

基于广义共同邻居的有向网络链路预测方法

赵学磊, 季新生, 刘树新, 赵宇

(信息工程大学, 河南 郑州 450001)

摘 要: 链路预测利用已知网络节点及结构等信息预测网络未知连接或未来连接。现有主流方法多应用于无向网络, 而且部分有向链路预测方法忽视了共同邻居在有向网络下的多样异构特征。针对上述问题, 提出了一种广义共同邻居算法。该方法首先对有向网络定义了广义共同邻居, 通过网络中有向邻居异构体的连边概率衡量不同结构对连边贡献程度, 然后利用该定义对现有的局部相似性指标进行了优化, 重定义 8 种基于广义共同邻居的有向相似性指标。在 12 个数据集上的实验表明, 所提方法在两个衡量指标下普遍提升了现有预测指标性能。

关键词: 有向网络; 链路预测; 共同邻居; 异构融合

中图分类号: TP301.6

文献标识码: A

doi: 10.11959/j.issn.2096-109x.2020059

Link prediction methods based on generalized common neighbor in directed network

ZHAO Xuelei, JI Xinsheng, LIU Shuxin, ZHAO Yu

Information Engineering University, Zhengzhou 450001, China

Abstract: Link prediction aims to predict missing or future links through currently observed information of network. Existing mainstream methods are mostly applied to undirected network, and some methods designed for directed network ignored the diverse heterogeneous features of common neighbor. For this problem, a generalized common neighbor algorithm was proposed. Firstly, a generalized common neighbor was defined for the directed network. Then the degree of contribution of different structures was measured by the joint edge probability of the directed neighbor isomers, and the existing undirected local similarity index is improved by the new definition, redefining eight kinds of directed similarity indicators based on generalized common neighbor. Experiments on 12 datasets show that proposed method generally improves the performance of existing predictive indicators under two metrics.

Key words: directed network, link prediction, common neighbor, heterogeneous fusion

收稿日期: 2019-09-03; 修回日期: 2020-02-13

通信作者: 赵学磊, 1375565768@qq.com

基金项目: 国家自然科学基金 (61803384)

Foundation Item: The National Natural Science Foundation of China (61803384)

论文引用格式: 赵学磊, 季新生, 刘树新, 等. 基于广义共同邻居的有向网络链路预测方法[J]. 网络与信息安全学报, 2020, 6(5): 89-100.

ZHAO X L, JI X S, LIU S X, et al. Link prediction methods based on generalized common neighbor in directed network [J]. Chinese Journal of Network and Information Security, 2020, 6(5): 89-100.

1 引言

近年来, 复杂网络逐渐成为探究真实复杂系统内在机理的抽象化工具, 其重要方向链路预测^[1-4]得到了研究人员的持续关注。链路预测旨在利用网络中已知信息预测网络中的未知连接^[5]、未来连接^[6]或错误连接^[7], 是复杂网络与信息科学交叉融合的重要桥梁, 对研究网络演化^[8-11]、未知链路探测^[12]等具有巨大的实际应用价值。

当前, 针对无向网络的链路预测研究已取得众多成果, 吕琳媛^[3-4]详细对比分析了各种方法, 并将其归类为局部相似性、全局相似性指标。局部相似性指标优势在于计算复杂度低, 适用大规模网络。全局相似性指标考虑全网拓扑信息预测精度更高, 但会付出较高计算代价。然而, 前期研究集中于无向网络, 并且现有局部或全局指标均难以直接应用于有向网络的预测。

局部相似性指标以共同邻居^[13](CN, common neighbor) 指标最具代表性, 因有向网络中连边具有方向性, 使无向共同邻居衍生多种异构形式, 传统预测方案忽略不同异构形式之间差异, 极大降低了有向网络中的预测精度, 甚至使预测失去意义。文献[14]研究发现, 有向网络存在特殊的互惠机制, 即节点倾向于建立反向的连接以回应其他节点对自己的连接, 故有向网络中存在部分双向的链接, 称之为互惠边, 互惠边在促进连边产生时通常比单向连边发挥更大作用, 表示节点间具有更为紧密的局部结构, 而现有指标仅利用网络单向连边进行预测。文献[15]则是通过对网络中不同邻居结构下的连边概率进行统计并作为相似分值, 在预测精度上取得了不错的效果。Valverde 等^[16]利用社交平衡理论研究证实了繁杂的人际关系将导致社交群体的不同演化方向等。张扬夫^[17]曾受益于网络特性分析, 提出一般化相似度指标, 在共同邻居基础上提升了预测性能。然而, 上述理论或方法仍无法使现有指标普遍适用于有向网络, 缺少对局部邻居异构多样性的量化区分, 且目前算法性能对比基准多采用无向网络局部相似性指标, 该类指标中共同邻居含义在有向网络中已然发生改变。因此, 有向网络的链

路预测应当利用网络中多样的局部邻居异构形式进行预测。

针对上述问题, 本文提出基于广义共同邻居的链路预测算法, 通过分析网络中共同邻居有向异构体的具体作用, 量化不同结构对连边产生的贡献程度, 并结合连边概率将 9 种局部拓扑异构体进行融合, 得到广义上的共同邻居计算方式, 并基于此, 针对有向网络重定义 8 种基于 CN 的相似性指标, 使其有效利用共同邻居中有向连边信息并应用于有向网络。在 12 个真实数据集下实验对比, 在 AUC 及 Ranking Score 衡量标准下所提方案大幅提升了预测准确度, 验证了所提方案的有效性。

2 相关工作介绍

经典的 CN 算法认为两个节点间若存在更多共同邻居, 则两节点更倾向于产生连边。若引入两节点度的影响, 可从不同的角度衍生其他相似性指标, 具体包括 Salton^[18]指标、Jaccard^[19]指标、Sørensen^[20]指标、HPI^[21]指标、HDI^[22]指标、LHN-I^[23]指标。另外, 考虑共同邻居节点的度时有 Adamic-Adar^[24](AA) 指标、资源分配^[25](RA, resource allocation) 指标。在文献[26]中, 杨瑞琪等将上述指标做了有向的变体使其尽可能适用于有向网络, 一定程度上提高了在有向网络中的预测准确性, 本节对其改进后的有向指标进行描述, 同时为与其他较优的有向算法仿真对比, 还列出 PA^[27]、LP^[28]、Katz^[29]这 3 种主流指标。现对相关指标及其有向方式的定义简介如下。

CN 指标: 以节点间共同邻居数目衡量连边可能性, 表示为

$$S_{xy} = |\Gamma_{out}(x) \cap \Gamma_{in}(y)| \quad (1)$$

其中, $\Gamma_{out}(x)$ 表示节点 x 的出度邻居节点, $\Gamma_{in}(y)$ 表示节点 y 的入度邻居节点。

Salton 指标: 在共同邻居基础上引入端节点度信息, 与端节点度之积开方成反比, 表示为

$$S_{xy} = \frac{|\Gamma_{out}(x) \cap \Gamma_{in}(y)|}{\sqrt{k_{out}(x) \cdot k_{in}(y)}} \quad (2)$$

