一种基于线性规划的有向网络链路预测方法

李劲松[①] 彭建华[①] 刘树新*[①] 季新生[①②]

[①](中国人民解放军战略支援部队信息工程大学 郑州 450001)

[②](清华大学计算机科学与技术系 北京 100084)

摘 要：大多数有向网络链路预测方法在计算节点相似性时没有充分考虑有向网络的结构特点，未区分不同有向邻居对连边形成具有的贡献差异，导致预测性能受到局限。鉴于此，该文提出一种基于线性规划的有向网络链路预测方法。该方法对3种有向邻居的信息贡献进行量化分析，结合结构特点建立线性规划模型，进而通过求解贡献矩阵的最优解构建相似性指标。9个真实有向网络中的实验结果表明，所提方法相比于9种现有方法在两种衡量标准下表现出较高的预测性能与良好的鲁棒性。

关键词：有向网络；链路预测；节点相似性；线性规划

中图分类号：N92; TP393 **文献标识码**：A **文章编号**：1009-5896(2020)10-2394-09

DOI: 10.11999/JEIT190731

A Link Prediction Method in Directed Networks Via Linear Programming

LI Jinsong[①] PENG Jianhua[①] LIU Shuxin[①] JI Xinsheng[①②]

[①](*Information Engineering University, People's Liberation Army Strategic Support Force, Zhengzhou* 450001, *China*)

[②](*Department of Computer Science and Technology, Tsinghua University, Beijing* 100084, *China*)

Abstract: Most existing link prediction methods in directed networks fail to consider the structural properties of directed networks when calculating node similarity, nor do they differentiate the contributions of directed neighbors on link formation, resulting in the limitation on prediction performance. To solve these problems, a novel link prediction method in directed networks based on linear programming is proposed. The contributions of three types of directed neighbors are quantified, then the linear programming problem is established based on network topological property. The similarity index is deduced by solving the optimal solution of the linear programming problem. Experimental results on nine real-world directed networks show that the proposed method outperforms nine benchmarks on both accuracy and robustness under two evaluation metrics.

Key words: Directed network; Link prediction; Node similarity; Linear programming

1 引言

近年来，随着网络科学领域发展的日趋成熟，复杂网络已逐渐成为研究热点[1,2]。常见的复杂网络包括社交网络、信息网络、航空交通运输网、食物链网络等。链路预测作为研究复杂网络结构与动力学特性的重要工具，旨在利用已观测的部分网络数据推测未知连边存在的可能性[3]。链路预测已在诸多领域展现出巨大应用价值，如在社交网络中的链路预测是实现好友推荐系统的重要方法，在电信网络中链路预测可用于揭露可疑、隐蔽的通联关系，及时发现电信诈骗源头[2,3]。

复杂网络链路预测问题自提出以来已取得诸多研究成果。其中，基于拓扑相似性的链路预测方法实现简单、预测效果好，受到研究者青睐[4–7]。根据拓扑信息范围可将基于拓扑相似性的链路预测方法分为局部相似性指标、准局部相似性指标和全局相似性指标[2]3类。局部相似性指标以共同邻居结构为研究对象，认为两节点间若存在更多共同邻居则更可能形成连边。共同邻居(Common Neighbor, CN)指标、资源配置(Resource Allocation, RA)指标、Adamic-Adar(AA)指标等都基于这一假设[2]。Liu等人[6]和王凯等人[7]在RA的基础上先后提出扩

收稿日期：2019-09-20；改回日期：2020-05-25；网络出版：2020-06-01

*通信作者： 刘树新 liushuxin11@126.com

基金项目：国家自然科学基金(61803384)

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

展资源配置指标和节点间资源承载度指标。全局指标利用完整的网络结构信息实现链路预测，包括Katz[8]提出的Katz指标，Chebotarev等人[9]提出的MFI指标等。准局部指标考虑较小范围的拓扑信息，其中局部路径(Local Path, LP)指标[10]是经典的准局部指标。

现有链路预测研究多以无向网络为对象，然而现实世界中的网络多为有向的[11]。对连边方向的预测具有重要现实意义，但目前该方面研究尚不充分。Zhang等人[12]采用前馈环模体定义有向网络共同邻居，扩展了CN, AA和RA指标。Wang等人[13]在LP指标基础上提出扩展局部路径指标，通过增加全局节点解决稀疏网络信息损失问题。Li等人[11]借助零模型分析互惠边在网络中的作用，进而利用互惠边改进了4种有向网络链路预测指标。Bütün等人[14]提出一种基于监督学习的有向网络链路预测框架，将三元组闭合指数作为特征进行训练。Salha等人[15]结合有向网络结构特点与牛顿定律扩展了图自动编码器和变分自动编码器模型，提出一种基于图嵌入的有向网络链路预测方法。共同邻居结构是多数现有方法的研究对象，然而未构成共同邻居的网络节点对连边形成同样具有间接影响。Gundala等人[16]研究了社交网络中节点间的间接联系，提出两种基于监督学习的链路预测方法。Pech等人[17]提出线性优化(Linear Optimization, LO)指标，将邻居节点的贡献程度作为未知量进行优化求解。此类方法虽在无向网络中表现良好，但没有考虑有向网络的结构特点，未区分不同结构有向邻居对连边形成的贡献差异，在有向网络中的适用性受限。

基于上述分析，本文从有向网络的局部结构特征出发，通过对邻居节点贡献度的量化，构建线性规划模型，进而通过贡献度矩阵最优解推导出链路预测指标。该方法充分考虑有向网络所有可能的局部结构，不需先验信息且实现简单。通过多个真实有向网络中的实验验证了所提方法的准确性与鲁棒性。

2　相关工作介绍

2.1　问题描述

给定一个有向网络$D(\boldsymbol{V},\boldsymbol{E})$, $\boldsymbol{V}$和$\boldsymbol{E}$分别为节点集合和连边集合，$N=|\boldsymbol{V}|$和$M=|\boldsymbol{E}|$分别表示节点数量和连边数量。设网络邻接矩阵为$\boldsymbol{A}=[a_{xy}]_{N\times N}$。设网络中所有可能连边集合表示为$\boldsymbol{U}$，则不存在边集合为$\boldsymbol{U}$-$\boldsymbol{E}$。链路预测算法为每条不存在边$e(x,y)\in\boldsymbol{U}$-$\boldsymbol{E}$赋予一个相似性指标$s_{xy}$，以表征该连边存在的可能性。

