$$\mathcal{H}(g_{j_1},g_{j_2},\cdots,g_{j_{k_i}})\to g_i \tag{4}$$

式中，$j_1,j_2,\cdots,j_{k_i}$ 是节点 i 的邻居。如果 $|V|$ 是有限

值，那么这个更新过程将会在迭代有限的次数之后达到一个稳定状态，表示为 $(g_1^{\infty},g_2^{\infty},\cdots,g_{|V|}^{\infty})$，即此时对于任何节点的更新都不会让该节点的 g 值发生变化：

$$g_i^{\infty}=\mathcal{H}(g_{j_1},g_{j_2},\cdots,g_{j_{k_i}}),\forall i\in V \tag{5}$$

并且对于任意节点，都有 $g_i^{\infty}=c_i$。

定理2说明，在异步更新中该方法仍然可以保证收敛，这使得可以通过一种非中心化的、局部化的方法来计算节点核数与网络层数[32]。

1.2 无向含权网络

权的H指数计算中只考虑了节点的度信息，并没有考虑节点之间连边的权重。然而，在实际中很多网络都是含权的，而且连边的权重往往具有重要意义。例如，在社交网络中，边权表示两个人之间关系的亲密程度；在国际贸易网络中，边权可以表示两个国家之间的贸易额等。无权的H指数方法相当于把所有的边权当作1来处理。

为了能够更加真实全面的反映含权网络的信息以及更加准确的评估节点的影响力[33]，将该方法扩展到含权网络上成为必然。定义节点 i 的含权H指数为一个作用在节点 i 的所有邻居节点的含权度(也称节点强度)上的函数，记为 H^w，即：

$$h_i^w=\mathcal{H}^w[(w_{ij_1},s_{j_1}),(w_{ij_2},s_{j_2}),\cdots,(w_{ij_{k_i}},s_{j_{k_i}})] \tag{6}$$

式中，$j_1,j_2,\cdots,j_{k_i}$ 是节点 i 的邻居；k_i 等于节点 i 的度，且二元数对 (w_{ij_*},s_{j_*}) 按照强度 s_{j_*} 的大小降序排列；$\mathcal{H}^w$ 函数返回一个最大实数 x，且 x 满足 $f(x)\geqslant x$，这里：

$$f(x)=\begin{cases}s_{j_1} & 0<x\leqslant w_{ij_1}\\ s_{jr} & \sum\limits_{m=1}^{r-1}w_{ij_m}<x\leqslant\sum\limits_{m=1}^{r}w_{ij_m}\text{且 } r\geqslant 2\end{cases} \tag{7}$$

接下来，分别讨论边权是整数和实数两种情况下的计算方法。

1.2.1 整数边权的 H^w 指数

本部分将以整数边权为例研究带有离散权重的含权网络。在整数边权的情况下，可以将每一条边权为 w 的边分解成 w 条权重均为1的边。设有节点 i，它的度为 k_i，它与邻居 j 的连边权重为 w_{ij}，它的邻居节点分别为 $(j_1,j_2,\cdots,j_{k_i})$，邻居节点的强度分别为 $(s_{j_1},s_{j_2},\cdots,s_{j_{k_i}})$且$s_{j_1}\geqslant s_{j_2}\geqslant\cdots\geqslant s_{j_{k_i}}$，则式(6)也可以表示为：

$$h_i^w=\mathcal{H}\left(\begin{array}{l}\underbrace{s_{j_1},s_{j_1},\cdots,s_{j_1}}_{w_{ij_1}},\underbrace{s_{j_2},s_{j_2},\cdots,s_{j_2}}_{w_{ij_2}},\\ \cdots,\underbrace{s_{j_{k_i}},s_{j_{k_i}},\cdots,s_{j_{k_i}}}_{w_{ij_{k_i}}}\end{array}\right) \tag{8}$$

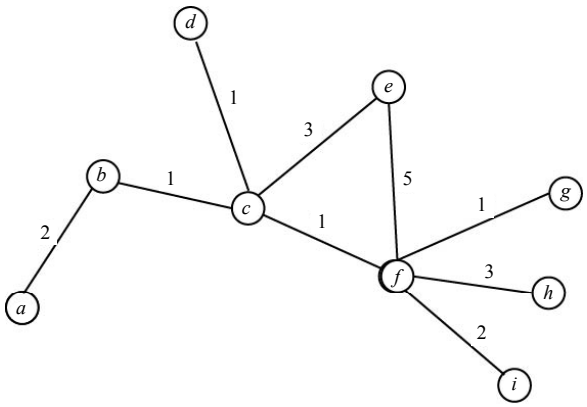


图1 含权网络示例

以图1为例，计算节点f的含权H指数 h_f^w 为：

$$\begin{aligned}h_f^w&=\mathcal{H}^w\left(\underbrace{s_e,s_e,\cdots,s_e}_{w_{ef}},\underbrace{s_c}_{w_{cf}},\underbrace{s_h,s_h,s_h}_{w_{fh}},\underbrace{s_i,s_i}_{w_{fi}},\underbrace{s_g}_{w_{fg}}\right)=\\ &\mathcal{H}^w\left(\underbrace{8,8,8,8,8}_{5\text{个}},\underbrace{6}_{1\text{个}},\underbrace{3,3,3}_{3\text{个}},\underbrace{2,2}_{2\text{个}},\underbrace{1}_{1\text{个}}\right)=6\end{aligned} \tag{9}$$

也可以借助直角坐标系来更清楚地展示含权 H^w 函数的计算过程。仍以图 1 中节点f为例，把它的邻居按照其强度大小做降序排序，并将每个节点重复 w_{ij} 次，如式(9)的序列，然后绘入坐标系中，横坐标为节点，纵坐标为强度，如图 2 所示。此时，这些点与 X 轴和 Y 轴围成的区域中存在一个以原点(0, 0)为顶点的面积最大的正方形，这个正方形的边长 h 就是节点 f 的 H^w 指数值。

