基于高阶信息的网络相似性比较方法*

陈浩宇[1)] 徐涛[1)] 刘闯[1)] 张子柯[2)3)] 詹秀秀[1)3)†]

1) (杭州师范大学复杂科学研究中心, 杭州 311121)

2) (浙江大学数字沟通研究中心, 杭州 310058)

3) (浙江大学传媒与国际文化学院, 杭州 310058)

(2023 年 7 月 5 日收到; 2023 年 10 月 6 日收到修改稿)

量化复杂网络之间的结构相似性是网络科学中一个基本且具有挑战性的问题, 在医学、社会学等多个领域发挥了至关重要的作用. 传统的网络比较方法通常基于简单的结构特征, 例如节点度分布、最短路径长度等, 这些方法可能无法充分捕捉网络的全局结构信息, 导致得到的网络相似性不精准. 本文提出了一种基于高阶信息的网络相似性比较方法, 该方法同时考虑了网络的全局结构和局部结构. 具体而言, 通过构建网络节点的高阶聚类系数分布和节点间距离分布, 并利用基于这两个分布的 Jensen-Shannon 散度来量化网络之间的相似性. 实验结果表明, 相较于其他基线方法, 本文提出的方法不仅能高效地比较不同网络的相似性, 且在对真实网络进行扰动的过程中也表现出鲁棒性.

关键词: 网络相似性, 高阶聚类系数, 距离分布

PACS: 89.75.–k, 89.75.Fb **DOI:** 10.7498/aps.73.20231096

1 引 言

大多数实际的复杂系统都可以表示成复杂网络, 从细菌、细胞和蛋白质系统, 到人类关系网络, 甚至再到大型的互联网和万维网[1,2]. 近年来, 研究各种复杂网络结构的相似性已经成为一个流行且跨学科的主题, 目前在社会科学[3]、医学[4]、生物学[5]和计算机科学[6]等领域得到了广泛应用. 比较网络之间的相似性在当今社会发展中扮演着不可或缺的角色[7–11]. 例如, 在医学领域[12], 研究人员可以利用基因网络比较不同组织之间的相似性, 以此发现与疾病相关的基因和信号通路; 在生物领域[13], 通过相似性方法来比较不同生物体或不同条件下的蛋白质相互作用网络, 可以发现共享的功能模块以及关键的蛋白质节点; 在社交领域[14], 利用相似性比较不同社交网络中的社区结构, 从而揭示社区的划分和特征.

网络比较问题最初源于图同构[15,16]问题, 也被称为图匹配[17]或网络对齐[18], 其研究的本质在于判断两个图中的节点集是否存在一对一映射的关系[19,20]. 通过比较网络的结构拓扑性质来量化网络之间的异同性, 例如边密度、度分布和节点距离分布[21]等较为简单的拓扑结构特征, 或者网络社区结构[22]、光谱熵[23,24]等更为复杂的拓扑结构特征[25–27]. 例如, Koutra 等[28]提出了一种基于 Matusita 距离来度量网络节点亲和力矩阵之间的相似度方法; De Domenico 和 Biamonte[29]提出了一套基于谱熵的网络比较信息理论工具; Schieber 等[30]通过考虑节点之间的最短路径距离的概率分布, 高效量化了网络之间的差异; Chen 等[31]提出了一种基于节点可通信性序列和谱熵的比较方法; Liu 等[32]

* 国家自然科学基金 (批准号: 72371224, 92146001)、浙江省自然科学基金 (批准号: LQ22F030008)、中央高校基本科研业务费、杭州师范大学科研启动项目 (批准号: 2021QDL030) 和兵团财政科技计划项目 (批准号: 2021AB034) 资助的课题.

† 通信作者. E-mail: zhanxiuxiu@hznu.edu.cn

提出了基于遗传算法和模型选择的机器学习的网络比较方法. 这些方法的基本思想都是选取网络特定的拓扑性质, 并选择特定的熵度量来比较网络的异同[33]. 然而, 单独选取特定的拓扑性质无法充分捕获网络的全局结构信息. 因此, 我们引进了网络节点的高阶聚类系数, 并通过构建节点的高阶聚类系数分布和最短路径分布, 再基于两个分布之间的 Jensen-Shannon 散度[34]来度量网络节点连通异质性. 该方法在综合考虑网络全局性质重要性的同时, 也兼顾了网络局部的拓扑结构的影响.

本文的其余部分结构如下: 第 2 节介绍本文用到的相关概念, 提出基于高阶信息的网络相似性比较方法, 并且介绍网络比较的常见基线算法; 第 3 节是实验部分, 分别在人工合成网络和真实网络上评估了此方法; 第 4 节是总结与展望, 并为今后的网络比较工作提供了全新思路.

2 基于高阶信息的网络相似性方法

2.1 基本定义

符号定义 为了方便讨论, 将网络统一记为 $G=(V,E)$, 节点集合 V 和边集合 E 分别表示为 $V=\{v_1,v_2,\cdots,v_N\}$, $E=\{e_k=(v_i,v_j),\ k=1,\cdots,M|v_i,v_j\in V\}$, 其中 N 和 M 分别表示网络 G 中节点和边的数量. 根据网络中节点与边的隶属关系, 构建邻接矩阵 $\boldsymbol{A}_{N\times N}$, 具体而言, 当节点 v_i 和节点 v_j 有连边时, A_{ij} 的值为 1; 否则为 0. 此外, 构建网络中节点的邻居集 $Q=\{q_1,q_2,\cdots,q_N\}$, 其中 q_i 为节点 v_i 的邻居集合, $|q_i|$ 为节点 v_i 的邻居个数.

2.2 高阶聚类系数

聚类系数是一种描述网络拓扑结构的指标, 它反映了节点周围邻居之间相互连接的紧密程度, 且可以衡量网络中节点的聚集程度. 然而, 在真实网络中, 由于节点间最短路径长度较短且聚类系数较大, 且这样的局部特征不能充分描述节点周围邻居之间的高阶联系, 往往导致无法捕捉网络中局部拓扑结构性质的差异. 为了解决这个问题, 本文引入了一种更高阶的聚类系数定义[35], 在标准聚类系数的基础上进行了扩展.

传统标准聚类系数只考虑三角形子图的数量来度量网络中节点的聚集程度, 而高阶聚类系数则进一步考虑了更高阶的子图, 从而能够更准确地刻画节点邻居间的联系, 捕捉到网络中的更复杂的拓扑结构. 它能帮助我们理解网络的聚集性和功能模块, 从而提高网络的鲁棒性和可靠性, 这对于研究网络的演化过程具有重要意义. 例如, 在分析社会团体网络时, 有利于发现更为复杂的关联模式, 揭示群体内部的交互模式和组织形式[36].

首先定义 $E_i(x)$ 为网络中去除 v_i 以及对应的连边后 v_i 的邻居中距离为 x 的节点对的数量. 因此, $s_i(x)$ 定义为去除节点 v_i 之后, 其邻居之间距离为 x 的节点对的比例, 表达式为

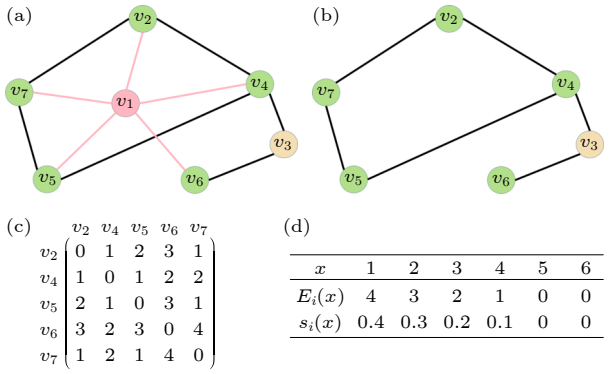
$$s_i(x)=\frac{2E_i(x)}{|q_i|\left(|q_i|-1\right)}. \tag{1}$$

本文称 $s_i(x)$ 为关于节点 v_i 的 x-阶聚类系数, 根据 x 数值的改变, 可以定义任意阶聚类系数. 因此, 对于网络中的所有节点可以构建一个完整的网络节点高阶聚类系数分布矩阵 $\boldsymbol{S}=\{\boldsymbol{S}_1,\boldsymbol{S}_2,\cdots,\boldsymbol{S}_N\}$. 具体来说, 节点 v_i 的聚类系数分布为 $\boldsymbol{S}_i=\{s_i(x)|1\leqslant x\leqslant N-1\}$. 由于在计算节点 v_i 的聚类系数分布时会在网络中移除该节点, 因此新生成的网络直径最大可能值为 $N-2$. 其中 $s_i(x)(1\leqslant x\leqslant N-2)$ 为节点 v_i 的 x-阶聚类系数值, 而 $\boldsymbol{S}_i$ 的最后一列值 $s_i(N-1)$ 为节点 v_i 的邻居之间不存在路径的节点对的比例.