2.2　相似性指标

节点相似性通常被看作连边形成的内在动力，以下介绍几种有向网络中常用的相似性指标。

(1) DCN(Directed Common Neighbor)[12]：利用节点x和y之间的共同邻居数量衡量相似度，定义为

$$s_{xy}^{\mathrm{DCN}}=|\Gamma_{\mathrm{out}}(x)\cap\Gamma_{\mathrm{in}}(y)| \tag{1}$$

(2) DAA(Directed Adamic-Adar)[12]：惩罚具有较大出度的共同邻居节点，定义为

$$s_{xy}^{\mathrm{DRA}}=\sum_{z\in\Gamma_{\mathrm{out}}(x)\cap\Gamma_{\mathrm{in}}(y)}\frac{1}{\lg(k_z^{\mathrm{out}})} \tag{2}$$

(3) DRA(Directed Resource Allocation)[12]：基于节点资源配置假设，认为共同邻居传递的资源量反比于其出度，用端节点获得的资源总量衡量节点相似度，定义为

$$s_{xy}^{\mathrm{DRA}}=\sum_{z\in\Gamma_{\mathrm{out}}(x)\cap\Gamma_{\mathrm{in}}(y)}\frac{1}{k_z^{\mathrm{out}}} \tag{3}$$

(4) DPA(Directed Preferential Attachment)[12]：基于偏好连接假设，认为连边概率正比于度，定义为

$$s_{xy}^{\mathrm{DPA}}=k_x^{\mathrm{out}}\times k_y^{\mathrm{in}} \tag{4}$$

(5) Bifan[18]：基于网络势能理论、聚类特性和同质性假设，筛选出最优网络模体Bifan，认为如果某连边的存在可导致网络中产生更多Bifan模体，则该连边存在的可能性越大。定义为

$$s_{xy}^{\mathrm{Bifan}}=|\Gamma_{\mathrm{in}}(\Gamma_{\mathrm{out}}(x))\cap\Gamma_{\mathrm{in}}(y)| \tag{5}$$

(6) LP(Local Path)[10]：在DCN指标的基础上考虑3阶有向路径上的邻居节点，定义为

$$\boldsymbol{S}=\boldsymbol{A}^2+\alpha\cdot\boldsymbol{A}^3 \tag{6}$$

(7) Katz[8]：考虑节点x和y之间的所有路径，将路径总数量化为节点相似度，定义为

$$\boldsymbol{S}^{\mathrm{Katz}}=(\boldsymbol{I}-\alpha\cdot\boldsymbol{A})^{-1}-\boldsymbol{I} \tag{7}$$

(8) MFI(Matrix Forest Index)[9]：基于矩阵森林理论的全局指标，设$\boldsymbol{L}$为网络拉普拉斯矩阵，定义为

$$\boldsymbol{S}^{\mathrm{MFI}}=(\boldsymbol{I}+\boldsymbol{L})^{-1} \tag{8}$$

(9) LO(Linear Optimization)[17]：将邻居贡献度作为未知量，构建并求解线性规划模型，定义为

$$\boldsymbol{S}^{\mathrm{LO}}=\alpha\boldsymbol{A}(\alpha\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}+\boldsymbol{I})^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} \tag{9}$$

3　基于线性规划的有向网络链路预测方法

3.1　有向网络局部结构

有向网络中从一个节点出发，可形成3种类型连边，即连入边、连出边和互惠边。考虑网络中某

节点$x \in \boldsymbol{V}$，设节点x的所有邻居集合为$\Gamma(x) = \Gamma_{\text{in}}(x) \cup \Gamma_{\text{out}}(x) \cup \Gamma_{\text{recip}}(x)$, $\Gamma_{\text{in}}(x)$表示连入邻居集合，$\Gamma_{\text{out}}(x)$表示连出邻居集合，$\Gamma_{\text{recip}}(x)$表示具有互惠连边的邻居集合。由于信息传导方向不同，3种邻居节点对连边形成具有不同程度的贡献。为区分其贡献大小，定义有向网络含权邻接矩阵$\boldsymbol{R}$。

定义1 对于一个有向网络$D(\boldsymbol{V}, \boldsymbol{E})$，定义含权邻接矩阵$\boldsymbol{R} = [r_{xy}]_{N \times N}$，令$\alpha \in \mathbb{R}$为可调参数，其元素满足

$$r_{xy} = \begin{cases} 1, & e(x,y) \in \boldsymbol{E},\ e(y,x) \notin \boldsymbol{E} \\ \alpha, & e(x,y) \notin \boldsymbol{E},\ e(y,x) \in \boldsymbol{E} \\ 1+\alpha, & e(x,y) \in \boldsymbol{E}, e(y,x) \in \boldsymbol{E} \end{cases} \tag{10}$$

与邻接矩阵$\boldsymbol{A}$不同，含权邻接矩阵$\boldsymbol{R}$通过参数α调节各类型连边的信息贡献。特别地，当$\alpha = 1$时，含权邻接矩阵$\boldsymbol{R}$中所有单向边的权值为1，互惠边权值为2，此时任意节点的连出和连入节点具有相同贡献。

3.2 有向网络线性规划指标

考虑两个节点$x, y \in \boldsymbol{V}$，设节点x的某一邻居节点$z \in \Gamma(x)$对连边$e(x,y)$的形成具有一定信息贡献，用c_{zy}表示。将连边$e(x,y)$的相似性指标定义为节点x的所有邻居节点对连边$e(x,y)$信息贡献的线性和，表示为

$$\begin{aligned} s_{xy} = &\sum_{z \in \Gamma(x)} r_{xz} \cdot c_{zy} = \sum_{z \in \Gamma_{\text{out}}(x)} c_{zy} \\ &+ \alpha \cdot \sum_{z \in \Gamma_{\text{in}}(x)} c_{zy} + (1+\alpha) \cdot \sum_{z \in \Gamma_{\text{recip}}(x)} c_{zy} \end{aligned} \tag{11}$$

其矩阵形式为

$$\boldsymbol{S} = \boldsymbol{AC} + \alpha \boldsymbol{A}^{\text{T}} \boldsymbol{C} = \boldsymbol{RC} \tag{12}$$

其中，$\boldsymbol{R} = \boldsymbol{A} + \alpha \boldsymbol{A}^{\text{T}}$, $\boldsymbol{C}$为贡献度矩阵。