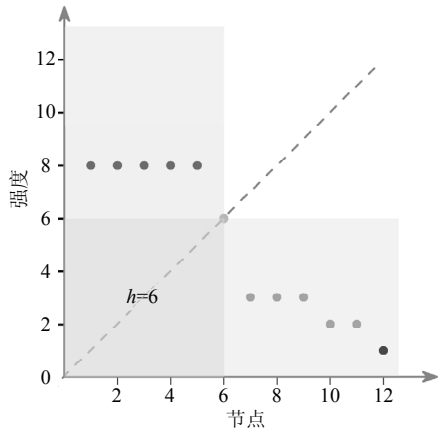


图2 含权 H^w 函数的图像表达

同样地，也可以把含权的 H^w 指数做迭代，类似于式(1)和式(2)，得到含权的 n 阶 H^w 指数：

$$h_i^{w(1)}=\mathcal{H}^w\left(\begin{array}{l}\underbrace{h_{j_1}^{w(0)},h_{j_1}^{w(0)},\cdots,h_{j_1}^{w(0)}}_{w_{ij_1}},\\ \underbrace{h_{j_2}^{w(0)},h_{j_2}^{w(0)},\cdots,h_{j_2}^{w(0)}}_{w_{ij_2}},\\ \cdots,\underbrace{h_{j_{k_i}}^{w(0)},h_{j_{k_i}}^{w(0)},\cdots,h_{j_{k_i}}^{w(0)}}_{w_{ij_{k_i}}}\end{array}\right) \tag{10}$$

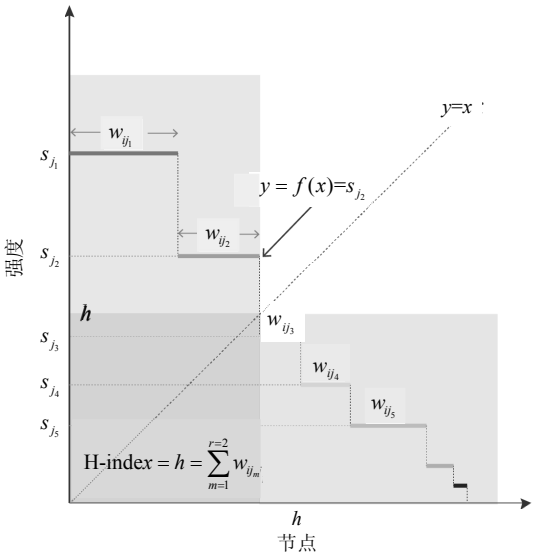
$$h_i^{w(n)}=\mathcal{H}^w\left(\begin{array}{l}\underbrace{h_{j_1}^{w(n-1)},h_{j_1}^{w(n-1)},\cdots,h_{j_1}^{w(n-1)}}_{w_{ij_1}},\\ \underbrace{h_{j_2}^{w(n-1)},h_{j_2}^{w(n-1)},\cdots,h_{j_2}^{w(n-1)}}_{w_{ij_2}},\\ \cdots,\underbrace{h_{j_{k_i}}^{w(n-1)},h_{j_{k_i}}^{w(n-1)},\cdots,h_{j_{k_i}}^{w(n-1)}}_{w_{ij_{k_i}}}\end{array}\right) \tag{11}$$

式中，节点的强度等于其0阶 H^w 指数 $h_i^{w(0)}=s_i$，节点的 H^w 指数等于其1阶 H^w 指数 $h_i^w=h_i^{w(1)}$。

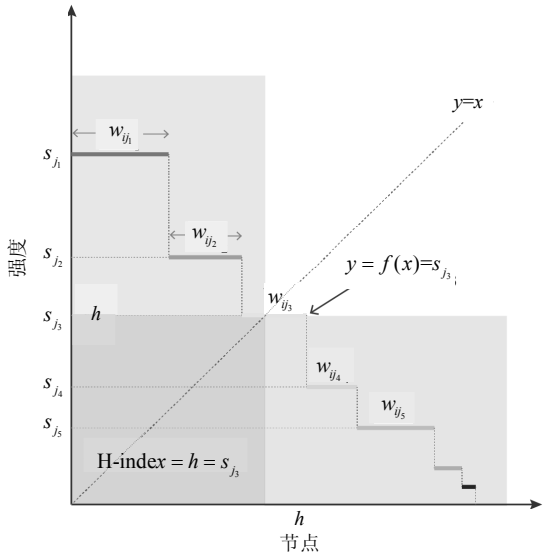
1.2.2　实数边权的 H^w 指数

在实数权重下计算含权 H^w 指数时，式(8)已经不再适用。然而利用坐标轴计算含权 H^w 指数在实数权重的情况下依然适用。根据上述方法，式(6)的计算过程为：对于节点 i (其度为 k_i)，如果 $k_i=1$，则 h_i^w 的值取 w_{ij_1} 与 s_{j_1} 中较小的值；当 $k_i>1$，则寻找一个整数 $k(k<k_i)$，使得式(12)成立：

$$\begin{cases}\sum\limits_{m=1}^{k}w_{ij_m}\leqslant s_{j_k}\\ \sum\limits_{m=1}^{k+1}w_{ij_m}\geqslant s_{j_{k+1}}\end{cases} \tag{12}$$



a. 满足式(12)时的情形1



b. 满足式(12)时的情形2

图3　连续权重下的含权 H^w 函数计算[33]

根据图3所示的两种情况可知，当 $\sum\limits_{m=1}^{k}w_{ij_m}\geqslant s_{j_{k+1}}$ 时，所得结果如图3a所示，即 $h_i^w=\sum\limits_{m=1}^{k}w_{ij_m}$；当 $\sum\limits_{m=1}^{k}w_{ij_m}<s_{j_{k+1}}$ 时，所得结果如图3b所示，即 $h_i^w=s_{j_{k+1}}$。

综上可知，节点 i 的含权 H^w 数为：

$$\begin{cases}\min\left\{w_{ij_1},s_{j_1}\right\} & k_i=1\\ \max\left\{\sum\limits_{m=1}^{k}w_{ij_m},\ s_{j_{k+1}}\right\} & \begin{cases}k_i>1\\ 0<k<k_i\end{cases}\ \text{且}\begin{cases}\sum\limits_{m=1}^{k}w_{ij_m}\leqslant s_{j_k}\\ \sum\limits_{m=1}^{k+1}w_{ij_m}\geqslant s_{j_{k+1}}\end{cases}\\ \sum\limits_{m=1}^{k_i}w_{ij_m} & \sum\limits_{m=1}^{k_i}w_{ij_m}<s_{j_{k_i}}\end{cases} \tag{13}$$

式中，$\min\left\{w_{ij_1},s_{j_1}\right\}$ 表示 h_i^w 取两者中的较小值，$\max\left\{\sum\limits_{m=1}^{k}w_{ij_m},\ s_{j_{k+1}}\right\}$ 表示 h_i^w 取两者中的较大值。