图 1 给出了高阶聚类系数的一个例子, 其中图 1(a) 是一个由 7 个节点构成的小网络, 在此网络中去除节点 v_1 及其相应的连边后可得到图 1(b). 计算图 1(b) 中节点 v_1 的邻居之间的距离可得矩阵, 如图 1(c) 所示, 例如节点 v_2 和节点 v_6 之间的距离为 3. 由图 1(c) 的距离矩阵可以进一步计算节点 v_1 的距离分布, 结果如图 1(d) 所示, 即阶数为 1 的聚类系数为 $s_1(1)=0.4$, 阶数为 2 的聚类系数为 $s_1(2)=0.3$, 以此类推, 阶数为 5 的聚类系数为 $s_1(5)=0$. 值得注意的是, 因为节点 v_1 的每两个邻居间都存在路径, 因此阶数为 6 的聚类系数 $s_1(6)=0$.

2.3 基于高阶信息的网络相似性算法

首先定义节点的距离分布, 即一个网络节点间距离分布矩阵由 $\boldsymbol{P}=\{\boldsymbol{P}_1,\boldsymbol{P}_2,\cdots,\boldsymbol{P}_N\}$ 表示, 每个节点 v_i 的距离分布向量为 $\boldsymbol{P}_i=\{p_i(x)|1\leqslant x\leqslant d+1\}$, 其中 $p_i(x)$ 表示与节点 v_i 距离为 x 的节点的比例, 而 $p_i(d+1)$ 则表示与节点 v_i 不存在路径的节点的比例, d 为网络直径.



(c)

	v_2	v_4	v_5	v_6	v_7
v_2	0	1	2	3	1
v_4	1	0	1	2	2
v_5	2	1	0	3	1
v_6	3	2	3	0	4
v_7	1	2	1	4	0

(d)

x	1	2	3	4	5	6
$E_i(x)$	4	3	2	1	0	0
$s_i(x)$	0.4	0.3	0.2	0.1	0	0

图 1 网络高阶聚类系数计算示意图 (a) 节点 v_1 及它的邻居形成的网络; (b) 去除节点 v_1 后的网络; (c) 图 1(b) 所示网络中节点 v_1 的邻居之间的距离矩阵; (d) 节点 v_1 的高阶聚类系数分布

Fig. 1. Illustration of the calculation of the higher-order clustering coefficient: (a) A network formed by node v_1 and its neighbors; (b) network after removing node v_1; (c) distance matrix between neighbors of node v_1 in the network shown in panel (b); (d) the higher-order clustering coefficient distribution of node v_1.

节点 v_i 的高阶聚类系数分布和距离分布分别考虑了 v_i 的邻居之间的聚集程度以及 v_i 与网络中其他节点的距离远近, 以下综合这两种分布来定义网络的相似性算法. 首先, 因为 $d+1 \leqslant N$, 可以将距离分布 $\boldsymbol{P}_i$ 通过补零的方式变成一个 $N-1$ 阶向量. 然后定义网络 G 的高阶信息分布 $\boldsymbol{T}_i = \{\gamma \boldsymbol{S}_i, (1-\gamma)\boldsymbol{P}_i\}$, 且 $\boldsymbol{T}_i$ 的维数为 $1 \times (2N-2)$, 其中 $\gamma \in [0,1]$, 可以调节 $\boldsymbol{S}_i$ 和 $\boldsymbol{P}_i$ 在 $\boldsymbol{T}_i$ 中所占的比例, γ 越大, 表示分布里面更注重高阶聚类系数信息, 如果 γ 趋于 0 表示更注重距离信息.

给定网络 G 以及其上的高阶信息分布 $\boldsymbol{T}$, 根据 Jensen-Shannon 散度来定义网络节点散度 NND (network node dispersion):

$$\mathrm{NND}(G) = \frac{\mathcal{J}(\boldsymbol{T}_1, \boldsymbol{T}_2, \cdots, \boldsymbol{T}_N)}{\log(N+1)}, \tag{2}$$

式中 $\mathcal{J}(\boldsymbol{T}_1, \boldsymbol{T}_2, \cdots, \boldsymbol{T}_N)$ 表示 N 个节点分布的 Jensen-Shannon 散度, 其表达式为

$$\mathcal{J}(\boldsymbol{T}_1, \boldsymbol{T}_2, \cdots, \boldsymbol{T}_N) = \frac{1}{N}\sum_{i,j} t_i(j)\log\left(\frac{t_i(j)}{\mu_j}\right), \tag{3}$$

其中 $t_i(j)$ 是高阶信息分布 $\boldsymbol{T}_i$ 的第 j 列的具体数值; 而 μ_j 为 N 个节点分布第 j 维的平均值, 其表达式为

$$\mu_j = \frac{1}{N}\left(\sum_{i=1}^{N} t_i(j)\right). \tag{4}$$

网络 NND 衡量了网络节点连通异质性的大小, 其值越大表示网络中节点连通异质性越大, 反之亦然.

网络之间的相似性 给定网络 G 和 G', 可以使用以下公式来计算两者在结构上的相似性 $D_{\mathrm{HC}}(G, G')$, 表达式为

$$D_{\mathrm{HC}}(G, G') = \beta\sqrt{\frac{\mathcal{J}(\mu_G, \mu_{G'})}{\ln 2}} + (1-\beta)\left|\sqrt{\mathrm{NND}(G)} - \sqrt{\mathrm{NND}(G')}\right|, \tag{5}$$

网络结构相似性 D_{HC} 由两个部分组成. 第 1 部分考虑了网络中节点的平均连通性, 其中 $\mathcal{J}(\mu_G, \mu_{G'})$ 表示两个网络的平均距离分布的 Jensen-Shannon 散度, $\mu_G = \{\mu_j | (1 \leqslant j \leqslant 2N-2)\}$, 它包含了基于节点高阶聚类系数和节点间距离分布的信息, 从而捕捉了网络的全局结构性质; 第 2 部分考虑了节点之间的异质性, 即网络中局部的结构性质. 其中参数 β 用于调节它们的权重. 如果两个网络之间的 D_{HC} 值越小, 说明它们的结构相似性越大; 反之, D_{HC} 值越大, 说明它们的结构相似性越小.

图 2 给出一种基于高阶信息的网络相似性比较方法计算流程图, 用于比较网络 G 和 G'. 图 2(a) 展示了网络 G 和 G' 的结构, 其中 G 是一个全连通网络, 而 G' 是一个含有一个孤立节点的网络. 通过以下步骤计算这两个网络之间的相似性, 具体过程如图 2(b) 和图 2(c) 所示: 首先, 计算节点的高阶聚类系数分布和节点的距离分布; 然后, 根据 (2) 式和 (5) 式计算两个网络之间的相似性. 最终, 得到网络 G 和 G' 之间的 D_{HC} 值为 0.39.