图1所示为在一个简单网络中计算相似性矩阵$\boldsymbol{S}$示意图。其中，节点1, 2为待考察端节点，s_{12}为待求相似性指标。节点3, 4, 5分别为节点1的连出邻居、连入邻居、互惠邻居，其连边权重分别设为1, α, $1+\alpha$，得到网络含权邻接矩阵$\boldsymbol{R}$。矩阵$\boldsymbol{C}$的第2列量化不同邻居节点对连边$e(1,2)$的贡献程度，假设节点3, 4, 5的贡献程度分别为0.25, 0.36, 0.47，根据式(12)可得：$s_{12} = 1 \times 0.25 + \alpha \times 0.36 + (1+\alpha) \times 0.47$。在实际网络中，贡献度矩阵$\boldsymbol{C}$的取值由多个因素决定，下面利用线性规划方法求解其最优取值。

注意到相似性指标s_{xy}与邻接矩阵元素a_{xy}具有关联。以图2为例，当$a_{12} = 1$, $a_{31} = 1$, $a_{23} = 0$时，期望的相似性指标应满足：$s_{12} \to 1$, $s_{31} \to 1$, $s_{23} \to 0$。因此邻接矩阵$\boldsymbol{A}$与相似性矩阵$\boldsymbol{S}$对应元素之差应尽可能小。采用特定矩阵范数度量矩阵距离时，应满足$\|\boldsymbol{S} - \boldsymbol{A}\| \to 0$。结合上述分析可构建关于变量$\boldsymbol{C}$的线性规划模型，优化目标为最小化$\boldsymbol{S}$与$\boldsymbol{A}$的矩阵范数，表示为

$$\arg\min_{\boldsymbol{C}} \|\boldsymbol{S} - \boldsymbol{A}\| \tag{13}$$

为减少有效特征数量以避免参数过拟合，在目标函数增加参数范数惩罚项$\lambda \|\boldsymbol{C}\|$。此时优化目标变为

$$\arg\min_{\boldsymbol{C}} \|\boldsymbol{S} - \boldsymbol{A}\| + \lambda \|\boldsymbol{C}\| \tag{14}$$

采用F-范数(Frobenius norm)的平方表示矩阵范数，有$\|\boldsymbol{A}\| = (\|\boldsymbol{A}\|_{\text{F}})^2 = \left(\sqrt{\text{tr}(\boldsymbol{A}^{\text{T}}\boldsymbol{A})}\right)^2 = \text{tr}(\boldsymbol{A}^{\text{T}}\boldsymbol{A})$。令

$$\begin{aligned} E &= \text{tr}\left((\boldsymbol{S}-\boldsymbol{A})^{\text{T}}(\boldsymbol{S}-\boldsymbol{A})\right) + \lambda \cdot \text{tr}(\boldsymbol{C}^{\text{T}}\boldsymbol{C}) \\ &= \text{tr}(\boldsymbol{S}^{\text{T}}\boldsymbol{S} - \boldsymbol{A}^{\text{T}}\boldsymbol{S} - \boldsymbol{S}^{\text{T}}\boldsymbol{A} + \boldsymbol{A}^{\text{T}}\boldsymbol{A}) + \lambda \cdot \text{tr}(\boldsymbol{C}^{\text{T}}\boldsymbol{C}) \\ &= \text{tr}(\boldsymbol{C}^{\text{T}}\boldsymbol{R}^{\text{T}}\boldsymbol{RC} - \boldsymbol{A}^{\text{T}}\boldsymbol{RC} - \boldsymbol{C}^{\text{T}}\boldsymbol{R}^{\text{T}}\boldsymbol{A} + \boldsymbol{A}^{\text{T}}\boldsymbol{A}) \\ &\quad + \lambda \cdot \text{tr}(\boldsymbol{C}^{\text{T}}\boldsymbol{C}) \end{aligned} \tag{15}$$

对E关于矩阵$\boldsymbol{C}$求导，利用矩阵迹的求导公式$\text{dtr}(\boldsymbol{XA}) = \boldsymbol{A}^{\text{T}}\text{d}\boldsymbol{X}$，可得