Jaccard 指标: 两节点邻居集合中共同邻居的占比, 表示为

$$S_{xy} = \frac{|\Gamma_{\text{out}}(x) \cap \Gamma_{\text{in}}(y)|}{|\Gamma_{\text{out}}(x) \cup \Gamma_{\text{in}}(y)|} \quad (3)$$

Sørensen: 共同邻居与端节点的度之和之比, 表示为

$$S_{xy} = \frac{2|\Gamma_{\text{out}}(x) \cap \Gamma_{\text{in}}(y)|}{k_{\text{out}}(x) + k_{\text{in}}(y)} \quad (4)$$

HPI 指标: 共同邻居与端节点最小度的比值, 表示为

$$S_{xy} = \frac{|\Gamma_{\text{out}}(x) \cap \Gamma_{\text{in}}(y)|}{\min\{k_{\text{out}}(x), k_{\text{in}}(y)\}} \quad (5)$$

HDI 指标: 与 HPI 定义相似, 表示为

$$S_{xy} = \frac{|\Gamma_{\text{out}}(x) \cap \Gamma_{\text{in}}(y)|}{\max\{k_{\text{out}}(x), k_{\text{in}}(y)\}} \quad (6)$$

LHN-I 指标: 与 Salton 定义相似, 分母为端节点度之积 (相比之下没有开根号), 表示为

$$S_{xy} = \frac{|\Gamma_{\text{out}}(x) \cap \Gamma_{\text{in}}(y)|}{k_{\text{out}}(x) \cdot k_{\text{in}}(y)} \quad (7)$$

AA 指标: 考虑共同邻居节点度信息为共同邻居节点赋予权重, 节点权重为节点度的对数分之一 (度越大权重越小), 表示为

$$S_{xy} = \sum_{z \in \Gamma_{\text{out}}(x) \cap \Gamma_{\text{in}}(y)} \frac{1}{\log k_{\text{out}}(z)} \quad (8)$$

其中, $k_{\text{out}}(z)$ 表示共同邻居 z 的出度。

RA 指标: 启发于资源分配过程, 将共同邻居视为传递媒介, 传递资源量与该节点的度成反比, 表示为

$$S_{xy} = \sum_{z \in \Gamma_{\text{out}}(x) \cap \Gamma_{\text{in}}(y)} \frac{1}{k_{\text{out}}(z)} \quad (9)$$

PA 指标: 节点与连边概率正比于二者度之积, 表示为

$$S_{xy} = k_{\text{out}}(x) \cdot k_{\text{in}}(y) \quad (10)$$

LP 指标: 在二阶共同邻居上考虑了三阶路径因素, 表示为

$$S_{xy} = A^2 + \alpha \cdot A^3 \quad (11)$$

Katz 指标: 全局考虑网络所有路径, 表示为

$$S_{xy} = \sum_{l=1}^{\infty} \alpha^l \cdot |\text{paths}_{x,y}^{<l>}| = \alpha A_{xy} + \alpha^2 (A^2)_{xy} + \alpha^3 (A^3)_{xy} + \cdots + \alpha^n (A^n)_{xy} \quad (12)$$

3 基于广义共同邻居的相似性算法

现有基于共同邻居的相似性指标在有向网络应用时, 其共同邻居均定义为 $z \in \Gamma_{\text{out}}(x) \cap \Gamma_{\text{in}}(y)$, 即节点对之间通过 $x \rightarrow z \rightarrow y$ 的单向路径连接的邻居节点。而有向网络的局部结构远复杂于无向网络, 连边方向的改变或互惠边的存在将使局部结构衍生多种异构形式。该共同邻居定义方式却缺失多样的有向异构连边信息, 导致基于共同邻居的预测指标在有向网络预测普遍精度较低。潘永昊等^[30]研究了网络内在结构的动力学特征, 认为不同结构对新链路的产生效果不同。常圣等^[31]研究发现, 应用势理论筛选出有向网络中不包含互惠边的 4 种可定义势子图可显著提高有向链路预测的精度。

因此, 共同邻居需面向有向网络进行重新定义, 融合有向连边异构的多种局部拓扑对连边的贡献。

3.1 有向异构下的广义共同邻居融合表示

有向网络复杂性在于, 节点 x 与 y 可通过出边、入边或互惠边的形式连接到共同邻居节点 z , 如图 1 所示, S1 结构即节点 x 与 y 均经由互惠边连接到邻居节点 z 。而 $z \in \Gamma_{\text{out}}(x) \cap \Gamma_{\text{in}}(y)$ 仅是 S9 结构的单一表示, 缺少对其他异构形式的充分利用。该定义中, $\Gamma_{\text{out}}(x)$ 表示 x 的出度邻居节点, 可认为在 S3、S8、S9 结构中 $x \rightarrow z$ 的指向性节点对, $\Gamma_{\text{in}}(y)$ 表示 y 的入度邻居节点, 即 S5、S7、S9 结构中 $z \rightarrow y$ 的指向性节点对。如此, 其余 8 种异构形式的共同邻居均可通过改变连边指向进行描述, 现将广义共同邻居 (GMCN, generalized mutual common neighbor) 定义如下。

定义 1 广义共同邻居: 给定有向网络 $G(V, E)$, 其中, V 代表网络节点集合, E 代表网络中连边集合, 该有向网络中两节点 x 与 y 的广义共同邻居定义如式 (13) 所示。

$$z \in \Gamma_i(x) \cap \Gamma_j(y), i, j \in \{\text{out}, \text{in}, \text{bila}\} \quad (13)$$

其中, $\Gamma_{\text{bila}}(x)$ 表示与节点 x 通过互惠边连接的邻居集合。

通过广义共同邻居定义方式, 多样的有向异构形式均得到统一表述, 保证了局部信息的充分

利用。然而, 将其应用预测算法时 9 类异构形式并非简单的数目作和。Brzozowski 等^[32]曾在社交网站 WaterCooler 研究发现, 好友关注关系网络下, S7 结构的数量最多, 但产生连边比例最低; S1 结构的数量最少, 但连边比例最高, 体现了互惠性在一些社交网络演化中的重要作用。但食物链网络中该现象却几乎不存在, 而 S9 结构产生连边的比例最高, 更加符合自然生物的层级捕食现象。

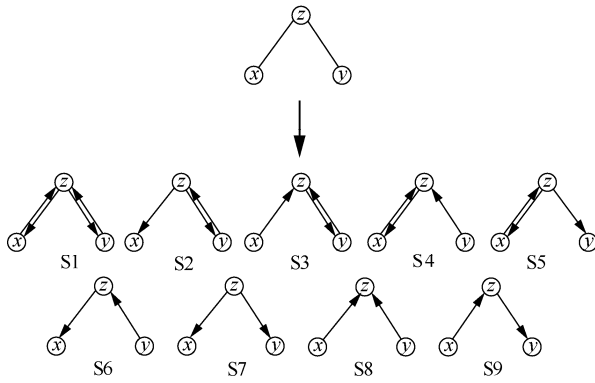


图 1 共同邻居结构的有向异构形式
Figure 1 Directed heterogeneous form of common neighbor structure