同样，节点 i 在连续权重下的 n 阶含权 H^w 指数就可以表示为：

$$\begin{aligned}h_i^{w(n)}=\mathcal{H}^w[&(w_{ij_1},h_{j_1}^{w(n-1)}),\\ &(w_{ij_2},h_{j_2}^{w(n-1)}),\cdots,(w_{ij_{k_i}},h_{j_{k_i}}^{w(n-1)})]\end{aligned} \tag{14}$$

式中，节点的强度等于其0阶 H^w 指数 $h_i^{w(0)}=s_i$，节点的 H^w 指数等于其1阶 H^w 指数 $h_i^w=h_i^{w(1)}$ (按照式(6)计算)。

1.2.3 含权核数与含权的k-壳(k-shell)分解

含权网络的 *k*-壳分解法[34]定义如下：找到网络中强度最小的节点集合 S，设此时最小强度值为 $s_{\min-1}$，删除网络中所有出现在该集合中的节点及其与邻居的连边；然后在经过上一步操作的网络上，继续找出所有强度小于等于 $s_{\min-1}$ 的节点加入集合 S，然后将它们及其与邻居的连边删除。重复以上操作，直至网络中所有节点的强度都大于 $s_{\min-1}$。此时集合 S 中的节点组成一层，称为原网络含权 $s_{\min-1}$-壳，记为 $k_s^w = s_{\min-1}$。继续上述方法，可以得到网络的含权 $s_{\min-2}$-壳。按照同样的方式反复操作，直到网络中没有节点为止。按照上述定义，原网络中的孤立节点(即度为 0 的点)就属于 0-壳，记为 $k_s^w = 0$。类似的，一个含权网络经过上述的含权 *k*-壳分解后，网络中所有的节点都有一个唯一的 k_s^w 值，并且 k_s^w 小于等于节点的强度，称这一值为节点的含权核数，记为 $c^w = s_{\min-1}$。

同时，把 *k*-壳分解后网络中所有 k_s^w 大于等于 k 的节点称为该网络的含权 *k*-核(*k*-core)。也就是说，网络的含权 *k*-核是网络中所有 k_s^w 大于等于 k 的含权 *k*-壳的并集。在一个边权为整数的连通的含权网络中，含权 1-核就包含了网络的所有节点。经过研究我们发现在含权网络中 DHC 定理依然有效。

定理 3 (含权网络中的DHC定理)对于无向含权网络 $G(V,E,W)$ 上的任一节点 $i \in V$，无论是在同步迭代更新还是异步迭代更新中，它的含权 H^w 指数 $h_i^{w(0)}, h_i^{w(1)}, h_i^{w(2)}, \cdots$ 最终都会收敛到节点 i 的含权核数 c_i^w，即：

$$c_i^w = \lim_{n\to\infty} h_i^{w(n)} \tag{15}$$

定理3证明见补充材料[35]。

1.3 有向含权网络

进一步，本文将该方法扩展到有向含权网络上。对于有向含权网络 $G<V,E,W>$，定义其上任意节点 i 的入向含权H指数(记为 $H^{w_{\text{in}}}$)为一个作用在节点 i 的所有邻居节点的含权入度(此处以入向强度为例)上的函数，即：

$$h_i^{w_{\text{in}}} = \mathcal{H}^{w_{\text{in}}}[(w_{j_1 i}, s_{j_1}^{\text{in}}), (w_{j_2 i}, s_{j_2}^{\text{in}}), \cdots, (w_{j_{k_i^{\text{in}}} i}, s_{j_{k_i^{\text{in}}}}^{\text{in}})] \tag{16}$$

式中，$j_1, j_2, \cdots, j_{k_i^{\text{in}}}$ 是节点 i 的入向邻居，按照节点的入强度 $s_{j_*}^{\text{in}}$ 降序排列；k_i^{in} 为节点的入度，即入向邻居的个数；$w_{j_* i}$ 是从邻居节点 j_* 指向节点 i 的连边的权重。零阶的 $H^{w_{\text{in}}}$ 指数定义为：

$$h_i^{w_{\text{in}}(0)} = s_i^{\text{in}} \tag{17}$$

而其 n 阶 $H^{w_{\text{in}}}$ 指数可迭代定义为：

$$h_i^{w_{\text{in}}(n)} = \mathcal{H}^{w_{\text{in}}}[(w_{j_1 i}, h_{j_1}^{w_{\text{in}}(n-1)}), (w_{j_2 i}, h_{j_2}^{w_{\text{in}}(n-1)}), \cdots, (w_{j_{k_i^{\text{in}}} i}, h_{j_{k_i^{\text{in}}}}^{w_{\text{in}}(n-1)})] \tag{18}$$

同理，得到节点 i 的出向 $H^{w_{\text{in}}}$ 指数：

$$h_i^{w_{\text{out}}(0)} = s_i^{\text{out}} \tag{19}$$

$$h_i^{w_{\text{out}}} = \mathcal{H}^{w_{\text{out}}}[(w_{ij_1}, s_{j_1}^{\text{out}}), (w_{ij_2}, s_{j_2}^{\text{out}}), \cdots, (w_{ij_{k_i^{\text{out}}}}, s_{j_{k_i^{\text{out}}}}^{\text{out}})] \tag{20}$$

$$h_i^{w_{\text{out}}(n)} = \mathcal{H}^{w_{\text{out}}}[(w_{ij_1}, h_{j_1}^{w_{\text{out}}(n-1)}), (w_{ij_2}, h_{j_2}^{w_{\text{out}}(n-1)}), \cdots, (w_{ij_{k_i^{\text{out}}}}, h_{j_{k_i^{\text{out}}}}^{w_{\text{out}}(n-1)})] \tag{21}$$

定理 4 对于有向含权网络 $G<V,E,W>$ 上的任一节点 $i \in V$，无论是在同步迭代更新还是异步迭代更新中，它的入向含权 $H^{w_{\text{in}}(n)}$ 指数序列和出向含权 $H^{w_{\text{out}}(n)}$ 指数序列最终都会分别收敛到节点 i 的入向含权核数 $c_i^{w_{\text{in}}}$ 和出向含权核数 $c_i^{w_{\text{out}}}$，即：

$$c_i^{w_{\text{in}}} = \lim_{n\to\infty} h_i^{w_{\text{in}}(n)} \tag{22}$$

$$c_i^{w_{\text{out}}} = \lim_{n\to\infty} h_i^{w_{\text{out}}(n)} \tag{23}$$

定理4证明类似定理3，不再赘述。

2 含权H指数算法与无权算法的对比分析

为了验证含权H指数的性能，接下来在C. elegans网络[36]、USAir网络[37]和将在本文第3部分提到的中国城市间的微博转发网络(Weibo)，这3个无向含权网络上进行实验，并将结果与无向无权H指数作出比较。C. elegans网络是秀丽隐杆线虫(C. elegans)的新陈代谢网络，节点表示神经元，连边表示它们之间的突触连接；USAir网络是美国航空网络，节点表示机场，连边代表他们之间存在直飞航线，连边权重代表这两个机场的距离(已做归一化处理)。3个网络的基本属性如表1所示。