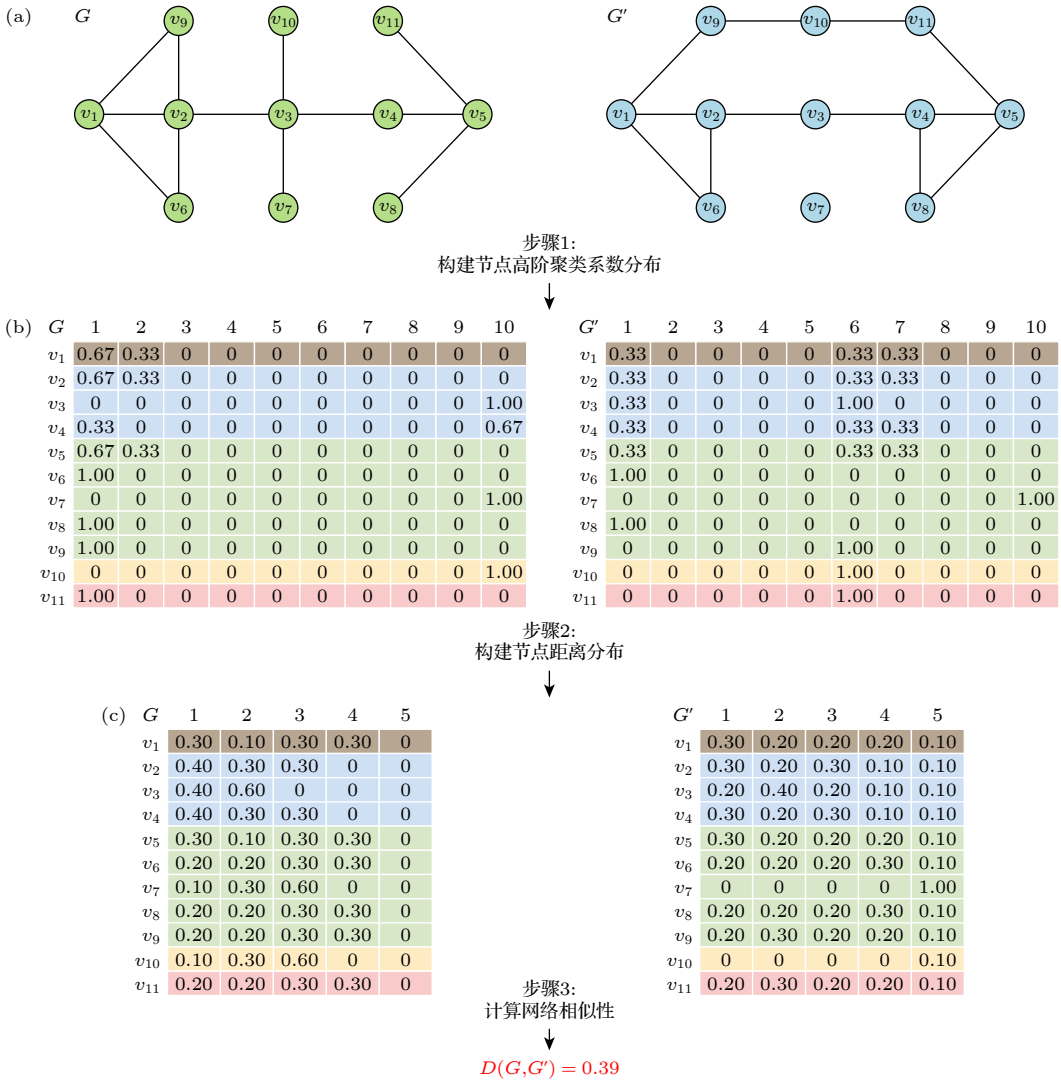
2.4 基线算法

基于距离分布的网络相似性方法[30]

根据上述节点 v_i 的距离分布 $\boldsymbol{P}_i = \{p_i(x)\}$, 由距离分布和 Jensen-Shannon 散度定义两个网络 G, G' 的相似性比较算法:

$$D_{\mathrm{SP}}(G, G') = w_1\sqrt{\frac{\mathcal{J}(\mu_G, \mu_{G'})}{\ln 2}} + w_2\left|\sqrt{\mathrm{NND}_{\mathrm{SP}}(G)} - \sqrt{\mathrm{NND}_{\mathrm{SP}}(G')}\right| + \frac{w_3}{2}\left(\sqrt{\frac{\mathcal{J}(P_{\alpha G}, P_{\alpha G'})}{\ln 2}} + \sqrt{\frac{\mathcal{J}(P_{\alpha G_c}, P_{\alpha G'_c})}{\ln 2}}\right), \tag{6}$$

D_{SP} 共基于 3 个距离的概率分布, 第 1 项表示以平均距离分布 (即 μ_G 和 $\mu_{G'}$) 为特征的差异性, 第 2



(b) G	1	2	3	4	5	6	7	8	9	10
v_1	0.67	0.33	0	0	0	0	0	0	0	0
v_2	0.67	0.33	0	0	0	0	0	0	0	0
v_3	0	0	0	0	0	0	0	0	0	1.00
v_4	0.33	0	0	0	0	0	0	0	0	0.67
v_5	0.67	0.33	0	0	0	0	0	0	0	0
v_6	1.00	0	0	0	0	0	0	0	0	0
v_7	0	0	0	0	0	0	0	0	0	1.00
v_8	1.00	0	0	0	0	0	0	0	0	0
v_9	1.00	0	0	0	0	0	0	0	0	0
v_{10}	0	0	0	0	0	0	0	0	0	1.00
v_{11}	1.00	0	0	0	0	0	0	0	0	0

G'	1	2	3	4	5	6	7	8	9	10
v_1	0.33	0	0	0	0	0.33	0.33	0	0	0
v_2	0.33	0	0	0	0	0.33	0.33	0	0	0
v_3	0.33	0	0	0	0	1.00	0	0	0	0
v_4	0.33	0	0	0	0	0.33	0.33	0	0	0
v_5	0.33	0	0	0	0	0.33	0.33	0	0	0
v_6	1.00	0	0	0	0	0	0	0	0	0
v_7	0	0	0	0	0	0	0	0	0	1.00
v_8	1.00	0	0	0	0	0	0	0	0	0
v_9	0	0	0	0	0	1.00	0	0	0	0
v_{10}	0	0	0	0	0	1.00	0	0	0	0
v_{11}	0	0	0	0	0	1.00	0	0	0	0

(c) G	1	2	3	4	5
v_1	0.30	0.10	0.30	0.30	0
v_2	0.40	0.30	0.30	0	0
v_3	0.40	0.60	0	0	0
v_4	0.40	0.30	0.30	0	0
v_5	0.30	0.10	0.30	0.30	0
v_6	0.20	0.20	0.30	0.30	0
v_7	0.10	0.30	0.60	0	0
v_8	0.20	0.20	0.30	0.30	0
v_9	0.20	0.20	0.30	0.30	0
v_{10}	0.10	0.30	0.60	0	0
v_{11}	0.20	0.20	0.30	0.30	0

G'	1	2	3	4	5
v_1	0.30	0.20	0.20	0.20	0.10
v_2	0.30	0.20	0.30	0.10	0.10
v_3	0.20	0.40	0.20	0.10	0.10
v_4	0.30	0.20	0.30	0.10	0.10
v_5	0.30	0.20	0.20	0.20	0.10
v_6	0.20	0.20	0.20	0.30	0.10
v_7	0	0	0	0	1.00
v_8	0.20	0.20	0.20	0.30	0.10
v_9	0.20	0.30	0.20	0.20	0.10
v_{10}	0	0	0	0	0.10
v_{11}	0.20	0.30	0.20	0.20	0.10

图 2 基于高阶信息的网络比较方法计算流程示意图 (a) 给定两个拥有 11 个节点的网络 G 和 G', 其中 G 有 14 条边, G' 有 12 条边; (b) 如何计算基于高阶信息的网络相似性的示例, 包含了节点高阶聚类系数分布和节点距离分布; (c) 网络相似值的计算, 其中 $\beta = 0.5$

Fig. 2. Schematic diagram of calculation flow of network comparison method based on high-order information: (a) Given two networks G and G' with 11 nodes, G has 14 edges and G' has 12 edges; (b) an illustration of how to compute the network similarity based on higher-order information, including the distribution of node higher-order clustering coefficients and node distance distribution; (c) calculation of the network similarity value $D_{\rm HC}$, where $\beta = 0.5$.