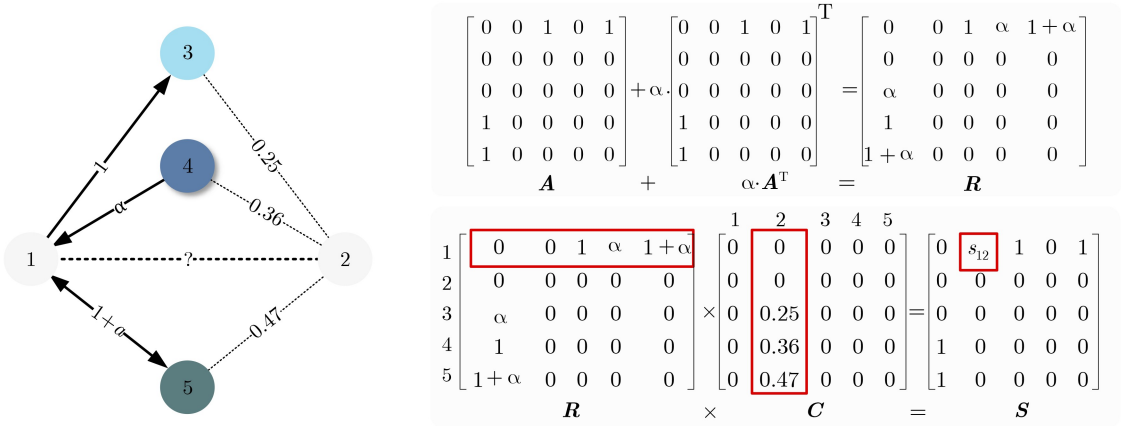


图 1 相似性矩阵的计算过程示意图

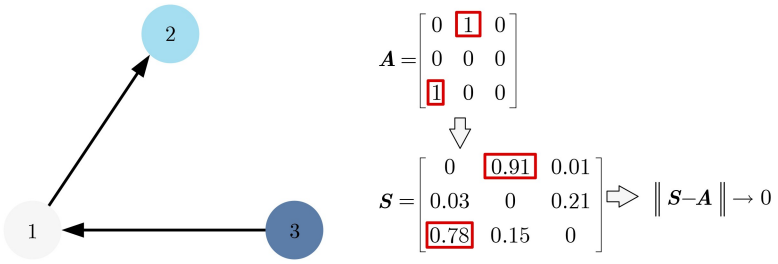


图 2 相似性指标与对应邻接矩阵元素之间的关系示意图

$$\begin{aligned}\frac{\mathrm{d}E}{\mathrm{d}\boldsymbol{C}}&=\frac{\mathrm{dtr}(\boldsymbol{C}^{\mathrm{T}}\boldsymbol{R}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{C})}{\mathrm{d}\boldsymbol{C}}-\frac{\mathrm{dtr}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{C})}{\mathrm{d}\boldsymbol{C}}-\frac{\mathrm{dtr}(\boldsymbol{C}^{\mathrm{T}}\boldsymbol{R}^{\mathrm{T}}\boldsymbol{A})}{\mathrm{d}\boldsymbol{C}}\\&\quad+\frac{\mathrm{dtr}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})}{\mathrm{d}\boldsymbol{C}}+\frac{\mathrm{dtr}(\lambda\cdot\boldsymbol{C}^{\mathrm{T}}\boldsymbol{C})}{\mathrm{d}\boldsymbol{C}}\\&=2\boldsymbol{R}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{C}-\boldsymbol{R}^{\mathrm{T}}\boldsymbol{A}-\boldsymbol{R}^{\mathrm{T}}\boldsymbol{A}+2\lambda\boldsymbol{C}\end{aligned}\tag{16}$$

根据有向网络结构易知，$\left|\boldsymbol{R}^{\mathrm{T}}\boldsymbol{R}+\lambda\cdot\boldsymbol{I}\right|\neq 0$。令上述导数$\mathrm{d}E/\mathrm{d}\boldsymbol{C}=0$，求得最优解$\boldsymbol{C}^*$为

$$\boldsymbol{C}^*=(\boldsymbol{R}^{\mathrm{T}}\boldsymbol{R}+\lambda\cdot\boldsymbol{I})^{-1}\boldsymbol{R}^{\mathrm{T}}\boldsymbol{A}\tag{17}$$

结合式(12)，利用最优解$\boldsymbol{C}^*$计算相似性矩阵$\boldsymbol{S}$，构建有向网络线性规划(Linear Programming index for Directed networks, LPD)指标，表示为

$$\boldsymbol{S}^{\mathrm{LPD}}=\boldsymbol{R}\boldsymbol{C}^*=\boldsymbol{R}(\boldsymbol{R}^{\mathrm{T}}\boldsymbol{R}+\lambda\cdot\boldsymbol{I})^{-1}\boldsymbol{R}^{\mathrm{T}}\boldsymbol{A}\tag{18}$$

其中，$\boldsymbol{R}=\boldsymbol{A}+\alpha\boldsymbol{A}^{\mathrm{T}}$。

4　衡量标准与数据集

链路预测在无权网络中可看作二分类问题。为评价链路预测方法性能，通常采用抽样方法将数据集划分为训练集和测试集。其中训练集用于计算链路预测指标，测试集用于计算该指标的性能衡量标准。

4.1　链路预测性能衡量标准

本文选用两种常用的衡量标准评价所提方法性能。受试者工作特性曲线(Receiver Operating characteristic Curve, ROC)可直观判断分类或预测效果[2]，它是一条以假正例率(False Positive Rate, FPR)作为横坐标，真正例率(True Positive Rate, TPR)作为纵坐标的曲线，满足

$$\mathrm{TPR}=\mathrm{TP}/N_+,\mathrm{FPR}=\mathrm{FP}/N_-\tag{19}$$

其中TP为真正例个数，FP为假正例个数，N_+为真正例和假正例之和，N_-为真正例和真负例之和。

AUC(Area Under receiver operating characteristic Curve)标准指ROC曲线与坐标轴围成的面积，可在整体角度衡量链路预测算法的准确性[2]。通过计算在测试集中随机选取的边的分数值比未连接边的分数值大的概率可以近似估计AUC。若测试集中边的分数值大于未连接边(假设有n'条)则加1分，若分数值相等(假设有n''条)则加0.5分，对应的AUC为

$$\mathrm{AUC}=\frac{n'+0.5n''}{n}\tag{20}$$

其中，n为比较次数。如果相似性指标为随机给出，AUC ≈ 0.5。AUC越接近1则意味着预测结果越准确。

4.2　网络数据

本文选取来自5个不同领域的9组有向网络数据进行实验。网络数据基本信息介绍如下：

(1) ADO (ADOlescent)[19]：一个根据1994-1995年某项调查构建的学生间好友关系网络。

(2) HIG (HIGh-school students)[19]：一个描述美国伊利诺伊州某高中男生间朋友关系的有向网络。

(3) RES (RESidence hall)[19]：一个包含217位住在澳大利亚国立大学宿舍区居民的好友关系网络。

(4) EMA (EMAil-enron-core)[19]：一个包含1999年至2003年间Enron公司员工之间的电子邮件网络。

(5) USA (US Airports)[19]：一个2010年美国各机场间的航空线路网络。

(6) OF (OpenFlight)[19]：一个包含世界各机场间航空线路的有向网络。

(7) CELE (C. ELEgans)[19]：一个秀丽隐杆线虫(Caenorhabditis elegans)的新陈代谢网络。

(8) FWFD (Food Web of FloriDa bay in fry season)[19]：一个南佛罗里达柏树湿地旱季食物链网络。

(9) LAK (Little rock lAKe)[19]：一个美国威斯康星州小石湖的食物链网络。

上述网络数据的基本统计参数如表1所示，其中包括网络类型，节点数$N=|\boldsymbol{V}|$，连边数$M=|\boldsymbol{E}|$，平均度$\langle k\rangle$，互惠系数ρ，平均集聚系数C, 90%有效直径$\langle d\rangle$，同配系数γ，幂率系数κ。

5　实验结果与分析

5.1　实验设置

实验选取9个主流有向网络链路预测指标作为参照，包括：4个局部指标(DCN, DAA, DRA, DPA)，2个准局部指标(LP, Bifan)和3个全局指标

(Katz, MFI, LO)。实验中，网络连边按比例f随机划分为训练集和测试集，二者边数之比为$f:(1-f)$。实验结果均取30次独立重复实验结果的平均值。

5.2 AUC结果与分析

表2示意了LPD指标与其他方法在9个真实网络中的AUC结果。其中LP-0.01表示LP指标中$\alpha=0.01$，Katz-0.01表示Katz指标中α=0.01, LO-0.01表示LO指标中α=0.01，LPD-0.01-1表示LPD指标中$\lambda^{-1}=0.01$, $\alpha=1$, LPD-max表示LPD指标的参数λ, α均取最优值。实验中，训练集划分比例$f=0.9$。