以 *C.elegans* 数据集为例进行具体分析, 首先统计该数据集中上述 9 类结构数目, 其次统计各结构下产生连边 $x \rightarrow y$ 的数目, 计算得到各异构体独自的连边率, 如表 1 所示。

表 1 *C.elegans* 数据集不同异构体数目及连边率统计
Table 1 Statistics of different isomers and edge rate of *C.elegans*

结构编号	结构数目	连边数目	连边率
S1	1 022	244	23.90%
S2	4 239	327	7.70%
S3	3546	772	21.80%
S4	3 546	327	9.20%
S5	4 239	1 232	29.10%
S6	15 396	395	2.60%
S7	20 438	2 596	12.70%
S8	39 786	3 056	7.70%
S9	15 396	2 151	14.00%

该数据集中, 结构 S8 的数目最多为 39 786, 基于该种结构的连边 $x \rightarrow y$ 数目为 3 056 条, 连边率仅 7.70%; 而结构 S1 总数目仅 1 022, 其连

边数目却可达到 244, 连边率为 23.90%。另外, 结构 S6 与 S9 具有一致数目, 二者连边率却相差极大, S9 连边率已接近于 S6 的 6 倍, 显然二者对连边产生的促进作用不等, 该差异性难以在简单统计数目作和的计算方式下有效区分。且 S6 结构表现出一种“反向层级”现象, 恰与原 CN 指标相反, 这甚至会抑制新连边的产生, 数目越大抑制作用越大。因此, 结构数目并非连边产生的决定因素, 其连边概率具有更加可信的参考价值。

对有向网络中共同邻居进行一般化定义后, 通过上述连边率的分析, 得出不同结构促进连边产生的贡献程度存在差异, 因此将各个异构形式的连边概率作为其贡献程度权值, 通过该贡献程度权值将各异构体进行融合计算, 基于此将广义共同邻居算法 (GMCN) 定义如下。

定义 2 广义共同邻居算法: 对于给定的有向网络 $G(V, E)$, 网络中任意两节点 x 与 y 之间的连边可能性计算为

$$S_{xy} = \sum p_k |\Gamma_i(x) \cap \Gamma_j(y)|, \quad i, j \in \{\text{out}, \text{in}, \text{bila}\}, \quad k = 1, 2, \dots, 9 \quad (14)$$

其中, p_k 表示 9 类结构在给定网络 $G(V, E)$ 下的连边概率, p_k 计算方式定义如下。

$$p_k = \frac{|\Gamma_j(\Gamma_i(x)) \cap \Gamma_{\text{out}}(x)|}{|\Gamma_i(x) \cap \Gamma_j(y)|}, \quad i, j \in \{\text{out}, \text{in}, \text{bila}\}, \quad k = 1, 2, \dots, 9 \quad (15)$$

$\Gamma_j(\Gamma_i(x))$ 表示节点 x 的二阶广义邻居, $\Gamma_j(\Gamma_i(x)) \cap \Gamma_{\text{out}}(x)$ 表示节点 x 与二阶广义邻居直接相连的数目。

3.2 基于广义共同邻居的相似性算法

广义共同邻居的重新定义, 有效涵盖了有向网络连边的多样异构形式, 而现有指标中存在的覆盖不足问题得以解决。将广义共同邻居定义应用于现有基于共同邻居的相似性指标, 重新定义基于广义共同邻居的相似性指标, 将各指标重新表示如下。

GMJaccard: 对 Jaccard 指标的重定义, 将多样连边形式的邻居集合统一表述, 量化区分, 具体计算为

$$S_{xy} = \sum_k p_k \frac{|\Gamma_i(x) \cap \Gamma_j(x)|}{|\Gamma_i(x) \cup \Gamma_j(x)|}, \quad i, j \in \{\text{out}, \text{in}, \text{bila}\},$$

$$k = 1, 2, \dots, 9 \quad (16)$$

GMSalton: Salton 指标的扩展, 除应用广义共同邻居外, 端节点的度同时进行了出入度等不同组合的统一表述, 通过互惠边连接的共同邻居在计算时需采用该节点的出入度之和, 具体计算为

$$S_{xy} = \sum_k p_k \frac{|\Gamma_i(x) \cap \Gamma_j(x)|}{\sqrt{k_i(x) \cdot k_j(y)}}, \quad i, j \in \{\text{out}, \text{in}, \text{bila}\},$$

$$k = 1, 2, \dots, 9 \quad (17)$$

上述公式中 $k_{\text{out}}(x)$ 表示节点 x 的出度, $k_{\text{in}}(x)$ 表示节点 x 的入度, $k_{\text{bila}}(x)$ 表示节点 x 的出入度之和, 下同。

GMSørensen: Sørensen 指标的扩展, 与 GMSalton 的扩展方式相同, 移除原 Sørensen 指标分子中的乘数 2, 表示为

$$S_{xy} = \sum_k p_k \frac{|\Gamma_i(x) \cap \Gamma_j(x)|}{k_i(x) + k_j(y)}, \quad i, j \in \{\text{out}, \text{in}, \text{bila}\},$$

$$k = 1, 2, \dots, 9 \quad (18)$$

GMHPI: HPI 指标的扩展, 表示为

$$S_{xy} = \frac{\sum_k p_k |\Gamma_i(x) \cap \Gamma_j(x)|}{\min\{k_i(x), k_j(y)\}}, \quad i, j \in \{\text{out}, \text{in}, \text{bila}\},$$

$$k = 1, 2, \dots, 9 \quad (19)$$

GMHDI: HDI 指标的扩展, 表示为

$$S_{xy} = \frac{\sum_k p_k |\Gamma_i(x) \cap \Gamma_j(x)|}{\max\{k_i(x), k_j(y)\}}, \quad i, j \in \{\text{out}, \text{in}, \text{bila}\},$$

$$k = 1, 2, \dots, 9 \quad (20)$$

GMLHN: LHN 指标的扩展, 表示为

$$S_{xy} = \sum_k p_k \frac{|\Gamma_i(x) \cap \Gamma_j(x)|}{k_i(x) \cdot k_j(y)}, \quad i, j \in \{\text{out}, \text{in}, \text{bila}\},$$

$$k = 1, 2, \dots, 9 \quad (21)$$

GMAA: AA 指标的扩展, 对广义下的共同邻居的出度进行加权计算, 表示为

$$S_{xy} = \sum_{z \in \Gamma_i(x) \cap \Gamma_j(y)} p_k \frac{1}{\log(k_{\text{out}}(z))},$$

$$i, j \in \{\text{out}, \text{in}, \text{bila}\}, \quad k = 1, 2, \dots, 9 \quad (22)$$