首先，比较含权H指数与不含权H指数对网络层数的划分情况。实际上，无权H指数与无权 *k*-壳分解作为节点排序算法的时候，一个主要缺陷就是节点的层级划分少，导致很多处于同一层的节点无法区分重要性。如表2所示，H、H^w、c、c^w 分别为无权H指数、含权H指数、无权核数、含权核数4种指标下的网络层数划分情况，n_∞ 与 n_∞^w 分别代表无权H指数与含权H指数的收敛时间。显然，无论是H指数还是核数，含权的方法都可以得到更精细的划分，对于节点的分辨能力更高；同时，含权网络的收敛时间并没有大幅上升。

表1　C.elegans网络、USAir网络与Weibo网络的基本属性

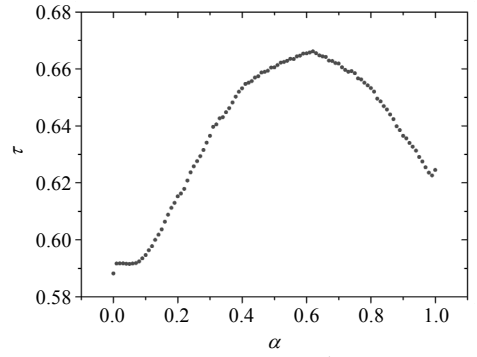
网络	节点数 N	边数 E	平均度<k>	平均强度<s>
C. elegans	297	2 148	14.465	29.626
USAir	332	2 126	12.807	0.462
Weibo	289	34 647	239.8	20 244.3

表2　不同指标的分层情况及网络的收敛时间

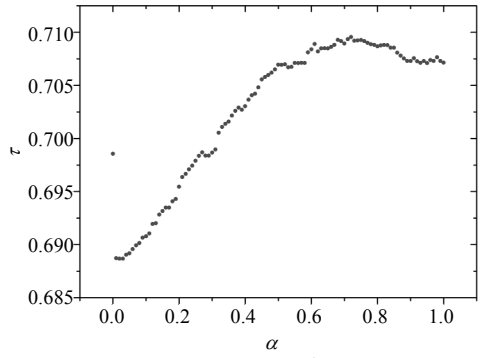
网络	H	H^w	c	c^w	n_∞	n_∞^w
C. elegans	22	80	10	58	16	12
USAir	32	290	23	240	6	12
Weibo	56	286	35	242	12	11

表3　不同指标与WSIR之间的Kendall Tau(τ)相关系数

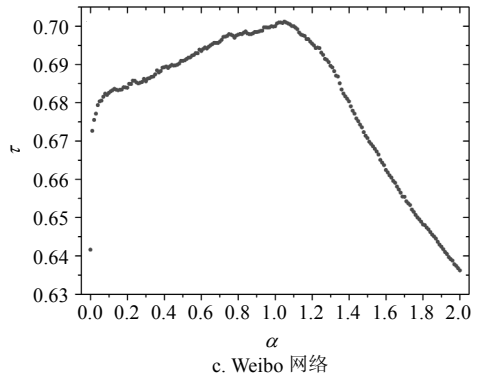
网络	H-WSIR	H^w - WSIR	c - WSIR	c^w - WSIR
C.elegans	0.588	**0.625**	0.549	**0.563**
USAir	0.699	**0.707**	0.690	**0.706**
Weibo	0.642	**0.700**	0.459	**0.700**



a. C.elegans 网络



b. USAir 网络



c. Weibo 网络

图4　使用参数 α 对权重重新分布情况下得到的H指数与WSIR模型所得结果的相关性 τ

接下来，继续分析含权与无权H指数这两种方法的重要节点识别的准确性。分别将含权与无权情况下的H指数和核数与WSIR模型[38]的仿真结果做相关性分析，计算它们之间的Kendall Tau(τ)相关系数[39]，结果如表3所示。结果表明，无论是以H指数，还是节点核数作为计算指标，含权的方法都比不含权的方法的重要节点排序准确性更高。

最后，本文分析权重因素对含权H指数的节点排序结果的准确率的影响。对边权进行重新分配，定义节点i和j的连边的新权重为：

$$w'_{ij} = w_{ij}^{\alpha} \quad \alpha \in [0,1] \tag{24}$$

式中，w_{ij} 是原网络中连边 e_{ij} 的权重。特别地，当α=0时原网络变为不含权的。接下来比较不同 α 时的含权H指数的排序结果与WSIR模型所得结果的Kendall Tau(τ)相关系数。τ 值越大，说明含权H指数方法对节点的排序越准确。实验结果如图4所示。从图中可以看出，随着边权调整因子 α 的变化，3个网络都存在一个峰值：在C.elegans网络上，当 α=0.62时，取到峰值 $\tau_{\max}$=0.666；在USAir网络上，当 α=0.72时，取到峰值 $\tau_{\max}$=0.710；而在城市间微博转发网络上，当 α=1.05 时，取到峰值 $\tau_{\max}$=0.701。由此，可以通过调整 α，进一步提升含权H指数方法的准确性。有意思的是，在原网络对应的不含权网络中(即 α=0 时)，节点排序的准确率都比较低，甚至为最低。这在某种程度上说明了一般情况下，不可以简单地将含权网络映射为同结构的不含权网络，否则将损失很多重要的信息。更进一步地，这也启示我们，很多不含权网络上的指标和方法不能简单地套用到含权网络上。实际存在的更多是含有更多信息的含权网络[40]。

3　含权H指数在微博转发网络分析中的应用

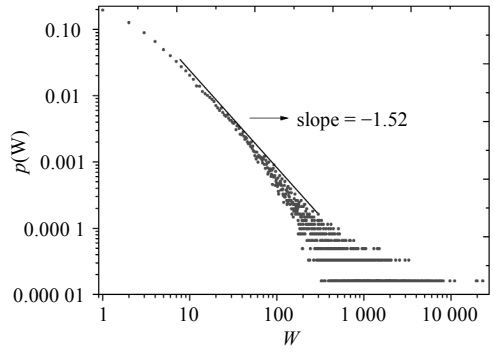
3.1　数据描述

为了研究社交网络上信息流动的空间特性，本文通过爬取新浪微博(www.weibo.com)上的微博转发数据，提取含有准确用户位置(精确到城市)的转发记录，构建了代表不同城市间信息流动的有向含权网络。例如注册城市为city1的用户转发了注册城市为city2的用户m条微博，则在最终生成的网络中表示为：“city2 →city1 m”，节点表示城市，连边方向