可以看到，4种局部指标的预测性能在不同网络中波动较大，2种准局部指标由于考虑了更多节点的有效信息，AUC得到一定提升。尤其是Bi-fan指标在CELE, FWFD, LAK 3个网络中性能甚至超过Katz和MFI。这说明在生物网络、食物链网络中个体之间的连接关系倾向形成Bifan模体。然而由于此方法基于对网络局部结构的既定假设，无法在各类网络中保持稳定性能。作为全局指标，Katz, MFI和LO表现出较高的预测性能，这得益于其充分考虑网络全局结构信息。同时注意到，MFI指标在RES, EMA中AUC较低，Katz指标在FWFD中相比其他指标AUC较低。这是由于Katz, MFI未考虑邻居节点对连边形成的间接影响，且未区分不同节点具有的信息贡献差异所致。LO指标的AUC相比Katz, MFI具有明显提升，然而在9组网络中均低于本文所提LPD指标。其原因在于LPD指标不仅考虑了不同节点的信息贡献差异，还对有向网络的不同邻居结构进行区分量化。综合看来，在除ADO以外的8组网络中，LPD指标的AUC均高于其他方法。仅在ADO网络中，LPD指标的AUC略低于Katz和MFI。LPD指标对不同类型网络结构特点的捕捉更加充分，因此预测性能更稳定。

5.3 ROC结果与分析

为深入对比本文所提LPD指标和其他方法的预

表 1 网络数据基本统计参数

序号	名称	类型	N	M	$\langle k\rangle$	$\rho(\%)$	C	$\langle d\rangle$	γ	κ
(1)	ADO	社交网络	2,539	12,969	10.22	38.8	14.2	5.30	0.251	8.25
(2)	HIG	社交网络	70	366	10.46	50.3	40.4	3.51	0.083	4.38
(3)	RES	社交网络	217	2,672	24.63	62.4	30.4	2.79	0.096	6.32
(4)	EMA	信息网络	1,005	25,571	50.89	71.1	25.7	3.01	–0.014	2.80
(5)	USA	交通网络	1,574	28,236	35.88	78.1	38.4	3.85	–0.113	1.85
(6)	OF	交通网络	2,939	30,501	20.76	97.2	25.5	5.19	0.051	1.74
(7)	CELE	生物网络	453	4,596	20.29	16.8	12.4	3.03	–0.226	2.62
(8)	FWFD	生态网络	128	2,137	33.39	2.9	31.4	1.88	–0.104	5.02
(9)	LAK	生态网络	183	2,494	27.26	4.09	33.2	2.65	–0.266	2.99

表 2 AUC结果对比

预测指标	ADO	HIG	RES	EMA	USA	OF	CELE	FWFD	LAK
DCN	0.716	0.860	0.889	0.949	0.969	0.968	0.804	0.745	0.942
DAA	0.714	0.862	0.895	0.953	0.972	0.969	0.807	0.749	0.940
DRA	0.716	0.861	0.895	0.956	0.973	0.971	0.811	0.752	0.939
DPA	0.680	0.621	0.650	0.887	0.953	0.924	0.810	0.854	0.946
Bifan	0.770	0.826	0.859	0.932	0.964	0.965	0.888	0.920	0.988
LP-0.001	0.781	0.882	0.899	0.954	0.976	0.984	0.860	0.769	0.964
Katz-0.001	**0.877**	0.883	0.898	0.954	0.976	0.984	0.874	0.778	0.970
MFI	0.872	0.845	0.792	0.772	0.908	0.975	0.836	0.734	0.947
LO-0.01	0.678	0.830	0.888	0.949	0.967	0.971	0.889	**0.972**	**0.995**
LO-0.001	0.679	0.830	0.868	0.956	0.977	0.974	0.892	0.934	0.990
LPD-0.01-1	0.803	**0.898**	**0.917**	0.956	**0.984**	**0.991**	**0.899**	0.956	0.994
LPD-0.001-1	0.812	0.883	0.902	**0.969**	0.968	0.990	0.886	0.907	0.988
LPD-max	0.814	**0.898**	**0.919**	**0.969**	**0.984**	**0.993**	**0.907**	**0.973**	**0.996**

测效果差异，对9组网络数据的ROC曲线进行分析。由于所提方法为全局指标，只选用4个全局指标(Katz, MFI, LO, LPD)进行对比。图3示意了9组网络数据中各指标的ROC曲线图。其中，LPD指标的参数为$\lambda^{-1}=0.01$, $\alpha=1$，Katz指标和LO指标中$\alpha=0.01$。可以看出，在9个真实网络数据集中，LPD指标的表现均明显优于其他3种全局指标。在RES, EMA, OF和FWFD网络中LPD指标的性能提升明显。在ADO网络中，LPD指标明显优于LO指标，相比Katz, MFI指标虽总体略差，但其在前30%的预测结果中准确度远高于Katz, MFI。从ROC曲线对比结果同样可以发现，MFI在不同网络中表现不一，在RES, EMA和FWFD网络中表现较差。总体而言，相比于其他3种全局指标，LPD指标在9组网络中表现较好，进一步验证了其有效性。

5.4 鲁棒性分析

下面分析链路预测算法在不同训练集划分比例下的鲁棒性。图4所示为训练集划分比例f从0.95减小至0.35时各方法的AUC变化曲线。从结果可知，当训练集比例减小时，随着网络已知信息的减少，大部分方法的AUC性能随之下降。总体而言全局指标的鲁棒性优于局部、准局部指标，在仅有35%可观测数据的情况下仍能保持较高的AUC。其中，Katz指标和MFI指标的波动范围虽然不大，但其AUC明显低于LPD指标。LO指标与LPD指标的变化趋势大致相同，但LO指标的AUC曲线始终低于LPD指标，在HIG, CELE, FWFD网络中差距尤为明显。

5.5 网络拓扑结构分析

下面对可调参数α的影响进行分析。图5所示为9组网络数据中LPD指标的AUC结果随参数α变化曲线图。参考Katz, LO指标的经验结果设置参数$\lambda^{-1}=0.01$，训练集划分比例$f=0.9$。注意到，当$\alpha=0$时式(9)和式(19)的表达是一致的，此时LPD指标等价于LO指标。由结果可知，在$\alpha=0$附近AUC变化明显，而在最优参数处LPD指标的预测