GMRA: 与 GMAA 类似, 是对 RA 指标的扩展, 表示为

$$S_{xy} = \sum_{z \in \Gamma_i(x) \cap \Gamma_j(y)} p_k \frac{1}{k_{\text{out}}(z)}, \quad i, j \in \{\text{out}, \text{in}, \text{bila}\},$$

$$k = 1, 2, \dots, 9 \quad (23)$$

3.3 算法复杂度分析

CN 指标因其简洁性著名, 该指标及基于 CN 的算法计算复杂度为 $O(N^3)$ ^[33]。基于 GMCN 的指标计算程序等同于 9 类有向邻居异构体的加权分值之和, 因 p_k 为各网络数据集的结构特性, 已知的数据集网络连边确定, p_k 为固定值, 仅对未知数据集首次计算时需统计 $S_1 \sim S_9$ 的连边概率, 因此已知网络数据集计算复杂度为 $O(9N^3) = O(N^3)$, 继承了原指标低复杂度的优势。

4 评价指标及数据集介绍

4.1 评价指标

链路预测评价指在公认计算准则下对比不同算法的得分。通常将网络数据集 E 划分为训练集 E^T 及测试集 E^P , $E^T \cup E^P = E$, 且 $E^T \cap E^P \neq \emptyset$ 。以训练集数据预测测试集连边, 以测试集和不存在边集计算评价指标得分, 相似分值越高, 连边概率越大。本文采用 AUC^[34]及 Ranking Score^[35] (RS) 两种主流方式对所提方法进行衡量。

AUC 可简单理解为在测试集中随机选择一条边的分数值比随机选择一条不存在边的分数值高的概率。从测试集 E^T 和不存在边集中各随机选取一条边, 比较二者的分数值, 若测试集中的连边预测分数值高于不存在边集中的预测分数值累加 1 (n' 表示), 若二者相等累加 0.5 (n'' 表示)。AUC 计算定义为

$$\text{AUC} = \frac{n' + 0.5n''}{n} \quad (24)$$

Ranking Score 对所有未连边按照相似分值大小排序, 得到测试集的边在最终排序中的位置, 测试集连边排名越高, 排序分值越小时, 说明该算法具有较好预测性能。以 H 为未连边的集合 (测试集中和不存在边集的集合), 网络中某条测试边的排序分为

$$RS_e = \frac{r_e}{H} \quad (25)$$

整个系统的排序分可通过遍历求得, 计算方式如下。

$$RS = \frac{1}{|E^P|} \sum_{e \in E^P} RS_e = \frac{1}{|E^P|} \sum_{e \in E^P} \frac{r_e}{H} \quad (26)$$

4.2 数据集介绍

为使实验验证具有代表性, 本文选定 12 个有向网络真实数据集, 涵盖社交网络、生物网络、引文网络、航空交通网络。数据集简介如下。

1) Highschool (HS)^[36]: 伊利诺伊州一所小型高中男生之间的友谊关系网络。

2) Residence hall (RH)^[37]: 澳大利亚国立大学校园 217 名居民之间的友谊关系网络。

3) Adolescent health (AH)^[38]: 美国青少年健康关系网络。

4) Physicians (PH)^[39]: 美国 241 名医生之间的相互朋友关系网络。

5) Usairport (UA)^[40]: 2010 年美国机场之间的定向航班网络。

6) Openflight (OF)^[41]: OpenFlights.org 项目收集的航班网络。

7) C.elegans (CE)^[42]: 线虫神经元之间的突触连接关系网络。

8) SciMet (SM)^[43]: 引用“科学计量学”主题的论文引用网络。

9) Kohonen (KH)^[43]: 有关“自组织映射”主题或“Kohonen”的论文引用网络。

10) Wikivote (WV)^[44]: 维基选票网站的投票选举关系网络。

11) Chess (CH)^[45]: 国际象棋 7 301 个玩家之间的 65 053 场对战关系网络。

12) Air traffic control^[45] (AC): 美国 FAA 飞行数据中心记录的规划航线网络。

预测实验时, 设定训练集比例 0.9, 测试集 0.1, 每数据集进行 50 次独立实验, AUC 及 Ranking Score 结果取其 50 次实验的均值。各数据集统计特征如表 2 所示, 包括网络节点规模 $|V|$ 、连边数目 $|E|$ 、最大出度 k_{out_max} 、最大入度 k_{in_max} 、平均度 $\langle k \rangle$ 、集聚系数 C 。

表 2 数据集基本特征参数
Table 2 Basic characteristics of datasets

数据集	$ V $	$ E $	k_{out_max}	k_{in_max}	$\langle K \rangle$	C
HS	70	366	18	12	5.23	0.46
RS	217	2 672	34	51	12.31	0.36
AH	2 539	12 969	27	10	5.11	0.15
PH	241	1 098	26	9	4.56	0.31
UA	1 574	28 236	294	302	17.94	0.5
OF	2 939	30 501	236	237	10.38	0.45
CE	297	2 345	134	39	7.9	0.29
SM	3 084	10 412	121	104	3.38	0.15
KH	4 469	12 729	51	735	2.85	0.21
WV	7 115	103 689	457	893	14.57	0.14
CH	7 301	65 001	140	152	8.9	0.18
AC	1 226	2 613	20	24	2.13	0.07

5 实验验证及分析

为了验证方法的有效性, 针对所提基于广义共同邻居的相似性指标, 本节对比了现有指标有向方式的 AUC 及 Ranking Score, 验证该方案整体预测精度的表现。

5.1 AUC 结果及分析

首先, GMCN 指标 AUC 结果如表 3 所示。GMCN 指标的 AUC 在 12 数据集上均有大幅度提升, 在 CH 数据集上相比 CN 指标 AUC 提升幅度最大, 由 0.787 提高至 0.906。整体来看, 社交网络中提升普遍较高, 4 种社交网络数据集的 AUC 平均提升 8.23%, 说明该类数据集中, 错综复杂的局部连边结构中的潜藏信息可有效提高预测精度。PA、LP、Katz 等指标虽在部分网络上表现突出, 但其 AUC 值波动较大, 如 PA 指标的“富者越富”机制在社交类网络中表现很差, AUC 值低于 0.7, 甚至不及 CN 指标。LP 指标考虑三阶路径因素, 仅在航空、引文类网络具有轻微优势; Katz 指标则考虑了全局路径信息, 约在半数网络上略高于广义共同邻居方法, 但其代价是付出极高的计算复杂度。因此, GMCN 指标与 CN 复杂度相当, 但与 PA、LP 相比却具有更高的精度, HS、RH、CE 数据集上, 该类指标甚至达到或高于利用全局路径信息的 Katz 指标的精度。

表 3 AUC 实验结果对比
Table 3 Comparison of AUC results

数据集	CN	GMCN	PA	LP	Katz
HS	0.843	0.893	0.646	0.872	0.878
RH	0.883	0.912	0.666	0.891	0.892
AH	0.717	0.807	0.684	0.779	0.877
PH	0.76	0.863	0.71	0.849	0.933
UA	0.971	0.975	0.953	0.976	0.964
OF	0.972	0.98	0.933	0.986	0.981
CE	0.795	0.882	0.806	0.85	0.867
SM	0.895	0.927	0.956	0.969	0.997
KH	0.922	0.931	0.982	0.982	0.999
WV	0.921	0.95	0.982	0.979	0.985
CH	0.787	0.906	0.886	0.919	0.944
AC	0.608	0.7	0.857	0.699	0.9