项表示网络节点分散度的差异性, 第 3 项则是网络 α-中心性分布的差异, 且 G_c 是 G 的补图. 这里

$$\mathrm{NND}_{\mathrm{SP}}(G) = \frac{\mathcal{J}(\boldsymbol{P}_1, \cdots, \boldsymbol{P}_N)}{\log(d+1)}, \tag{7}$$

式中, $\mathcal{J}(\boldsymbol{P}_1, \cdots, \boldsymbol{P}_N)$ 是 N 个节点距离分布的 Jensen-Shannon 散度; w_1, w_2, w_3 和 α 是可调参数, 且满足 $w_1 + w_2 + w_3 = 1$. 本文设置权重 $w_1 = w_2 = 0.45$ 和 $w_3 = 0.1$ 来进行网络相似性比较.

基于可通信性序列熵的网络相似性方法 [31]

通过构造网络可通信性矩阵 $\boldsymbol{C}$ 来测量节点之间的通信能力, 其定义如下:

$$\boldsymbol{C} = \mathrm{e}^{A} = \sum_{y=0}^{\infty} \frac{1}{y!} A^y = \begin{Bmatrix} C_{11} & C_{12} & \cdots & C_{1N} \\ C_{21} & C_{22} & \cdots & C_{2N} \\ \vdots & \vdots & \ddots & \\ C_{N1} & C_{N2} & \cdots & C_{NN} \end{Bmatrix}, \tag{8}$$

其中 C_{ij} 表示为节点 v_i 和 v_j 之间的可通信性, 它反映了网络中节点之间的信息传递能力. 假设 $L = \{L_1, L_2, \cdots, L_M\}$ 是标准化的可通信性序列, 其中

$$L_y = C_{ij} \Big/ \Big(\sum\nolimits_{i=1}^{N} \sum\nolimits_{j=1}^{N} C_{ij} \Big)$$

$(1 \leqslant y \leqslant M, 1 \leqslant i \leqslant j \leqslant N$和$M = N(N+1)/2)$,

L 序列的香农熵 $H(L)$ 定义如下:

$$H(L)=-\sum_{i=1}^{M}L_i\log_2 L_i, \tag{9}$$

给定两个网络 G 和 G', 归一化可通信性序列分别由 L^G 和 $L^{G'}$ 给出. 按升序对 $L^G(L^{G'})$ 中的值进行排序, 并获得新的可通信序列为 $\tilde{L}^G(\tilde{L}^{G'})$. 因此, 基于可通信性序列熵的网络相似性被定义为 $D_C(G,G')$:

$$D_C(G,G')=H\left(\frac{\tilde{L}^G+\tilde{L}^{G'}}{2}\right)-\frac{1}{2}\left[H(\tilde{L}^G)+H(\tilde{L}^{G'})\right]. \tag{10}$$

基于拉普拉斯特征值的网络相似性方法 [37]

该方法通过构造网络拉普拉斯矩阵, 并利用邻接矩阵和度矩阵的特征值计算得到光谱距离. 通过比较网络的光谱距离差异性, 该方法能够更全面地表示网络的拓扑结构, 从而选择更具稳定性的比较方式. 基于拉普拉斯特征值的网络相似性被定义为 $D_M(G,G')$:

$$D_M(G,G')=\frac{1}{N}\sum_{i=1}^{N}|\lambda_i-\lambda_i'|, \tag{11}$$

其中 λ 和 λ' 分别为两个网络中拉普拉斯矩阵的特征值.

3 实 验

3.1 人工合成网络比较

为了验证本文提出的方法在量化网络相似性方面的能力, 进行了参数 β 和 γ 的敏感性分析, 旨在展示基于高阶信息的网络比较方法的鲁棒性. 在实验中选择用 WS 网络和 BA 网络来测试方法的鲁棒性, 其中网络的平均度均为 10. WS 网络中选取重连概率为 $p=0.3$, 而 BA 网络中选取每一步的加边数 $m=5$. 图 3(a) 和图 3(c) 中的每一个点分别表示 $N=1000$ 的 WS 网络与 $N=[1500,5000]$, 间隔为 500 的 WS 网络相似性值 ($D_{\rm HC}$), 不同参数 β 和 γ 的取值分别对应不同颜色的曲线. 图 3(b)

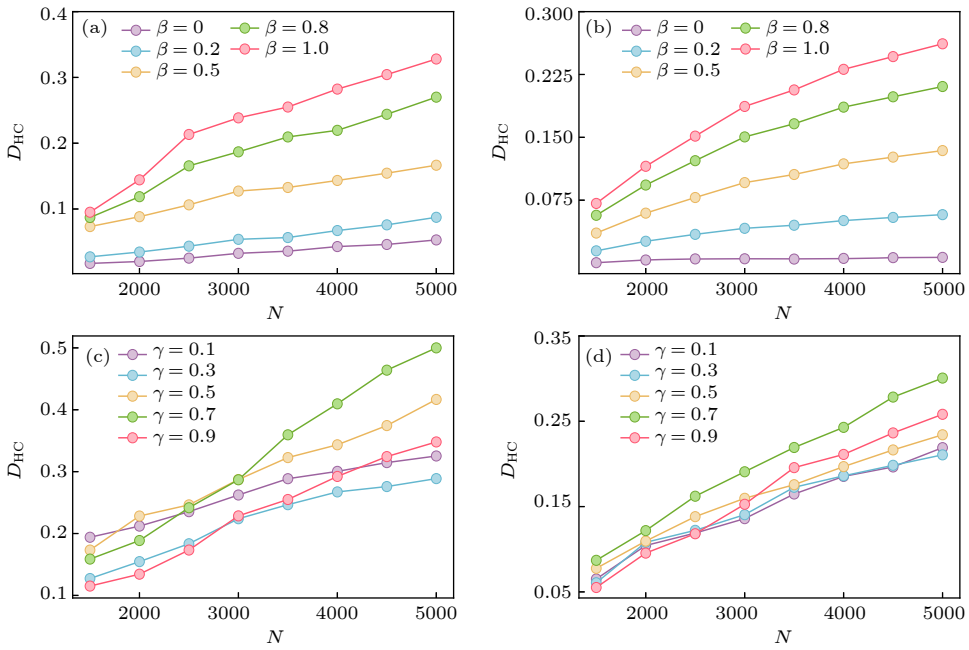


图 3 人工合成网络下 (WS, BA) 的参数敏感性分析 (a) 不同参数 β 下 $N=1000$ 的 WS 网络与 $N=[1500,5000]$, 间隔为 500, 重连概率 $p=0.3$ 的 WS 网络之间的相似性; (b) 不同参数 β 下 $N=1000$ 的 BA 网络与 $N=[1500,5000]$, 间隔为 500 的 BA 网络之间的相似性, 其中每个 BA 网络每一步加边数 $m=5$; (c) 不同参数 γ 下 WS 网络之间的相似性, 参数与 (a) 图一样; (d) 不同参数 γ 下 BA 网络之间的相似性, 参数与 (b) 图一样. 所有的结果均基于 100 次实验的平均值

Fig. 3. Parameter sensitivity analysis of synthetic networks generated by the WS and BA model: (a) Similarity between the WS network of $N=1000$ and the WS networks of $N=[1500,5000]$ with the interval is 500 under different parameters β, where the probability of rewiring $p=0.3$; (b) similarity between the BA network of $N=1000$ and the BA networks of $N=[1500,5000]$ with an interval of 500 under different parameters β, where each BA network adds edges at each step with number of $m=5$; (c) similarity between WS networks under different parameters γ, the parameters are the same as with those in panel (a); (d) similarity between BA networks under different parameters γ, the parameters are the same as those in panel (b). All results are based on an average of 100 realizations.

和图 3(d) 对相同规模的 BA 网络进行了类似的分析. 由图 3(a)—(d) 可以看出, WS (或 BA) 网络与节点数相差小的网络相似性更大, 且 β (或 γ) 并不会影响曲线的趋势. 此外, 图 3(a) 和图 3(b) 显示 β 的值越大网络差异性就越明显, 因此在后续的所有实验分析中将 β 设置为 1. 从图 3(c) 和图 3(d) 可以观察到, 两种不同人工合成网络下当 $\gamma = 0.7$ 时 (即在高阶信息分布中较注重高阶聚类系数的拓扑信息), 得到的相似性效果最为显著. 因此后续的所有实验分析都将 γ 设置为 0.7.