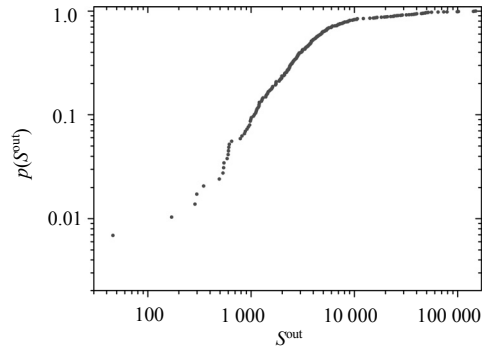
表示信息的流动方向，边权“1”表示位置为city1的用户对位置为city2的用户的微博的转发总次数为1。该有向网络的统计特性如表4所示，其中，L^{out}、C、D和Q分别代表网络的出向平均路径长度、聚类系数、图密度和模块度。图5中分别展示了连边权重的概率分布，节点的出向强度和入向强度的累积概率分布。图6为权重与H指数的阶数对于网络层数的影响。

表4 城市间微博转发网络的基本属性

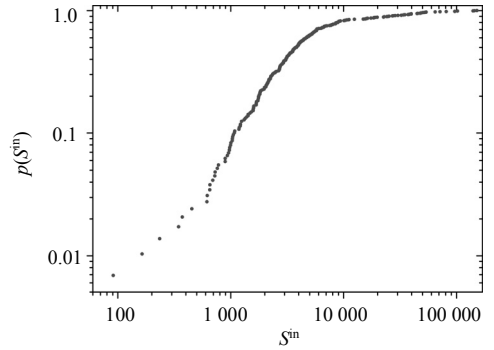
N	E	$<k>$	$<s>$	L^{out}	C	D	Q
289	61 232	212	10 122	1.3	0.8	0.7	0.3



a. 连边权重的概率分布

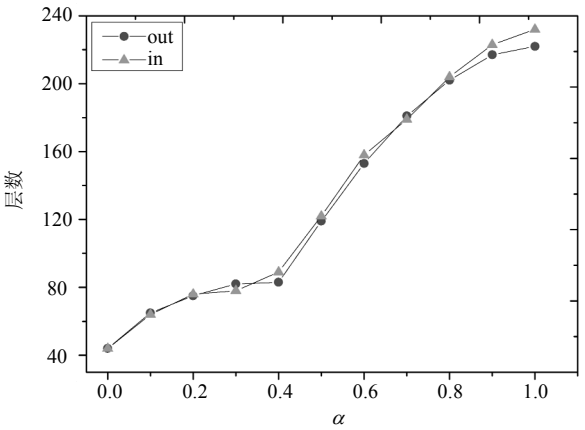


b. 节点出向强度累计概率分布

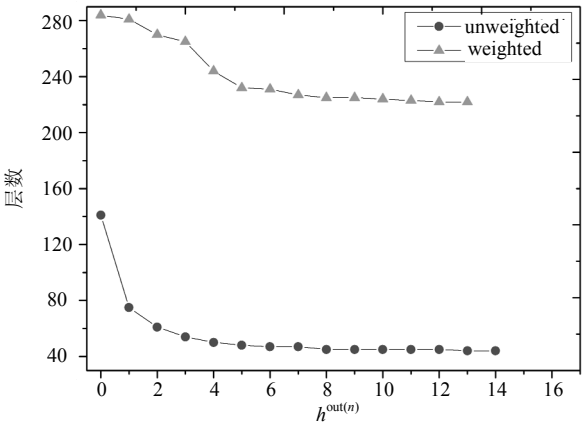


c. 节点入向强度累计概率分布

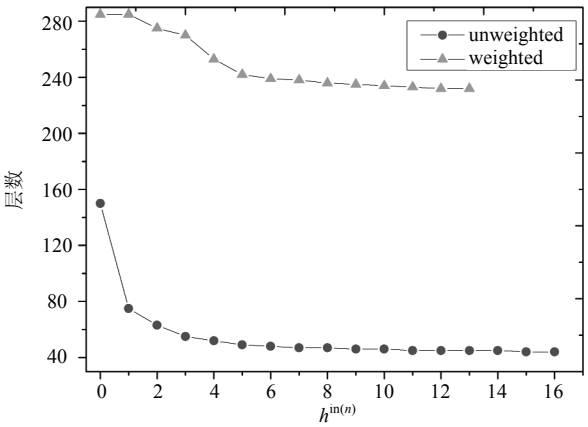
图5 城市间微博转发网络的基本属性



a. 不同边权比例时的网络层数



b. 出向不同阶 H 指数的网络层数



c. 入向不同阶 H 指数的网络层数

图6 权重与H指数的阶数对于网络层数的影响

图7给出了城市间微博转发网络的邻接矩阵热度图。这是一个不对称的邻接矩阵，图中已将相同省份的城市连续编号，而且同一省份的城市按照其出强度大小降序编号。图中已对转发量取自然数e为底的对数，并用颜色深浅表示其大小，行累加与列累加分别为一个城市的出强度与入强度。从图中可以发现省内城市之间的微博互动要比不同省份的城市之间的互动更多一些。此外，省会城市几乎全