其次,将广义共同邻居定义应用现有指标,得到基于 GMCN 的链路预测算法 AUC 结果,如表 4 所示。综合来看,在 12 个数据集上,重定义指标 AUC 值同样有大幅提升,各指标在社交类网络上提升明显,特别在 AH 数据集上 AUC 普遍由 0.7 提升至 0.8 之上,这一现象体现出社交平衡理论中多样的正负关系的确具有重要价值,现有指标仅通过单一的改进无

法使经典指标适用于有向网络,应从结构上区分不同连边的异构体作用,从而有效提高预测精度。

纵向来看,各指标之间差异化并不明显,GMSalton、GMJaccard、GMSørenson、GMHPI、GMHDI、GMLHN 指标与 GMCN 预测精度相当,AUC 值保持在较小波动范围,这也同无向网络中该类指标预测效果相符。另外,采用局部节点度信息的 AA、RA 指标因仅计算到前述 S9 形式的共同邻居,疏漏其余 8 种异构邻居的度信息,除在航空网络有较高预测精度外,其余各类型网络不及 0.9。而重定义优化之后的 GMAA 及 GMRA 则通过引入广义共同邻居,将数个网络的 AUC 值提升至 0.9 以上。

不同数据集中 GM 类指标与原指标 AUC 对比如图 2 所示。这里直观地显示了基于广义共同邻居的指标有效提升了 AUC 预测性能,在多种类型数据集的 AUC 值普遍保持在 0.9 上下,AH、PH 及 AC 数据集上 AUC 值虽较低,但相比原指标,AUC 提升近 10%。GMCN 指标相比 CN 指标提升效果最明显,在上述 12 个数据集上平均 AUC 提升 6.01%,其余几种指标 AUC 值普遍提高 5%左右,GMLHN 指标稍有提升为 0.47%。

表 4 基于 GMCN 的相似性指标 AUC 结果
Table 4 AUC results of similarity index based on GMCN

数据集	Salton	GMSalton	Jaccard	GMJaccard	Sørenson	GMSørenson	HPI	GMHPI	HDI	GMHDI	LHN	GMLHN	AA	GMAA	RA	GMRA
HS	0.845	0.899	0.844	0.888	0.847	0.879	0.843	0.887	0.844	0.888	0.862	0.924	0.845	0.881	0.843	0.901
RH	0.887	0.919	0.885	0.91	0.886	0.904	0.882	0.901	0.882	0.906	0.872	0.899	0.886	0.903	0.881	0.89
AH	0.717	0.808	0.715	0.809	0.716	0.808	0.715	0.808	0.717	0.807	0.711	0.805	0.714	0.751	0.716	0.735
PH	0.755	0.861	0.756	0.853	0.758	0.857	0.758	0.861	0.756	0.854	0.751	0.875	0.746	0.836	0.744	0.775
UA	0.953	0.961	0.949	0.954	0.951	0.952	0.923	0.936	0.949	0.952	0.872	0.872	0.972	0.977	0.974	0.984
OF	0.968	0.976	0.967	0.975	0.967	0.976	0.962	0.97	0.966	0.975	0.954	0.96	0.974	0.979	0.974	0.975
CE	0.785	0.856	0.785	0.821	0.784	0.844	0.789	0.84	0.781	0.816	0.777	0.787	0.802	0.871	0.801	0.811
SM	0.893	0.925	0.895	0.921	0.893	0.925	0.894	0.925	0.893	0.923	0.893	0.921	0.87	0.909	0.888	0.923
KH	0.923	0.94	0.924	0.932	0.923	0.939	0.922	0.945	0.924	0.932	0.925	0.931	0.894	0.917	0.923	0.93
WV	0.918	0.941	0.919	0.93	0.919	0.944	0.914	0.94	0.919	0.932	0.911	0.914	0.922	0.943	0.919	0.943
CH	0.789	0.903	0.79	0.904	0.788	0.903	0.79	0.902	0.789	0.904	0.789	0.9	0.787	0.904	0.789	0.909
AC	0.606	0.7	0.606	0.699	0.606	0.696	0.606	0.698	0.609	0.699	0.607	0.697	0.603	0.677	0.615	0.693