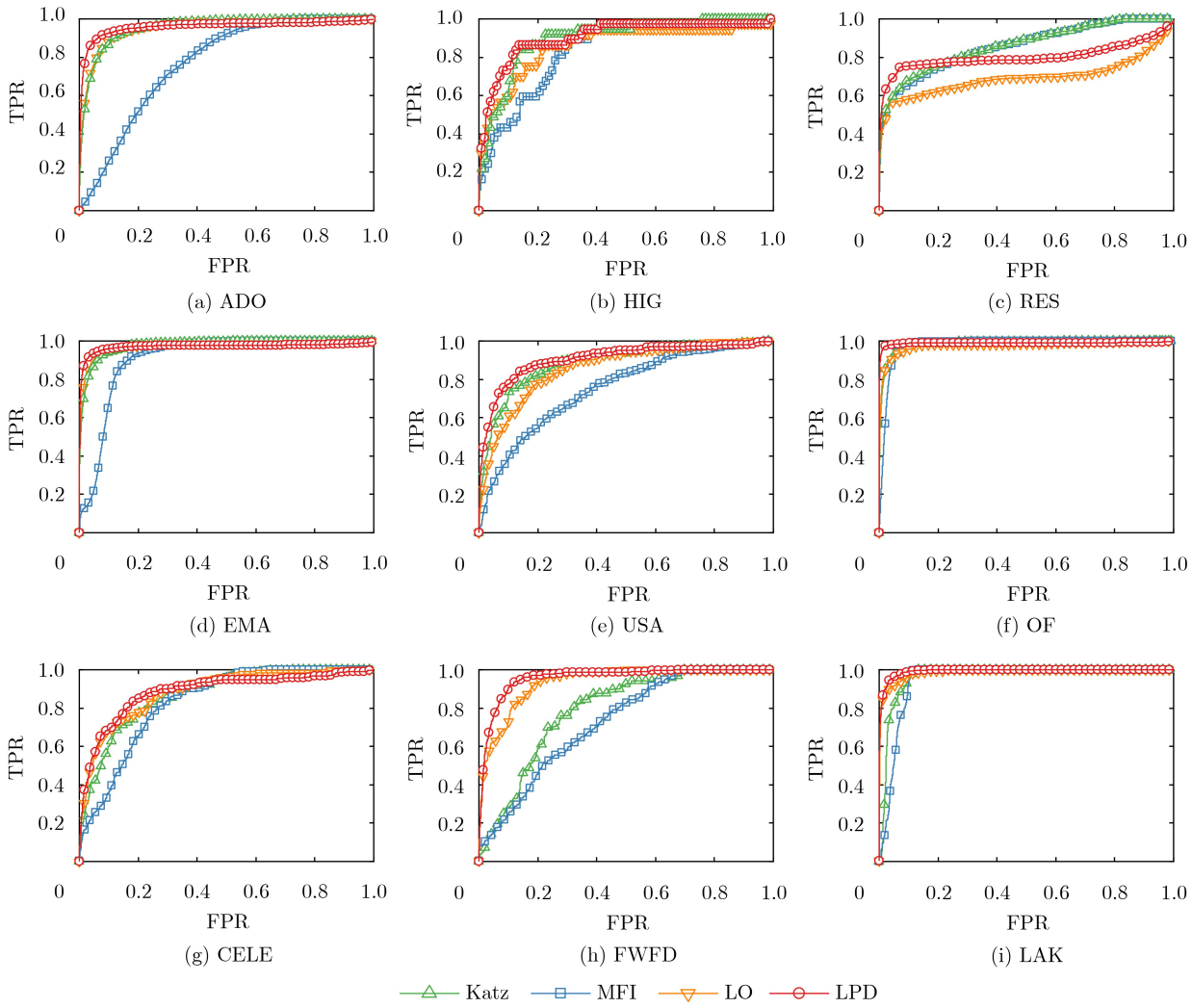


图 3 9组网络中ROC曲线对比图

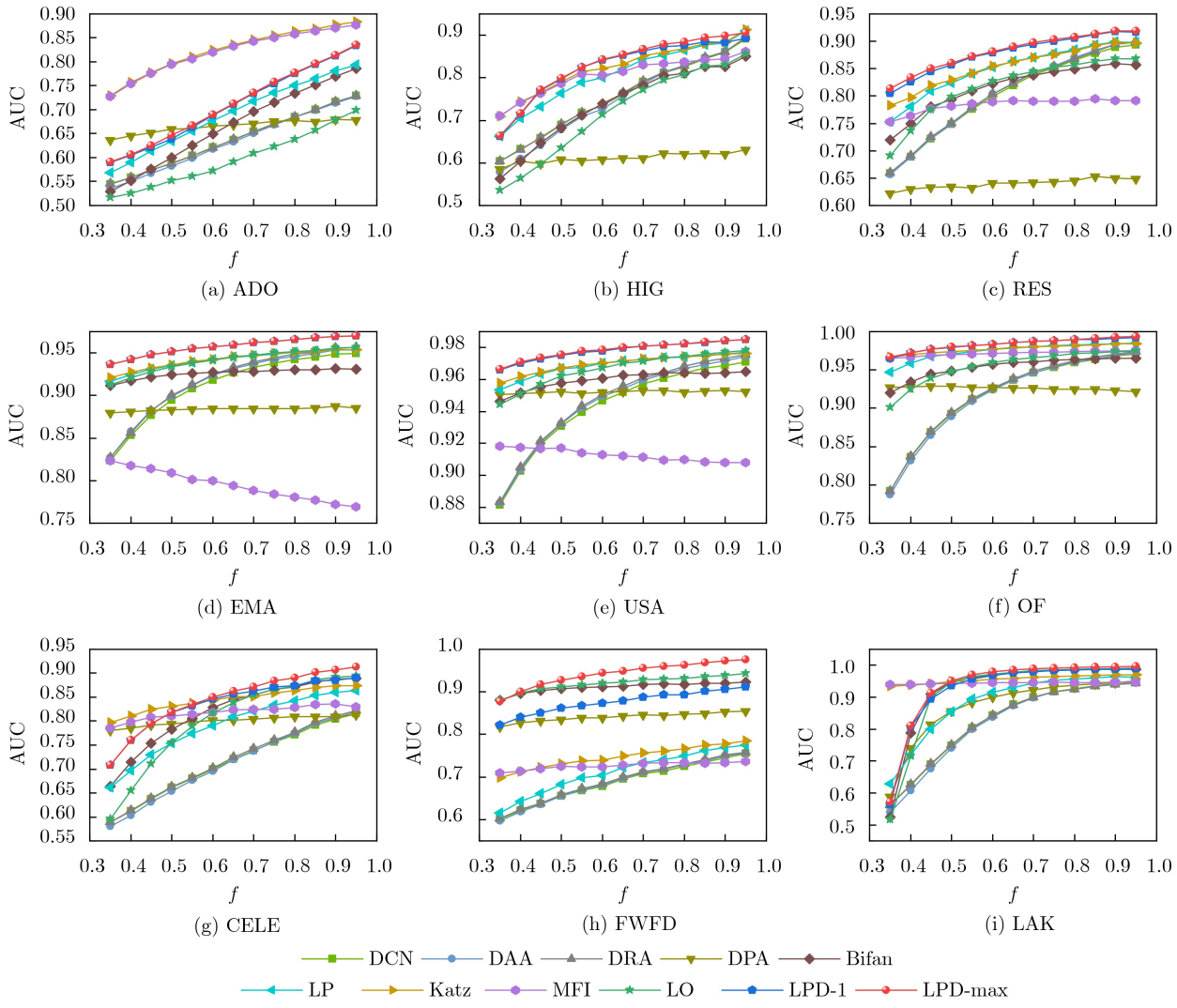
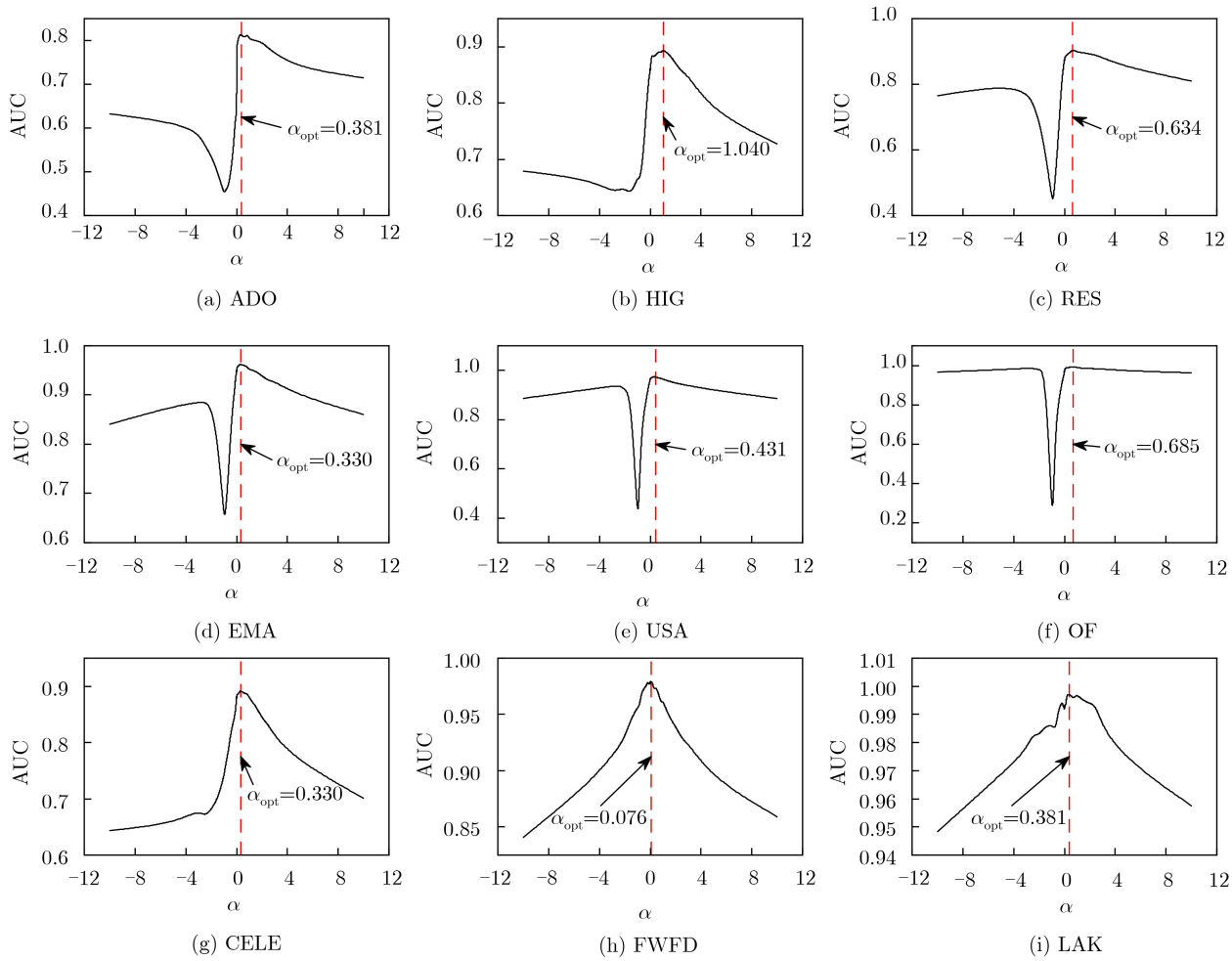


图 4 AUC随训练集划分比例变化曲线图

性能均优于LO指标。不同类型网络中最优参数α_{opt}的取值不同。在ADO, HIG, RES, EMA网络中，AUC在$\alpha=0$处呈上升趋势，在$(0,1]$区间取到最大值，而在$[-1,0)$区间性能急剧下降。这表明在此类网络中连出边、连入边和互惠边均具有近乎同等的地位，因此相比于只考虑连出边的LO指标，LPD指标可显著提升预测性能。这反映了互惠边是社交网络中的一种重要信息传递途径，对连边形成具有决定性作用。在CELE, FWFD和LAK网络中，AUC在$\alpha=0$左右达到最大值，之后随$|\alpha|$的增加性能逐渐降低。这表明在此类网络中连出边的影响比连入边更加显著，互惠边的影响几乎可以忽略不计。同时这也体现出蛋白质网络和食物链网络中存在较多前馈环模体，而互惠边对信息传递的贡献非常有限。在航空网络USA和OF中，最优参数α_{opt}不仅出现在$\alpha>0$区间，在$\alpha<-2$区间也可近似达到最佳性能。这表明在航空网络中连出边和互惠边的影响可能不尽相同，有些边的存在甚至具有负面影响。通常情况下，参数α的建议取值在0.5左右。