图 4 在由模型生成的人工网络上 (即 ER, WS, BA 网络) 对比了提出的基于高阶信息的网络比较方法与其他基线方法在网络相似性上的性能, 其中网络规模统一设置为 $N = 1000$. 在 ER 模型中, 通

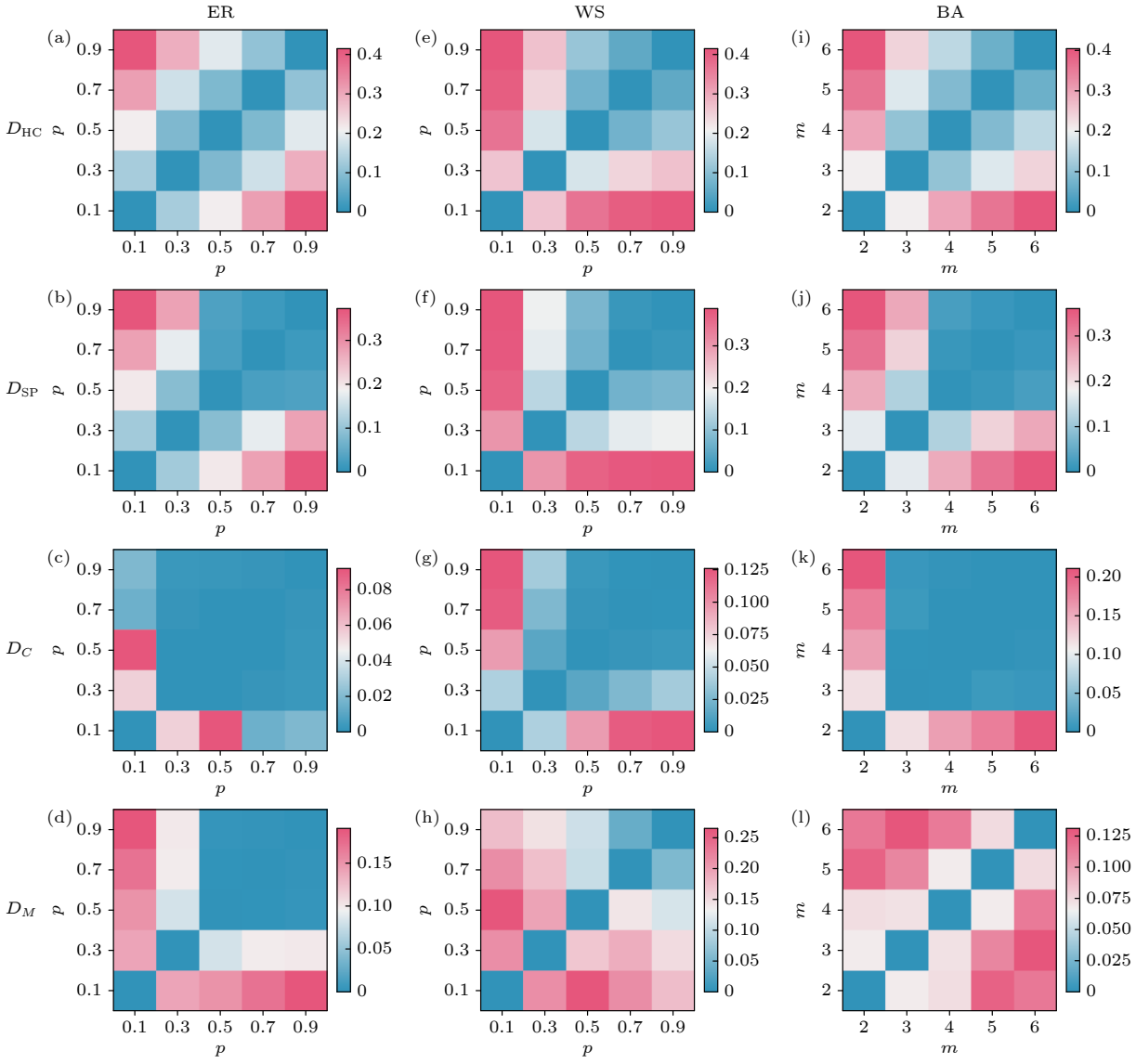


图 4 四种相似性方法在人工合成网络上的效果评估 (网络规模均为 $N = 1000$) (a)—(d) 不同重连概率 p 下 ER 模型生成的每对网络的相似性, 其中相似性方法分别为 D_{HC}, D_{SP}, D_C 以及 D_M; (e)—(h) 不同重连概率 p, 平均度为 10 下 WS 模型生成的每对网络的相似性, 其中相似性方法分别为 D_{HC}, D_{SP}, D_C 以及 D_M; (i)—(l) 不同加边数 $m \in \{2,3,4,5,6\}$ 下 BA 模型生成的每对网络的相似性, 其中相似性方法分别为 D_{HC}, D_{SP}, D_C 以及 D_M. 所有的结果均基于 100 次实验的平均值

Fig. 4. Effectiveness of four similarity methods in comparing synthetic networks. The network size is set to $N = 1000$: (a)–(d) Similarity between each pair of networks generated by the ER model under different rewiring probabilities p, where the network comparing methods are D_{HC}, D_{SP}, D_C and D_M; (e)–(h) similarity between each pair of networks generated by the WS model with different rewiring probabilities p and an average degree of 10, where the network comparing methods are D_{HC}, D_{SP}, D_C and D_M; (i)–(l) similarity between each pair of networks generated by the BA model under different edge numbers $m \in \{2,3,4,5,6\}$ added at each time step, where the similarity methods are D_{HC}, D_{SP}, D_C and D_M. All results are based on an average of 100 realizations.

过重连概率 $p \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ 构造网络; 在 WS 模型中, 通过相同重连概率构造网络, 且平均度均为 10; 在 BA 模型中, 通过每一步的加边数 $m \in \{2, 3, 4, 5, 6\}$ 构造网络. 图 4(a)—(d) 给出了 $D_{\rm HC}$, $D_{\rm SP}$, D_C 和 D_M 等 4 种方法在 ER 网络上的性能比较. 直观上来讲, 由相似重连概率生成的网络的相似性会比较高, 而重连概率相差较大的网络的相似性差距较大. 由结果可知, $D_{\rm HC}$ 能够很好地展现这一结论, 而 $D_{\rm SP}$ 方法在 $p \leqslant 0.5$ 表现比较好, 但是 $p \geqslant 0.5$ 的网络之间的相似性差别不大. 同样, D_C 和 D_M 分别在 $p \geqslant 0.3$ 和 $p \geqslant 0.5$ 无法区分由不同重连概率生成的网络. 此外, 在 WS 和 BA 网络上, 基于高阶信息的网络比较方法 $D_{\rm HC}$ 也能够很好地区分不同参数生成的网络, 而 $D_{\rm SP}$ 无法区分 $p \geqslant 0.5$ 的 WS 网络和 $m \geqslant 4$ 的 BA 网络. D_C 无法区分 $p \geqslant 0.5$ 的 WS 网络和 $m \geqslant 3$ 的 BA 网络. 虽然 D_M 能够区分不同概率 p (或不同加边数 m) 下的 WS (或 BA) 网络, 但是其相似性的数值与直观的理解是相反的. 例如, 在 WS 网络上, 直观上来讲, $p = 0.1$ 的 WS 网络与 $p = 0.5$ 的 WS 网络的相似性应该大于 $p = 0.1$ 的 WS 网络与 $p = 0.7$ (或 $p = 0.9$) 的 WS 网络的相似性, 但 D_M 给出的数值恰恰相反. 此外, 在 BA 网络上 D_M 的性能与 WS 网络相似. $D_{\rm HC}$ 和 $D_{\rm SP}$ 相比, 同时应用了节点的距离分布来比较网络, 不同的是 $D_{\rm HC}$ 考虑了节点的高阶聚类系数, 这表明高阶聚类系数这一拓扑性质能够帮助比较不同结构的网络. 综上所述, 与其他方法相比, 基于高阶信息的网络比较方法能够很好地区分由不同参数生成的 ER (或 WS, BA) 网络.