排在各自省(或自治区，下同)内的第一位，且图中明显可以看出这些省会城市在省际之间的活跃度要远高于非城会城市。

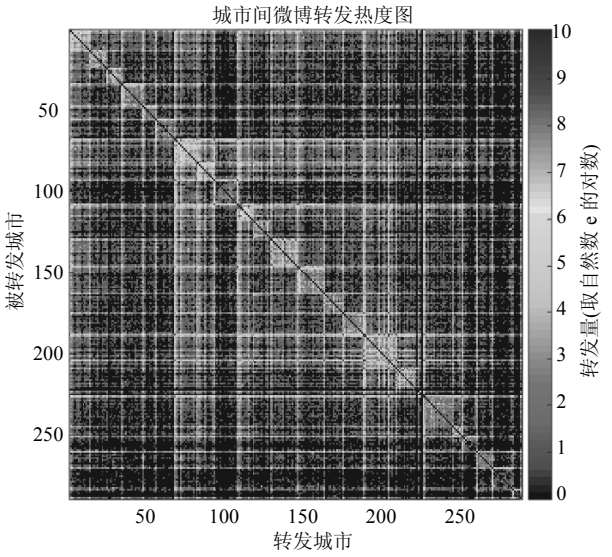


图7　城市间微博转发热度图

3.2　结果分析

本文在城市间微博转发网络上分别计算了节点的出向强度 s^{out}，出向含权H指数 $h^{w_{\text{out}}}$，出向含权核数 $c^{w_{\text{out}}}$，入向强度 s^{in}，入向含权H指数 $h^{w_{\text{in}}}$，入向含权核数 $c^{w_{\text{in}}}$，并和对应的无权指标一一加以对比，发现含权的指标在排序准确率和分辨率上都优于无权的指标。进一步用式(24)的方法对边的权重进行调整时，网络层数的变化情况如图6a所示，横坐标为边权调整因子 α，纵坐标为含权H指数收敛后网络的层数，α 为0时等价于无权网络，为1时等价于原含权网络。此外，本文还考察了 H^w 指数的阶数对网络层数的影响，如图6b和图6c所示，横坐标为 H^w 指数的不同阶数，纵坐标为对应阶数下的网络层数，结果发现含权 H^w 指数在挖掘重要节点的准确性(如图4所示)和分辨率上均优于无权H指数。

本文分别按照节点的强度，含权 H^w 指数，含权核数大小对城市进行排序，如表5所示(以核数为主序列)。网络中相同颜色的城市含权核数相同。进一步，又拿出出向强度排名前40的城市，比较了它们的出向强度、出向H指数、出向核数这3个指标的排名情况，如图8所示。横坐标为出向强度排名，纵坐标分别为出向H指数和出向核数排名，虚线表示两对指标排名相同。从表5和图8中可以发现泉州、厦门、苏州、沈阳、太原、哈尔滨等城市的出向强度排名与出向H指数和出向核数的排名差异较大，前面3个城市的出向H指数和出向核数排名高于出强度，而后3个城市相反。这说明泉州、厦门、苏州等城市虽然用户总体被转发量(即节点出强度)较少，但实际上位于这些城市的用户的微博影响力(出向H指数与出向核数)却要更高一些；而沈阳、太原、哈尔滨等城市情况相反。

表5　不同指标的Top30+城市列表

出向网络				入向网络			
城市	s^{out}	$h^{w_{\text{out}}}$	$c^{w_{\text{out}}}$	城市	s^{in}	$h^{w_{\text{in}}}$	$c^{w_{\text{in}}}$
北京	1	1	1	北京	1	1	1
广州	2	2	1	广州	2	2	1
深圳	3	3	1	深圳	3	3	1
上海	4	4	1	上海	4	4	1
杭州	5	5	5	杭州	5	5	5
阜阳	7	6	5	成都	6	6	5
成都	6	7	5	阜阳	7	6	5
重庆	8	7	5	重庆	9	8	5
郑州	9	7	5	郑州	8	9	5
长沙	10	11	10	长沙	10	10	5
合肥	16	12	10	合肥	16	12	5
东莞	18	13	10	东莞	18	13	5
武汉	15	14	13	武汉	15	14	13
福州	14	10	14	福州	14	11	14
西安	12	15	14	西安	13	15	14
厦门	23	16	14	厦门	23	16	14
青岛	17	17	14	青岛	17	17	14
南京	19	18	14	南宁	11	18	14
南宁	11	19	14	济南	12	19	14
苏州	24	20	14	苏州	24	20	14
泉州	27	21	14	南京	19	21	14
济南	13	22	14	南昌	21	22	14
南昌	20	23	23	佛山	27	23	14
佛山	28	25	23	泉州	29	24	14
昆明	22	24	25	昆明	22	25	25
宁波	32	28	26	天津	26	26	26
天津	26	26	27	宁波	32	28	26
温州	33	29	27	温州	34	29	28
中山	35	30	29	贵阳	28	30	29
贵阳	29	31	30	中山	35	31	30
沈阳	21	27	31	沈阳	20	27	31
兰州	30	34	34	兰州	30	34	35
太原	25	41	41	太原	25	37	40

为了进一步说明以上结果的原因，将转发厦门、苏州、太原、哈尔滨的微博最多的各自前20个邻居城市分别展示在图9中。目标城市用五角星表示，其邻居节点的大小表示它自己的出向强度大小，颜色表示它自己的出向核数大小。这些邻居对目标城市的转发量从目标城市左侧沿着顺时针方向依次递减。从图中可以清楚地看到转发厦门、苏州微博最多的这些城市大部分为高影响力的城市，而转发太

原、哈尔滨微博最多的这些城市都是影响力较弱的城市。据此，便解释了这些城市出向强度排名与出向H指数排名和出向核数核名存在较大差异的原因，也进一步证明了H指数与核数在刻画城市影响力时具有更高的准确性。

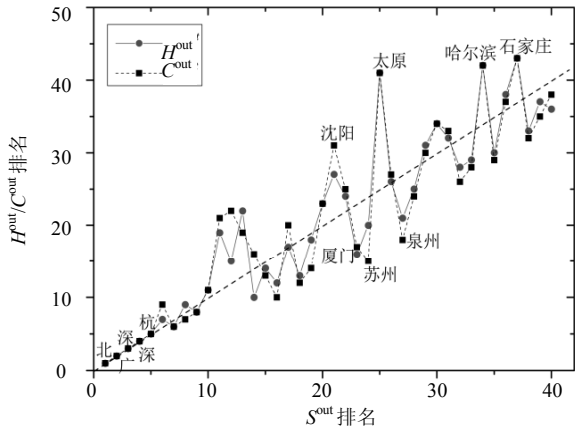
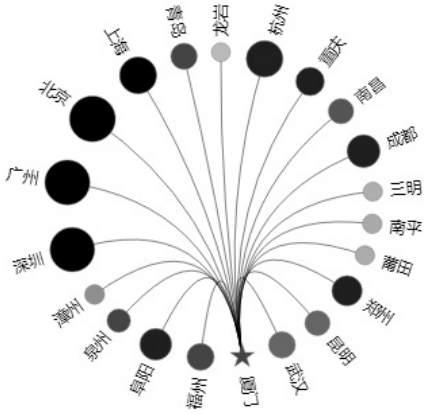
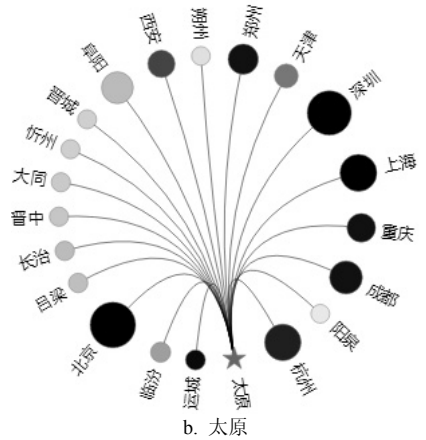