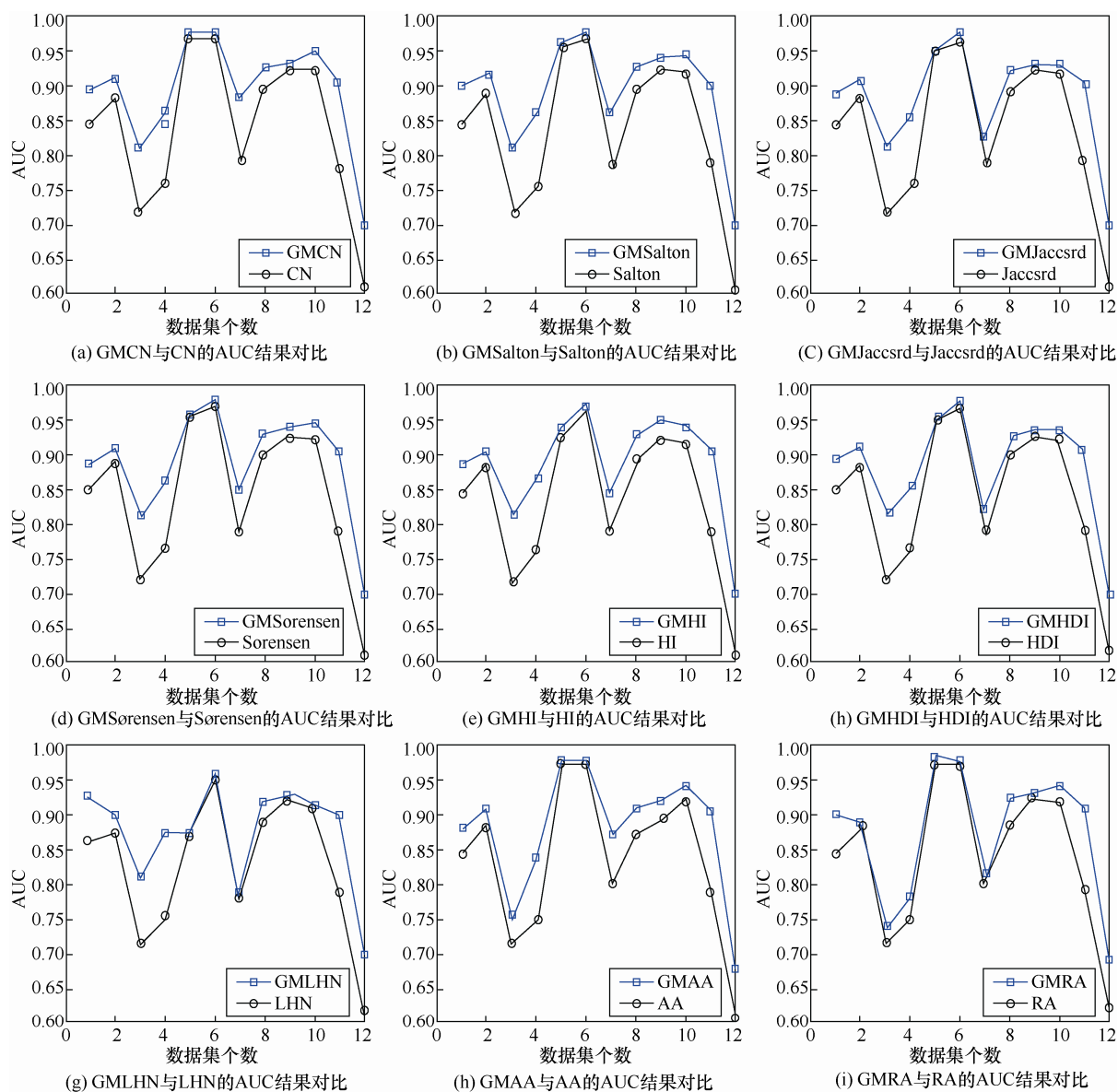


图2 不同数据集下GM类指标与原指标AUC结果
Figure 2 AUC of GM-indexs and original indexs under different datasets

5.2 RS 结果及分析

GMCN 与 CN 的 Ranking Score 对比结果在表 5 中给出。首先从整体来看, GMCN 指标的预测性能仍然全面优于 CN 指标, 并且在社交网络上表现突出, 相比 CN 指标的 RS 分值平均降幅为 30.9%, 低于 PA 及 LP 指标, 在 HS 及 RH 等部分数据集上低于 Katz 指标, 再次验证了基于 GMCN 的相似性指标在社交类网络的优良预测性能。另外, 在 UA、CE、WV、CH 等数据集上 GMCN 也可与 LP、PA、Katz 主流指标达到相当

的分值, 体现了该类指标整体预测排名准确性的优势。

同时将 GMSalton 等 8 种基于 GMCN 的指标在 12 个数据集下 RS 分值在表 6 中进行对比。该 8 种指标 RS 得分普遍远低于原指标, 均在各自网络数据集上表现出更好预测效果, 同时相比 AUC 衡量标准对比优势更为突出。如在 HS 数据集上原指标 RS 分值由 0.15 降至 0.09 左右, 优化指标的优势提升近 40%, 社交类网络整体降幅明显。其次, 该类指标在 WV 投票网络、CH 对战网络

表 5 Ranking Score 结果对比
Table 5 Comparison of Ranking Score results

数据集	CN	GMCN	PA	LP	Katz
HS	0.159	0.093	0.397	0.152	0.139
RH	0.12	0.094	0.358	0.111	0.112
AH	0.256	0.17	0.316	0.201	0.123
PH	0.188	0.102	0.297	0.11	0.06
UA	0.027	0.023	0.048	0.024	0.035
OF	0.025	0.019	0.068	0.015	0.018
CE	0.178	0.114	0.199	0.143	0.13
SM	0.131	0.098	0.042	0.045	0.003
KH	0.084	0.067	0.017	0.018	0.001
WV	0.077	0.052	0.02	0.02	0.015
CH	0.139	0.071	0.119	0.07	0.054
AC	0.232	0.192	0.145	0.197	0.089

表 6 基于 GMCN 的相似性指标 Ranking Score 结果
Table 6 Ranking Score results of similarity index based on GMCN

数据集	Salton	GMSalton	Jaccard	GMJaccard	Sørenson	GMSørenson	HPI	GMHPI	HDI	GMHDI	LHN	GMLHN	AA	GMAA	RA	GMRA
HS	0.15	0.085	0.151	0.087	0.151	0.097	0.154	0.091	0.153	0.091	0.137	0.118	0.135	0.115	0.136	0.093
RH	0.117	0.098	0.118	0.098	0.118	0.104	0.119	0.105	0.121	0.102	0.119	0.103	0.104	0.089	0.104	0.082
AH	0.256	0.17	0.256	0.17	0.256	0.17	0.256	0.17	0.256	0.17	0.255	0.175	0.257	0.181	0.255	0.178
PH	0.187	0.101	0.188	0.11	0.188	0.104	0.187	0.103	0.189	0.109	0.177	0.115	0.176	0.12	0.175	0.112
UA	0.046	0.039	0.048	0.047	0.048	0.047	0.074	0.064	0.05	0.049	0.126	0.124	0.027	0.022	0.025	0.02
OF	0.03	0.024	0.03	0.025	0.03	0.025	0.035	0.029	0.031	0.026	0.042	0.039	0.024	0.017	0.022	0.016
CE	0.187	0.156	0.19	0.18	0.19	0.153	0.183	0.164	0.192	0.182	0.241	0.188	0.166	0.128	0.165	0.113
SM	0.131	0.101	0.131	0.103	0.131	0.101	0.131	0.101	0.131	0.103	0.136	0.106	0.166	0.128	0.135	0.105
KH	0.084	0.068	0.084	0.078	0.084	0.072	0.084	0.067	0.084	0.079	0.079	0.069	0.119	0.092	0.078	0.066
WV	0.081	0.055	0.08	0.071	0.08	0.057	0.083	0.061	0.08	0.068	0.086	0.073	0.075	0.059	0.076	0.057
CH	0.14	0.072	0.139	0.071	0.139	0.072	0.14	0.074	0.139	0.072	0.14	0.076	0.139	0.072	0.138	0.071
AC	0.233	0.193	0.233	0.194	0.233	0.193	0.233	0.193	0.233	0.194	0.244	0.198	0.248	0.206	0.244	0.197

表现不俗，CH 网络 RS 得分由 0.14 降至 0.07。航空网络、生物网络上，该类指标在 RS 分值胜于原指标。GMRA 及 GMRA 相比其他新指标降幅稍低，但已优于 AA 及 RA 指标，预测排名整体有了极大提升。因此，在 RS 分值评价方式下，基于广义共同邻居的相似性指标对于各种类型的网络数据均具有更好的预测表现。

各新定义指标与原有指标在 12 个数据集上的 Ranking Score 效果对比如图 3 所示。对比图 2 和图 3 可以直观看到，RS 结果与 AUC 的提升效

果基本相对应，其 AUC 值提升明显的网络的 RS 分值也更低。9 个重定义的新指标在各数据集中的 RS 分值均在原指标曲线下方，说明数据集中的连边在整体排序中更为靠前。对于 4 种社交网络，GM 类指标的 RS 分值平均降低 30.9%，Physicians 网络上 RS 分值降低了 41%，效果最明显的国际象棋对战网络 RS 降低达 48.3%。另外，几类网络的降低普遍在 15%~20%，同样证实了基于广义共同邻居的相似性指标在 RS 评价指标下对不同网络的适用性。