图 5 给出了 4 种由模型生成的人工合成网络上的进一步研究, 对比了提出的基于高阶信息的网络比较方法与其他基线方法在网络相似性方面的表现. 这 4 种网络分别是 K-regular 网络、WSC (通过 1% 重连边的 K-regular 网络生成得到)、WSK (通过 10% 重连边的 K-regular 网络生成得到) 和 BA 网络 (每一步的加边数 $m = 5$), 其中网络规模统一设置为 $N = 1000$, $|E| = 5000$, 平均度为 10. 由 4 种网络的生成方式可知, K-regular 网络与其他 3 种网络的相似性由高到低分别是 WSC, WSK 和 BA, 而本文提出的基于高阶信息的方法 $D_{\rm HC}$ 结果符合这一结论. 此外, $D_{\rm HC}$ 在比较不同模型生成的网络相似性方面表现出明显的效果差异, 并且具有较小的误差范围; 而 $D_{\rm SP}$ 方法在 K-regular 网络和 WSC (WSK)、K-regular 网络和 BA 网络比较中结果非常相近, 表明该方法不能有效区分这些网络之间的相似性; D_C 方法在 4 种网络比较上大体趋势上正确, 但总体效果不够明显; D_M 则在 K-regular 网络和 WSC 网络以及 K-regular 网络和 WSK 网络比较中得出的相似性与实际情况相反, 且实验结果存在较大的误差和波动, 缺乏稳定性.

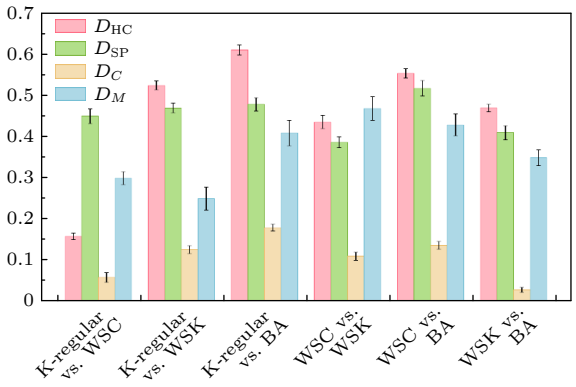


图 5 分别使用 $D_{\rm HC}$, $D_{\rm SP}$, D_C 和 D_M 这 4 种方法对 4 种人工合成网络 (K-regular, WSC, WSK 和 BA) 进行相互比较. 所有的结果均基于 100 次实验的平均值

Fig. 5. Comparison of the four synthetic networks, i.e., K-regular, WSC, WSK, and BA, by using four methods of $D_{\rm HC}$, $D_{\rm SP}$, D_C and D_M. All results are based on an average of 100 realizations.

3.2 真实网络比较

进一步在来自多个领域的 12 个真实网络上评估了基于高阶信息的网络比较方法的有效性. 这 12 个真实数据集的具体描述如下: Chesapeake 网络是美国河口切萨皮克湾的中盐营养网络, 节点表示一组生物体, 例如浮游植物或纤毛虫, 边表示碳交换; Windsurfers 网络描述了 1986 年秋季南加州风帆冲浪者之间的人际交往网络; Contiguous 网络描述了美国的 48 个相邻州和哥伦比亚特区, 边表示两个州共享边界; Jazz 网络是爵士音乐家之间的合作网络; Infectious 网络是 2009 年在都柏林科学展览馆举办的“INFECTIOUS: STAY AWAY”展览期间的一个面对面的接触网络; Yeast 和 Metabolic 网络分别是蛋白质相互作用网络和线虫的代谢网络; Rovira 网络是位于西班牙加泰罗尼亚南部塔拉戈纳的 Rovira i Virgili 大学的一个电子邮件通信网络; Petsterc 和 Petster 网络描述了网站用户之间的友谊和家庭联系网络; Irvine 网络

表 1 真实网络的拓扑结构性质, 其中 N 为节点数, $|E|$ 为边数, Ad 为平均度, Avl 为平均路径长度, Ld 为网络密度, C 为聚类系数, d 为直径

Table 1. Topology properties of the real networks, where N is the number of nodes, $|E|$ is the number of edges, Ad is the average degree, Avl is the average path length, Ld is the network link density, and C is the clustering coefficient, and d is the diameter.

Networks	N	$\|E\|$	Ad	Avl	Ld	C	d
Chesapeake	39	170	8.72	1.83	0.2294	0.450	3
Windsurfers	43	336	15.63	1.69	0.3721	0.653	3
Contiguous	49	107	4.37	4.16	0.0910	0.497	11
Jazz	198	2742	27.69	2.24	0.1406	0.617	6
Infectious	410	2765	13.49	3.63	0.0330	0.456	9
Metabolic	453	2025	8.94	2.68	0.0198	0.646	7
Rovira	1133	5451	9.62	3.61	0.0085	0.220	8
Petster	1858	12534	13.49	3.45	0.0073	0.141	14
Yeast	1870	2203	2.44	6.81	0.0013	0.067	19
Irvine	1899	59835	14.57	3.06	0.0079	0.109	8
Petsterc	2426	16631	13.71	3.59	0.0057	0.538	10
Pgp	10680	24316	4.55	7.49	0.0004	0.266	24

是加州大学 Irvine 分校学生在线社区用户之间的消息传递网络; Pgp 网络是一种基于 Pretty Good Privacy (Pgp) 算法的用户交互网络. 表 1 展示了这些数据集的全部拓扑结构信息, 包括了节点数 N、边数 ($|E|$) 、平均度 (Ad) 、平均路径长度 (Avl) 、网络密度 (Ld) 、聚类系数 (C) 和直径 (d) .

图 6 给出了 12 个真实网络与对应的零模型之间的相似性. 对于一个特定的网络, 本文考虑了 3 种不同的 k 阶零模型, 分别记作 $Dk1.0$, $Dk2.0$ 和 $Dk2.5$ [8]. 其中, 零模型的 k 值表示网络拓扑结构保留的程度. 具体来说, 当 $k=1.0$ 时, 生成的网络保留了原始网络的度序列; 当 $k=2.0$ 时, 在重连过程中保持了度序列和度相关性属性不变; 当 $k=2.5$ 时, 则保留了原始网络的聚类谱属性. 也就是说, k 越大保留的原始网络的性质越多, 即零模型与原始网络越相近. 如图 6 所示, 本文基于高阶信息的网络比较方法 D_{HC} 在不同的网络上比较了原始网络与其零模型的相似性, 结果表明随着 k 值的增加, 真实网络与其随机化网络之间的相似性趋向于变大, 即 k 越大表明随机化后的网络与原始网络共享的拓扑属性越多. 这说明本文的方法能够很好地区分原始网络和零模型之间的不同.

图 7 给出对真实网络以及经过扰动后的网络进行比较的过程. 扰动的过程如下: 对于给定的网络, 随机添加或删除一定比例的边 $f\in[-0.9,0.9]$, 然后比较原网络与扰动后网络之间的相似性. 其中, 正值的 f 表示添加边的过程, 负值的 f 表示删除边的过程. 直观上来讲, 网络扰动程度越大 ($|f|$ 越大), 原网络与扰动后网络的相似性就越小. 实验结果显示 D_{HC} 在删边和加边的过程中都能保持良好的上升趋势, 即 $|f|$ 越大, 原网络与扰动后的网络相似性越小. 相比之下, 从图 7 可以看出, D_{SP}, D_C 和 D_M 这 3 种方法只在删边时能较好地反映网络结构变化, 在添加边时效果欠佳. 例如, 在网络 Chesapeake 上, 随着 $|f|$ 的增大, D_{SP}, D_C 和 D_M 的值变化都很小, 也就是说在这 3 种方法下, 无论

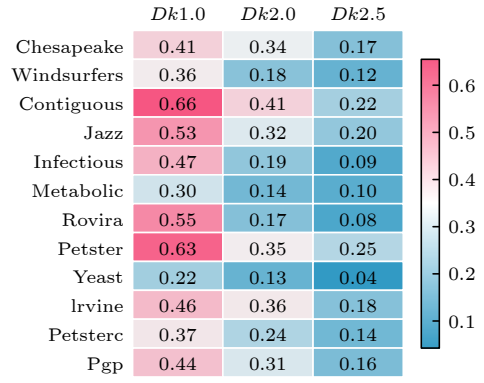


图 6 真实网络与其零模型生成的网络相似性. 考虑了具有不同 k 值 (1.0, 2.0 和 2.5) 的 Dk 零模型, 图中的值表示 D_{HC} 的值的大小. 所有的结果均基于 100 次实验的平均值

Fig. 6. Similarity between real networks and their null-models. We considered the Dk null model with different values k (1.0, 2.0, and 2.5), and the values in the figure indicate the value of D_{HC}. All results are based on an average of 100 realizations.