图8 出度方向上不同指标的比较

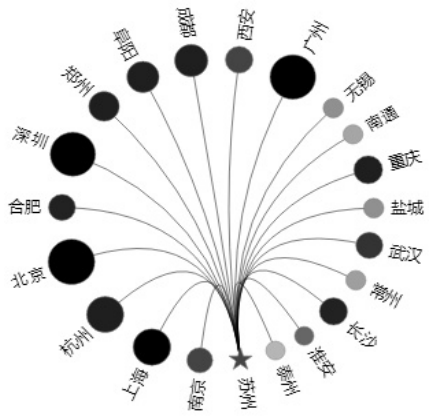
一个城市的入强度等于这个城市从其他城市转发的微博数之和，代表了这个城市的活跃度，反之，一个城市的出强度等于其他城市从该城市转走的微博数之和，代表了这个城市的影响力——转发该城市微博的邻居城市影响力越大，转发的数量越多，则说明该城市的影响力也越大。



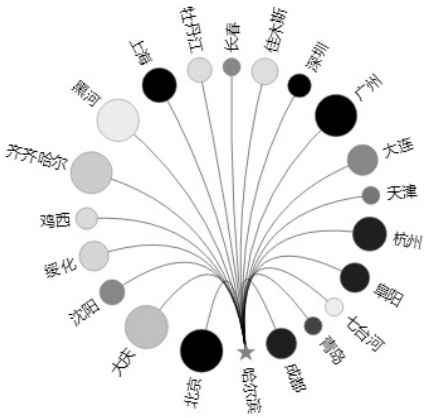
a. 厦门



b. 太原



c. 苏州



d. 哈尔滨

图9 转发厦门、太原、苏州、哈尔滨微博最多的各自Top20个邻居城市

本文进一步对于每个城市的具体转发量按出向和入向进行了分析，最终得出以下结论：

1) 北京、上海、广州、深圳4个城市，无论是它们自身的活跃度(入强度居于前4位)还是对其他城市的影响力都是最高的。具体来说，网络中除了一个孤立节点(三沙市)之外，其他所有城市都至少与北、上、广、深中的一个存在交互；此外，有37个城市从北、上、广、深中的某一个城市转发的微博在其所有转发的微博中占比最高，占所有城市数量的12.8%；有65个城市从北、上、广、深4个城市转发的微博总数高于从其他城市转发的微博数，占所有城市数量的22.5%，且这些城市的平均出强度排名为80.7。

2) 对一个城市而言，与它交互最多的城市是同一省份的另一个城市，这样的城市有255个，占比88%(图7中也可以看出来)，而这其中与它交互最强的城市是它的省会城市的有229个，在所有城市中占比79.2%。这体现了在线社交媒体具有很强的“局域性”特征。

3) 共有89个城市(均为非省会城市)的省内转发总数大于省外转发总数，他们的平均出强度排名为159，也就是说这89个城市大都是影响力较弱的城市。在所有城市按照各自的出向H指数(或者入向H指数)降序排列的列表中，排名越靠前的城市，上述的“局域性”特征越弱，反之则越强。可以在一定程度上说，一个城市的整体影响力越大，则它“局域性”特征越弱，反之则越强；

4) 无论出向还是入向，所有省会城市以及直辖市的含权H指数的平均出强度排名约为30，且与它们交互最强的城市是非省内城市的占其中三分之一。所以一般而言省会城市的影响力更大，“局域性”特征较弱；非省会城市的影响力较弱，“局域性”特征较强。

之后，本文考察了无权无向H指数和有向含权H指数分别与城市就业人口、城市生产总值之间的相关性，也发现有向含权H指数的相关性要高于无权无向H指数。最后，又参照最新的中国城市等级划分标准[41]，发现有向含权H指数也比无向无权H指数能更加准确地反映城市的规模划分。

4 结 束 语

本文首先将H指数这一重要节点挖掘方法拓展到了有向含权网络上，并证明了无论是同步更新还是异步更新，DHC定理在有向含权网络中仍然成立。这为含权网络上的k-壳分解提供了新的计算方法。接着在3个真实网络C. elegans网络、USAir网络和中国城市间微博转发网络上对比分析了含权与不含权两种情况下各指标在识别重要节点方面的表现，并计算了其与WSIR模型仿真结果的相关性，发现含权的各项指标的性能都好于不含权的情况。最后，应用有向含权H指数和核数对中国城市的在线媒体影响力进行排名，识别具有高影响力的城市，并总结了不同城市对于信息转发的偏好特征。

参 考 文 献

[1] BARABÁSI A L. Network science[M]. Cambridge: Cambridge University Press, 2016.

[2] 任晓龙, 朱燕燕, 王思云, 等. 在线社交网络结构与区域经济关联性研究[J]. 电子科技大学学报, 2015, 44(5): 643-651.
REN Xiao-long, ZHU Yan-yan, WANG Si-yun, et al. Online social network analysis and the relation with regional economic development[J]. Journal of University of Electronic Science and Technology of China, 2015, 44(5): 643-651.

[3] HUBERMAN B A, ADAMIC L A. Internet: Growth dynamics of the world-wide web[J]. Nature, 1999, 401(6749): 131.

[4] GUIMERA R, AMARAL L A N. Modeling the world-wide airport network[J]. The European Physical Journal B-Condensed Matter and Complex Systems, 2004, 38(2): 381-385.

[5] WILLIAMS R J, BERLOW E L, Dunne J A, et al. Two degrees of separation in complex food webs[J]. Proceedings of the National Academy of Sciences of the United States of America, 2002, 99(20): 12913-12916.

[6] 冷炳荣, 杨永春, 李英杰, 等. 中国城市经济网络结构空间特征及其复杂性分析[J]. 地理学报, 2011, 66(2): 199-211.
LENG Bing-rong, YANG Yong-chun, LI Ying-jie, et al. Spatial characteristics and complex analysis: a perspective from basic activities of urban networks in China[J]. Acta Geographica Sinica, 2011, 66(2): 199-211.