6 结束语

近年来，随着网络科学与链路预测的发展，有向网络中的链路预测问题逐渐成为研究热点。现有的有向网络链路预测方法大多未考虑不同类型邻居节点对连边形成的间接影响及其差异，导致有效信息缺失，预测性能受局限。针对上述问题，本文从有向网络特有的局部结构出发，通过量化分析不同有向邻居的贡献度建立线性规划模型，进而利用邻居节点贡献度矩阵最优解推导出LPD指标。多个真实网络中的实验结果表明，在两种衡量标准下，所提LPD指标不仅能提升预测精度，更表现出较好的鲁棒性与普适性。这为进一步揭示有向网络演化机制和内在连边机理提供了新思路。未来工作中我们将针对特定领域的网络进行深入分析，结合网络结构特点改进LPD指标，并探索其预测性能与网络复杂性的关系。



(a) ADO　(b) HIG　(c) RES

(d) EMA　(e) USA　(f) OF

(g) CELE　(h) FWFD　(i) LAK

图 5 AUC随参数α变化曲线图

参 考 文 献

[1] REN Zhuoming, ZENG An, and ZHANG Yicheng. Structure-oriented prediction in complex networks[J]. *Physics Reports*, 2018, 750: 1–51. doi: 10.1016/J.PHYSREP.2018.05.002.

[2] LÜ Linyuan and ZHOU Tao. Link prediction in complex networks: A survey[J]. *Physica A*: *Statistical Mechanics and its Applications*, 2011, 390(6): 1150–1170. doi: 10.1016/J.PHYSA.2010.11.027.

[3] 王凯, 李星, 兰巨龙, 等. 一种基于资源传输路径拓扑有效性的链路预测方法[J]. 电子与信息学报, 2020, 42(3): 653–660. doi: 10.11999/JEIT190333.

WANG Kai, LI Xing, LAN Julong, *et al*. A new link prediction method for complex networks based on topological effectiveness of resource transmission paths[J]. *Journal of Electronics & Information Technology*, 2020, 42(3): 653–660. doi: 10.11999/JEIT190333.

[4] LIU Shuxin, JI Xinsheng, LIU Caixia, *et al*. Similarity indices based on link weight assignment for link prediction of unweighted complex networks[J]. *International Journal of Modern Physics B*, 2017, 31(2): 1650254. doi: 10.1142/S0217979216502544.

[5] AGHABOZORGI F and KHAYYAMBASHI M R. A new similarity measure for link prediction based on local structures in social networks[J]. *Physica A*: *Statistical Mechanics and its Applications*, 2018, 501: 12–23. doi: 10.1016/J.PHYSA.2018.02.010.

[6] LIU Shuxin, JI Xinsheng, LIU Caixia, *et al*. Extended resource allocation index for link prediction of complex network[J]. *Physica A*: *Statistical Mechanics and its Applications*, 2017, 479: 174–183. doi: 10.1016/J.PHYSA.2017.02.078.

[7] 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(5): 1225–1234. doi: 10.11999/JEIT180553.

WANG Kai, LIU Shuxin, CHEN Hongchang, *et al*. A new link prediction method for complex networks based on resources carrying capacity between nodes[J]. *Journal of Electronics & Information Technology*, 2019, 41(5): 1225–1234. doi: 10.11999/JEIT180553.

[8] KATZ L. A new status index derived from sociometric analysis[J]. *Psychometrika*, 1953, 18(1): 39–43. doi: 10.1007/BF02289026.

[9] CHEBOTAREV P and SHAMIS E. The matrix-forest theorem and measuring relations in small social groups[J]. *Automation and Remote Control*, 1997, 58(9): 1505–1514.

[10] ZHOU Tao, LÜ Linyuan, and ZHANG Yicheng. Predicting missing links via local information[J]. *The European Physical Journal B*, 2009, 71(4): 623–630. doi: 10.1140/EPJB/E2009-00335-8.

[11] LI Jinsong, PENG Jianhua, LIU Shuxin, *et al*. Link prediction in directed networks utilizing the role of reciprocal links[J]. *IEEE Access*, 2020, 8: 28668–28680. doi: 10.1109/ACCESS.2020.2972072.

[12] ZHANG Xue, ZHAO Chengli, WANG Xiaojie, *et al*. Identifying missing and spurious interactions in directed networks[J]. *International Journal of Distributed Sensor Networks*, 2015, 11(9): 507386. doi: 10.1155/2015/507386.

[13] WANG Xiaojie, ZHANG Xue, ZHAO Chengli, *et al*. Predicting link directions using local directed path[J]. *Physica A*: *Statistical Mechanics and its Applications*, 2015, 419: 260–267. doi: 10.1016/J.PHYSA.2014.10.007.

[14] BÜTÜN E and KAYA M. A pattern based supervised link prediction in directed complex networks[J]. *Physica A*: *Statistical Mechanics and its Applications*, 2019, 525: 1136–1145. doi: 10.1016/J.PHYSA.2019.04.015.

[15] SALHA G, LIMNIOS S, HENNEQUIN R, *et al*. Gravity-inspired graph autoencoders for directed link prediction[C]. The 28th ACM International Conference on Information and Knowledge Management, Beijing, China, 2019: 589–598. doi: 10.1145/3357384.3358023.

[16] GUNDALA L A and SPEZZANO F. Estimating node indirect interaction duration to enhance link prediction[J]. *Social Network Analysis and Mining*, 2019, 9(1): 17. doi: 10.1007/s13278-019-0561-2.

[17] PECH R, HAO D, LEE Y L, *et al*. Link prediction via linear optimization[J]. *Physica A*: *Statistical Mechanics and Its Applications*, 2019, 528: e121319. doi: 10.1016/j.physa.2019.121319.

[18] ZHANG Qianming, LÜ Linyuan, WANG Wenqiang, *et al*. Potential theory for directed networks[J]. *PLoS One*, 2013, 8(2): e55437. doi: 10.1371/JOURNAL.PONE.0055437.

[19] KUNEGIS J. KONECT network dataset[EB/OL]. http://konect.uni-koblenz.de/networks/, 2017.

李劲松：男，1992年生，博士生，研究方向为复杂网络，链路预测，网络安全.

彭建华：男，1966年生，研究员，研究方向为网络安全，云安全，复杂网络.

刘树新：男，1987年生，助理研究员，研究方向为复杂网络，链路预测，网络演化.

季新生：男，1969年生，教授，研究方向为网络安全，云安全，复杂网络.

责任编辑：余 蓉