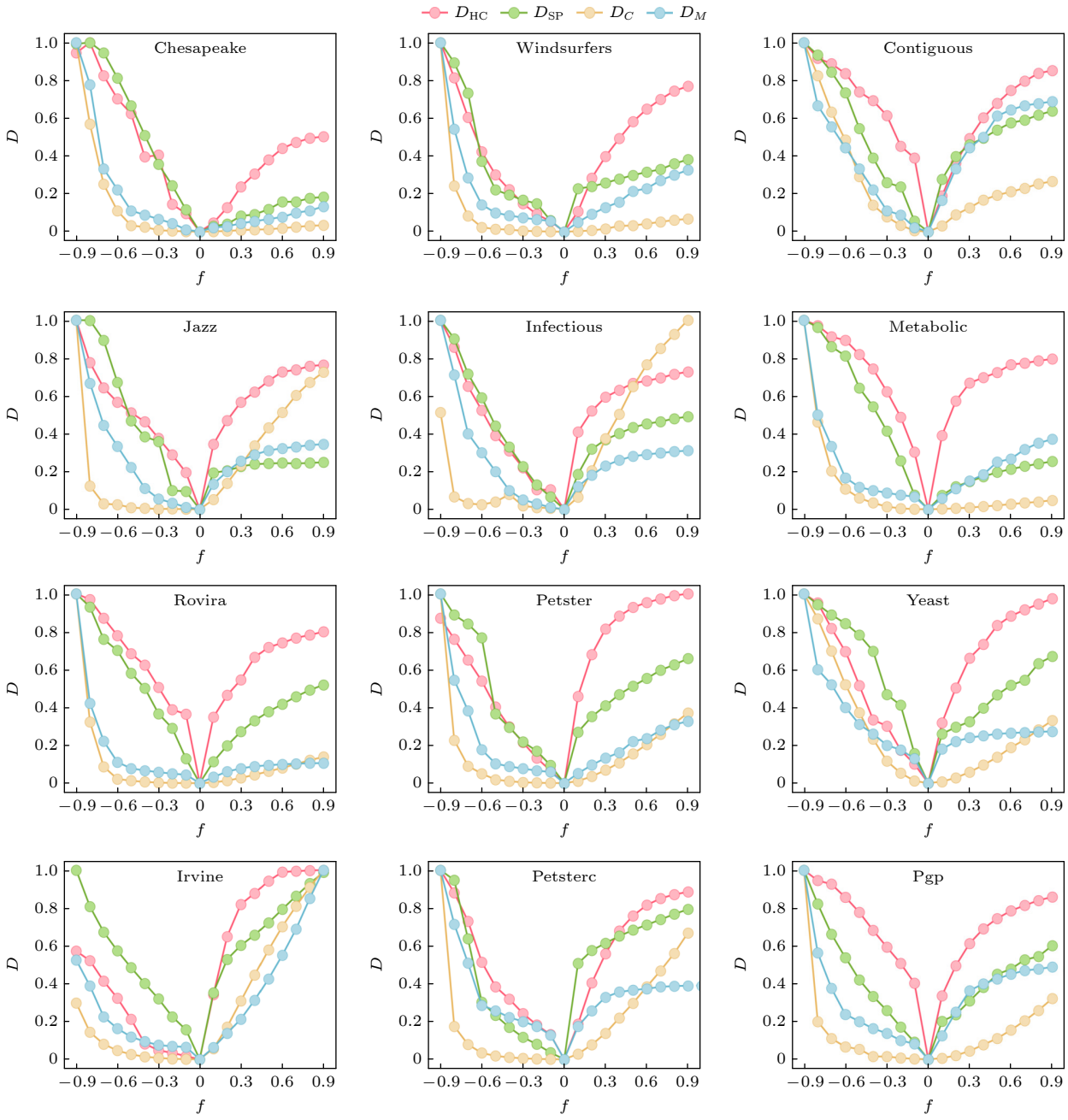


图 7 原始真实网络和经过扰动后生成的网络之间的相似性, 其中 f 的负值对应于给定比例的边的随机删除过程, 正值表示随机增边的过程. 所有的结果均基于 100 次实验的平均值.

Fig. 7. Similarity between the original real network and the network after perturbation, where negative values of f correspond to the deletion of $|f|$ fraction of edges and positive values of f indicate the addition of f fraction of edges. All results are based on an average of 100 realizations.

增加多少边, 原网络与增边后的网络相差不大, 这与事实不符. 此外, 在其他网络上, 与 D_{HC} 相比, 这 3 种方法都或多或少不能合理地表示出网络的相似性. 主要原因在于, D_M 主要依赖于邻接矩阵和度矩阵, 而忽略了高阶结构信息; D_{SP} 则无法同时考虑网络的全局结构特征和局部结构特征. 综合来说, 基于高阶信息的网络比较方法 D_{HC} 在网络扰动中展现出良好的性能, 能够更全面地反映网络的结构变化.

4 总结与展望

本文提出了一种利用基于高阶信息的网络相似性比较方法 D_{HC}, 来弥补现有网络相似性比较算法的不足. 具体而言, 首先计算每个节点的高阶聚

类系数分布和节点间距离分布, 然后使用分布的 Jensen-Shannon 散度来度量网络距离分布的异质性, 从而得到不同网络之间的相似性.

D_{HC} 在不同人工合成网络 (ER, WS, BA) 上均展示了良好的性能, 并在 12 个具有不同拓扑性质的真实网络扰动上也保持了稳定性. 此外, 将该方法与网络比较领域内一些具有代表性的基线算法 (D_{SP} , D_C , D_M) 分别在人工合成网络和真实网络上进行了性能对比, 也取得了令人满意的结果. 未来, 我们也希望将其扩展到更多类型的网络, 如有向网络[38]、超网络[39] 和时序网络[40] 等.