[7] SCHICH M, SONG C, AHN Y Y, et al. A network framework of cultural history[J]. Science, 2014, 345(6196): 558-562.

[8] KRIOUKOV D, KITSAK M, SINKOVITS R S, et al. Network cosmology[J]. Scientific Reports, 2012, 2(20): 793.

[9] NEWMAN M E J. Power laws, Pareto distributions and Zipf's law[J]. Contemporary Physics, 2005, 46(5): 323-351.

[10] LÜ Lin-yuan, ZHANG Zi-ke, ZHOU Tao. Zipf's law leads to Heaps' law: Analyzing their relation in finite-size systems[J]. PloS One, 2010, 5(12): e14139.

[11] LOTKA A J. The frequency distribution of scientific productivity[J]. Journal of the Washington Academy of Sciences, 1926, 16(12): 317-323.

[12] DE SOLLA PRICE D J. Little science big science[M]. New York: Columbia Press, 1963.

[13] 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: 1175-1197.
REN Xiao-long, LÜ Lin-yuan. Review of ranking nodes in complex networks (in Chinese)[J]. Chin Sci Bull (Chin Ver), 2014, 59: 1175-1197.

[14] LÜ Lin-yuan, CHEN Duan-bing, REN Xiao-Long, et al. Vital nodes identification in complex networks[J]. Physics Reports, 2016, 650: 1-63.

[15] 赫南, 李德毅, 淦文燕, 等. 复杂网络中重要性节点发掘综述[J]. 计算机科学, 2007, 34(12): 1-5.
HE Nan, LI De-yi, GAN Wen-yan, et al. Mining vital nodes in complex networks[J]. Computer Science, 2007, 34(12): 1-5.

[16] 刘建国, 任卓明, 郭强, 等. 复杂网络中节点重要性排序的研究进展[J]. 物理学报, 2013, 62(17): 178901.
LIU Jian-guo, REN Zhuo-ming, GUO Qiang, et al. Node importance ranking of complex networks[J]. Acta Physica Sinica, 2013, 62(17): 178901.

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

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

[19] CHEN Duan-bing, LÜ Lin-yuan, SHANG Ming-sheng, et

al. Identifying influential nodes in complex networks[J]. Physica A: Statistical mechanics and its applications, 2012, 391(4): 1777-1787.

[20] KITSAK M, GALLOS L K, HAVLIN S, et al. Identification of influential spreaders in complex networks[J]. Nature Physics, 2010, 6(11): 888-893.

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

[22] FREEMAN L C. A set of measures of centrality based on betweenness[J]. Sociometry, 1977: 35-41.

[23] RUHNAU B. Eigenvector-centrality—a node-centrality?[J]. Social Networks, 2000, 22(4): 357-365.

[24] BRIN S, PAGE L. The anatomy of a large-scale hypertextual web search engine[J]. Computer networks and ISDN Systems, 1998, 30(1): 107-117.

[25] LÜ Lin-yuan, ZHANG Yi-cheng, YEUNG C H, et al. Leaders in social networks, the delicious case[J]. PloS One, 2011, 6(6): e21202.

[26] 陈勇, 胡爱群, 胡啸. 通信网中节点重要性的评价方法[J]. 通信学报, 2004, 25(8): 129-134.
CHEN Yong, HU Ai-qun, HU Xiao. Evaluation method for node importance in communication networks[J]. Journal of China Institute of Communications, 2004, 25(8): 129-134.

[27] 谭跃进, 吴俊, 邓宏钟. 复杂网络中节点重要度评估的节点收缩方法[J]. 系统工程理论与实践, 2006, 26(11): 79-83.
TAN Yue-jin, WU Jun, DENG Hong-zhong. Evaluation method for node importance based on node contraction in complex networks[J]. System Engineering Theory and Practice, 2006, 26(11): 79-83.

[28] RESTREPO J G, OTT E, HUNT B R. Characterizing the dynamical importance of network nodes and links[J]. Physical Review Letters, 2006, 97(9): 094102.

[29] HIRSCH J E. An index to quantify an individual's scientific research output[J]. Proceedings of the National academy of Sciences of the United States of America, 2005: 16569-16572.

[30] KORN A, SCHUBERT A, TELCS A. Lobby index in networks[J]. Physica A: Statistical Mechanics and Its Applications, 2009, 388(11): 2221-2226.

[31] LÜ Lin-yuan, ZHOU Tao, ZHANG Qian-ming, et al. The H-index of a network node and its relation to degree and coreness[J]. Nature Communications, 2016, 7: 10168.

[32] LEE Yan-li, ZHOU Tao. Fast asynchronous updating algorithms for k-shell indices[J]. Physica A: Statistical Mechanics and Its Applications, 2017(9), 482: 524-531.

[33] LÜ Lin-yuan, CHEN Duan-bing, REN Xiao-long, et al. Vital nodes identification in complex networks[J]. Physics Reports, 2016, 650: 1-63.

[34] EIDSAA M, ALMAAS E. S-core network decomposition: a generalization of k-core analysis to weighted networks[J]. Physical Review E, 2013, 88(6): 062819.

[35] FAN Tian-long, ZHU Yan-yan, WU Lei-lei, et al. DHC theorem proving[EB/OL]. [2017-03-27]. http://linkprediction.org/index.php/link/resource/code.

[36] DUCH J, ARENAS A. Community detection in complex networks using extremal optimization[J]. Physical Review E, 2005, 72(2): 027104.

[37] ZHOU Tao, LÜ Lin-yuan, ZHANG Yi-cheng. Predicting missing links via local information[J]. The European Physical Journal B-Condensed Matter and Complex Systems, 2009, 71(4): 623-630.

[38] GARAS A, ARGYRAKIS P, ROZENBLAT C. 2010 Worldwide spreading of economic crisis[J]. New Journal of Physics, 2010, 12(2): 185-188.

[39] KENDALL M G. A new measure of rank correlation[J]. Biometrika, 1938, 30(1-2): 81-93.

[40] BARRAT A, BARTHELEMY M, PASTOR-SATORRAS R, et al. The architecture of complex weighted networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(11): 3747-3752.

[41] 国务院. 国务院关于调整城市规模划分标准的通知[EB/OL]. (2014-11-20). http://www.gov.cn/zhengce/content/2014-11/20/content_9225.htm.
The State Council of the People's Republic of China. Circular on adjusting the standard of city size division [EB/OL]. (2014-11-20). http://www.gov.cn/zhengce/content/2014-11/20/content_9225.htm.

编 辑 蒋 晓