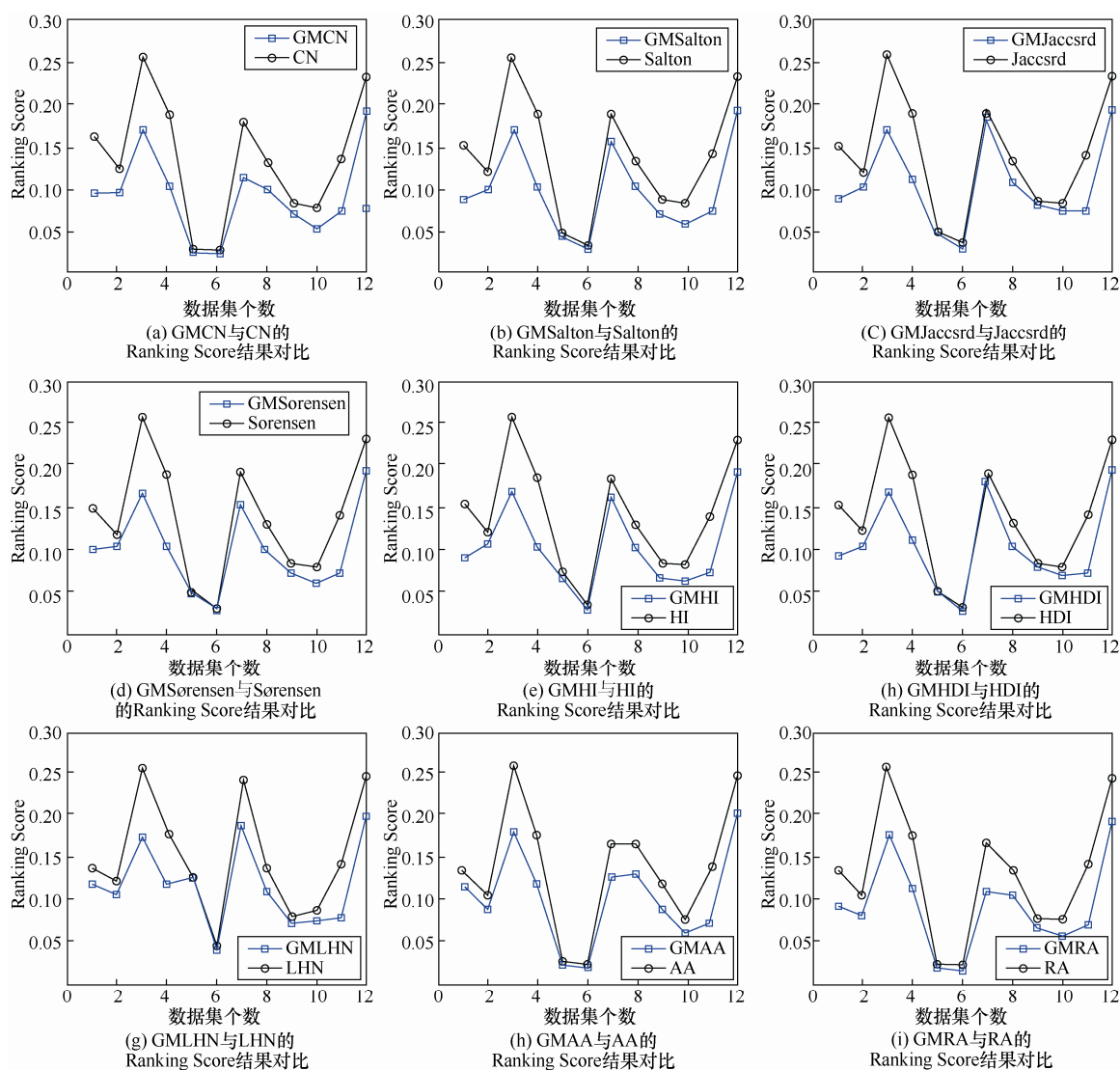


图 3 不同数据集下 GM 类指标与原指标 Ranking Score 结果
Figure 3 Ranking score of GM-indices and original indices under different datasets

6 结束语

本文针对有向网络中共同邻居存在多种有向异构形式，而现有指标缺少对多种异构形式下局部连边信息有效综合利用问题，提出了有向网络中广义共同邻居的定义，并基于广义共同邻居重定义了基于 CN 的 8 种相似性指标。以结构连边概率作为权重衡量促进连边的贡献程度使其针对多种类型网络具有普适性。十余个真实网络数据集仿真表明，广义共同邻居方式充分融合利用了共同邻居多种异构体间的局部结构信息，在 AUC、Ranking Score 衡量指标下，各新定义指标预测性能得到了

大幅提升。此外该类方法继承了 CN 算法低复杂度的优势，适用较大规模网络链路预测。

参考文献：

- [1] BENSON A, KLEINBERG J. Link prediction in networks with core-fringe data[C]//The World Wide Web Conference. 2019: 94-104.
- [2] PUJARI M, KANAWATI R. Link prediction in complex networks[M]//Advanced Methods for Complex Network Analysis. IGI Global, 2016: 58-97.
- [3] LYU L Y, ZHOU T. Link prediction in complex networks: a survey[J]. Physica A Statistical Mechanics & Its Applications, 2011, 390(6): 1150-1170.