参考文献

[1] Gursoy A, Keskin O, Nussinov R 2008 *Biochem. Soc. Trans.* **36** 1398
[2] Cheng X, Scherpen J M A 2021 *Annu. Rev. Control Robot. Auton. Syst.* **4** 425
[3] Dorogovtsev S N, Mendes J F F 2002 *Adv. Phys.* **51** 1079
[4] Goh K I, Cusick M E, Valle D, Childs B, Vidal M, Barabási A L 2007 *Proc. Natl. Acad. Sci. USA* **104** 8685
[5] Liu C, Ma Y F, Zhao J, Nussinov R, Zhang Y C, Cheng F X, Zhang Z K 2020 *Phys. Rep.* **846** 1
[6] Woolley S M, Posada D, Crandall K A 2008 *PLoS One* **3** e1913
[7] Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang D U 2006 *Phys. Rep.* **424** 175
[8] Orsini C, Dankulov M M, Colomer-de-Simón P, Jamakovic A, Mahadevan P, Vahdat A, Krioukov D 2015 *Nat. Commun.* **6** 8627
[9] Tantardini M, Ieva F, Tajoli L, Piccardi C 2019 *Sci. Rep.* **9** 17557
[10] Zhou X, Zhang F M, Li K W, Hui X B, Wu H S 2012 *Acta Phys. Sin.* **61** 190201 (in Chinese) [周漩, 张凤鸣, 李克武, 惠晓滨, 吴虎胜 2012 物理学报 **61** 190201]
[11] Liu J G, Ren Z M, Guo Q, Wang B H 2013 *Acta Phys. Sin.* **62** 178901 (in Chinese) [刘建国, 任卓明, 郭强, 汪秉宏 2013 物理学报 **62** 178901]
[12] Bracken C P, Scott H S, Goodall G J A 2016 *Nat. Rev. Genet.* **17** 719
[13] Pržulj N 2007 *Bioinformatics* **23** e177
[14] Rong H G, Huo S X, Hu C H, Mo J X 2014 *J. Commun.* **35** 2 (in Chinese) [荣辉桂, 火生旭, 胡春华, 莫进侠 2014 通信学报 **35** 2]
[15] Zemlyachenko V N, Korneenko N M, Tyshkevich R I 1985 *J. Sov. Math.* **29** 1426
[16] Grohe M, Schweitzer P 2020 *Commun. ACM* **63** 128
[17] Caetano T S, McAuley J J, Cheng L, Le Q V, Smola A J 2009 *IEEE Trans. Pattern Anal. Mach. Intell.* **31** 1048
[18] Klau G W 2009 *BMC Bioinf.* **10** S59
[19] Lischka J, Karl H 2009 *Proceedings of the 1st ACM Workshop on Virtualized Infrastructure Systems and Architectures* Barcelona, 17 August, 2009 p81
[20] Yang B, Liu D Y, Jin D, Ma H B 2009 *J. Softw.* **20** 54 (in Chinese) [杨博, 刘大有, 金弟, 马海宾 2009 软件学报 **20** 54]
[21] Aliakbary S, Motallebi S, Rashidian S, Habibi J, Movaghar A 2015 *Chaos* **25** 023111
[22] Liu X, Yi D Y 2011 *Acta Anat. Sin.* **37** 1520 (in Chinese) [刘旭, 易东云 2011 自动化学报 **37** 1520]
[23] Nascimento M C, De Carvalho A C 2011 *Eur. J. Oper. Res.* **211** 221
[24] Wilson R C, Zhu P 2008 *Pattern Recognit* **41** 2833
[25] Wang Z P, Zhan X X, Liu C, Zhang Z K 2022 *iScience* **25** 104446
[26] Wang X F, Liu Y B 2009 *J. Univ. Electron. Sci. Technol. China.* **38** 537 (in Chinese) [汪小帆, 刘亚冰 2009 电子科技大学学报 **38** 537]
[27] Lv L Y 2010 *J. Univ. Electron. Sci. Technol. China.* **39** 651 (in Chinese) [吕琳媛 2010 电子科技大学学报 **39** 651]
[28] Koutra D, Vogelstein J T, Faloutsos C 2013 *Proceedings of the 2013 SIAM International Conference on Data Mining (SDM)* Austin, May, 2013 p162
[29] De Domenico M, Biamonte J 2016 *Phys. Rev. X* **6** 041062
[30] Schieber T A, Carpi L, Díaz-Guilera A, Pardalos P M, Masoller C, Ravetti M G 2017 *Nat. Commun.* **8** 13928
[31] Chen D, Shi D D, Qin M, Xu S M, Pan G J 2018 *Phys. Rev. E* **98** 012319
[32] Liu Q, Dong Z, Wang E 2018 *Sci. Rep.* **8** 5134
[33] Deng X L, Wang B, Wu B, Yang S Q 2012 *J. Comput. Res. Dev.* **49** 725 (in Chinese) [邓小龙, 王柏, 吴斌, 杨胜琦 2012 计算机研究与发展 **49** 725]
[34] Menéndez M L, Pardo J A, Pardo L 1997 *J. Franklin Inst.* **334** 307
[35] Fronczak A, Hołyst J A, Jedynak M, Sienkiewicz J 2002 *Physica A* **316** 688
[36] Wang L, Dai G Z 2005 *Sci. & Tech. Rev.* **23** 62 (in Chinese) [王林, 戴冠中 2005 科技导报 **23** 62]
[37] Zager L A, Verghese G C 2008 *Appl. Math. Lett.* **21** 86
[38] Sarajlić A, Malod-Dognin N, Yaveroğlu Ö N, Pržulj N 2016 *Sci. Rep.* **6** 35098
[39] Wang L, Egorova E K, Mokryakov A V 2018 *J. Comput. Syst. Sci. Int.* **57** 109
[40] Holme P, Saramäki J 2012 *Phys. Rep.* **519** 97

Network similarity comparison method based on higher-order information*

Chen Hao-Yu [1)] Xu Tao [1)] Liu Chuang [1)] Zhang Zi-Ke [2)3)] Zhan Xiu-Xiu [1)3)†]

1) (*Complex Science Research Center*, *Hangzhou Normal University*, *Hangzhou* 311121, *China*)

2) (*Digital Communication Research Center*, *Zhejiang University*, *Hangzhou* 310058, *China*)

3) (*School of Media and International Culture*, *Zhejiang University*, *Hangzhou* 310058, *China*)

(Received 5 July 2023; revised manuscript received 6 October 2023)

Abstract

Quantifying structural similarity between complex networks presents a fundamental and formidable challenge in network science, which plays a crucial role in various fields, such as bioinformatics, social science, and economics, and serves as an effective method for network classification, temporal network evolution, network generated model evaluation, etc. Traditional network comparison methods often rely on simplistic structural properties such as node degree and network distance. However, these methods only consider the local or global aspect of a network, leading to inaccuracies in network similarity assessments. In this study, we introduce a network similarity comparison method based on the high-order structure. This innovative approach takes into account the global and the local structure of a network, resulting in a more comprehensive and accurate quantification of the network difference. Specifically, we construct distributions of higher-order clustering coefficient and distance between nodes in a network. The Jensen-Shannon divergence, based on these two distributions, is used to quantitatively measure the similarity between two networks, offering a more refined and robust measure of network similarity. To validate the effectiveness of our proposed method, we conduct a series of comprehensive experiments on the artificial and the real-world network, spanning various domains and applications. By meticulously fine-tuning the parameters related to three different artificial network generation models, we systematically compare the performances of our method under various parameter settings in the same network. In addition, we generate four different network models with varying levels of randomization, creating a diverse set of test cases to evaluate the robustness and adaptability of the method. In artificial networks, we rigorously compare our proposed method with other baseline techniques, consistently demonstrating its superior accuracy and stability through experimental results; in real networks, we select datasets from diverse domains and confirm the reliability of our method by conducting extensive similarity assessments between real networks and their perturbed reconstructed counterparts. Furthermore, in real networks, the rigorous comparison between our method and null models underscores its robustness and stability across a broad spectrum of scenarios and applications. Finally, a meticulous sensitivity analysis of the parameters reveals that our method exhibits remarkable performance consistency across networks of different types, scales, and complexities.

Keywords: network similarity, higher-order clustering coefficient, distance distribution

PACS: 89.75.–k, 89.75.Fb **DOI:** 10.7498/aps.73.20231096

* Project supported by the National Natural Science Foundation of China (Grant Nos. 72371224, 92146001), the Natural Science Foundation of Zhejiang Province, China (Grant No. LQ22F030008), the Fundamental Research Fund for the Central Universities, China, the Scientific Research Foundation for Scholars of Hangzhou Normal University, China (Grant No. 2021QDL030), and the Bingtuan Science and Technology Program, China (Grant No. 2021AB034).

† Corresponding author. E-mail: zhanxiuxiu@hznu.edu.cn

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

基于高阶信息的网络相似性比较方法

陈浩宇 徐涛 刘闯 张子柯 詹秀秀

Network similarity comparison method based on higher-order information

Chen Hao-Yu Xu Tao Liu Chuang Zhang Zi-Ke Zhan Xiu-Xiu

引用信息 Citation: Acta Physica Sinica, 73, 038901 (2024) DOI: 10.7498/aps.73.20231096
在线阅读 View online: https://doi.org/10.7498/aps.73.20231096
当期内容 View table of contents: http://wulixb.iphy.ac.cn