- [4] 吕琳媛. 复杂网络链路预测[J]. 电子科技大学学报, 2010, 39(5): 651-661.
LYU L Y. Complex network link prediction[J]. Journal of Electronic Science and Technology, 2010, 39(5): 651-661.
- [5] MERING C V, JENSEN L J, SNEL B, et al. STRING: known and predicted protein-protein associations, integrated and transferred across organisms[J]. Nucleic Acids Research, 2005, 33: 433-7.
- [6] SCELLATO S, NOULAS A, MASCOLO C. Exploiting place features in link prediction on location-based social networks[C]//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2011: 1046-1054.
- [7] DORAN D. Stochastic blockmodeling[M]//Network Role Mining and Analysis. 2017: 63-76.
- [8] 王凯, 刘树新, 陈鸿昶, 等. 一种基于节点间资源承载度的链路预测方法[J]. 电子与信息学报, 2019, 41(05):214-223.
WANG K, LIU S X, CHEN H C, et al. A new link prediction method for complex networks based on resources carrying capacity between nodes[J]. Journal of Electronics and Information Technology, 2019, 41(5): 1225-1234.
- [9] 刘树新, 季新生, 刘彩霞, 等. 局部拓扑信息耦合促进网络演化[J]. 电子与信息学报, 2016, 38(9):2180-2187.
LIU S X, JI X A, LIU C X, et al. Information coupling of local topology promoting the network evolution[J]. Journal of Electronics and Information Technology, 2016, 38(9): 2180-2187.
- [10] LIU S, JI X, LIU C, et al. Extended resource allocation index for link prediction of complex network[J]. Physica A Statistical Mechanics & Its Applications, 2017, 479:174-183.
- [11] LIU S, JI X, LIU C, et al. Similarity indices based on link weight assignment for link prediction of unweighted complex networks[J]. International Journal of Modern Physics B, 2016, 31(2):1650254.
- [12] CHENG X Q, REN F X, SHEN H W, et al. Bridgeness: a local index on edge significance in maintaining global connectivity[J]. Journal of Statistical Mechanics: Theory and Experiment, 2010, 2010(10): P10011.
- [13] LORRAIN F, WHITE H C. Structural equivalence of individuals in social networks[J]. Social Networks, 1977, 1(1): 67-98.
- [14] SHANG K K, SMALL M, YAN W S. Link direction for link prediction[J]. Physica A: Statistical Mechanics and its Applications, 2016, 469: 767-776.
- [15] SCHALL, DANIEL. Link prediction in directed social networks[J]. Social Network Analysis and Mining, 2014, 4(1): 157-334.
- [16] VALVERDE-REBAZA J C, ROCHE M, PONCELET P, et al. The role of location and social strength for friendship prediction in location-based social networks[J]. Information Processing & Management, 2018, 54(4): 475-489.
- [17] 张扬夫. 有向与加权网络的链路预测[D]. 湘潭: 湘潭大学, 2011.
ZHANG Y F. Link prediction for directed and weighted networks[D]. Xiangtan: Xiangtan University, 2011.
- [18] SALTON G, MCGILL M J. Introduction to modern information retrieval[M]. 1983.
- [19] JACCARD P. Étude comparative de la distribution florale dans une portion des Alpes et des Jura[J]. Bull Soc Vaudoise Sci Nat, 1901, 37: 547-579.
- [20] SØRENSEN T A. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons[J]. Biol. Skar., 1948, 5: 1-34.
- [21] RAVASZ E, SOMERA A L, MONGRU D A, et al. Hierarchical organization of modularity in metabolic networks[J]. Science, 2002, 297(5586): 1551-1555.
- [22] ZHOU T, LYU L, ZHANG Y C. Predicting missing links via local information[J]. The European Physical Journal B, 2009, 71(4): 623-630.
- [23] LEICHT E A, HOLME P, NEWMAN M E J. Vertex similarity in networks[J]. Physical Review E, 2006, 73(2): 026120.
- [24] ADAMIC L A, ADAR E. Friends and neighbors on the Web[J]. Social Networks, 2003, 25(3):211-230.
- [25] OU Q, JIN Y D, ZHOU T, et al. Power-law strength-degree correlation from resource-allocation dynamics on weighted networks[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2007, 75(2 Pt 1): 021102.
- [26] 杨瑞琪, 张月霞. 一种时序有向社会网络中的链路预测算法[J]. 计算机工程, 2019, 45(3): 203-207.
YANG R Q, ZHANG Y X. A link prediction algorithm for temporal directed social network[J]. Computer Engineering, 2019, 45(3): 203-207.
- [27] BARABÁSI A L, ALBERT R. Emergence of scaling in random networks[J]. Science, 1999, 286(5439): 509-512.
- [28] LYU L, JIN C H, ZHOU T. Similarity index based on local paths for link prediction of complex networks[J]. Physical Review E, 2009, 80(4): 046122.
- [29] KATZ L. A new status index derived from sociometric analysis[J]. Psychometrika, 1953, 18(1): 39-43.
- [30] 潘永昊, 于洪涛, 吴翼腾. 基于复杂网络动力学模型的链路预测方法[J]. 网络与信息安全学报, 2019, 5(6): 67-74.
PAN Y H, YU H T, WU Y T. Link prediction method based on complex network dynamics model[J]. Chinese Journal of Network and Information Security, 2019, 5(6): 67-74.
- [31] 常圣, 马宏, 刘树新. 基于三元组结构的有向网链路预测方法[J]. 网络与信息安全学报, 2019, 5(5): 39-47.
CHANG S, MA H, LIU S X. New method for link prediction in directed networks based on triad patterns[J]. Chinese Journal of Network and Information Security, 2019, 5(5): 39-47.
- [32] BRZOZOWSKI M J, ROMERO D M. Who should I follow? re-

- commending people in directed social networks[C]//Fifth International AAAI Conference on Weblogs and Social Media. 2011.
- [33] MARTÍNEZ V, BERZAL F, CUBERO J C. A survey of link prediction in complex networks[J]. ACM Computing Surveys (CSUR), 2016, 49(4): 1-33.
- [34] ZENG G, ZENG E. On the three-way equivalence of AUC in credit scoring with tied scores[J]. Communications in Statistics-Theory and Methods, 2019, 48(7): 1635-1650.
- [35] ZHOU T, REN J, MEDO M, et al. Bipartite network projection and personal recommendation[J]. Physical Review E, 2007, 76(4): 046115.
- [36] KISH, L. Introduction to mathematical sociology[J]. Social Forces, 1965, 44(2): 252-253.
- [37] FREEMAN L C, WEBSTER C M, KIRKE D M. Exploring social structure using dynamic three-dimensional color images[J]. Social networks, 1998, 20(2): 109-118.
- [38] MOODY J. Peer influence groups: identifying dense clusters in large networks[J]. Social Networks, 2001, 23(4): 261-283.
- [39] COLEMAN J, MENZEL K H. The diffusion of an innovation among physicians[J]. Sociometry, 1957, 20(4): 253-270.
- [40] OPSAHL T. Why anchorage is not (that) important: binary ties and sample selection[EB].
- [41] OPSAHL T, AGNEESSENS F, SKVORETZ J. Node centrality in weighted networks: generalizing degree and shortest paths[J]. Social Networks, 2010, 32(3): 245-251.
- [42] WATTS D J, STROGATZ S H. Collective dynamics of 'small-world' networks[J]. Nature, 1998, 393(6684): 440.
- [43] BATAGELJ V, MRVAR A. Pajek—analysis and visualization of large networks[J]. Lecture Notes in Computer Science, 2002, 2265: 77-103.
- [44] LESKOVEC J, HUTTENLOCHER D, KLEINBERG J. Governance in social media: a case study of the Wikipedia promotion process[C]//Fourth International AAAI Conference on Weblogs and Social Media. 2010.
- [45] KUNEGIS J. Handbook of Network Analysis [KONECT -- the

Koblenz Network Collection][J]. Computer Science, 2014(2): 1343-1350.

[作者简介]



赵学磊（1996—），男，河北石家庄人，信息工程大学硕士生，主要研究方向为复杂网络链路预测。



季新生（1968—），男，河南驻马店人，信息工程大学教授、博士生导师，主要研究方向为网络空间安全、移动通信网络安全、拟态安全。



刘树新（1987—），男，山东潍坊人，信息工程大学助理研究员，主要研究方向链路预测、通信网络安全。



赵宇（1984—），男，吉林辽源人，信息工程大学讲师，主要研究方向为移动通信网络、新型网络体系结构。