计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210100107

基于路径连接强度的有向网络链路预测方法

赵学磊　季新生　刘树新　李英乐　李海涛
中国人民解放军战略支援部队信息工程大学　郑州 450002
(zxl96@alu.uestc.edu.cn)

摘　要　链路预测旨在利用可获得的网络拓扑信息预测未知的连接关系。基于路径联系的预测方法在无向网络中取得了较好的效果。然而,在有向网络下,相同长度的路径因路径中连边方向不同会造成节点连接强度不同,传统预测方法难以区分路径异构造成的差异。鉴于此,首先以边权矩阵量化各类有向边连接强度的差异,进而为节点间不同异构的多类路径计算其连接强度,然后区分同一长度路径下各类路径的作用大小,最后综合多阶不同长度路径贡献,提出了一种基于路径连接强度的有向网络链路预测方法。在9个真实网络数据集上进行了实验,结果表明,考虑路径连接强度差异有效提高了在AUC及Precision衡量标准下的预测性能。

关键词:复杂网络;链路预测;有向路径;连接强度
中图法分类号　TP301.6

Link Prediction Method for Directed Networks Based on Path Connection Strength

ZHAO Xue-lei,JI Xin-sheng,LIU Shu-xin,LI Ying-le and LI Hai-tao
PLA Strategic Support Force Information Engineering University,Zhengzhou 450002,China

Abstract　Link prediction aims to predict unknown links using available network topology information. Prediction methods based on paths perform well in undirected networks. However,paths of the same length have different node connection strength due to different type of links through the path in directed network. Traditional methods is difficult to distinguish the path heterogeneity. Given this,the difference in the strength of three types of directed links is first quantified in terms of the link weight matrix,then the connection strength of different heterogeneous classpaths between nodes is calculated and the effect of different paths under the same length path is distinguished. Finally,a directed network link prediction method based on the path connection strength is proposed by integrating the contribution of multi-order paths of different lengths. Validation of 9 real networks shows that accounting for differences in path connection strength effectively improves prediction performance under the AUC and Precision metrics.

Keywords　Complex network,Link prediction,Directed paths,Connection strength

1　引言

近年来,复杂网络[1]越来越多地被用于分析现实网络的抽象理论而成为网络科学研究的热点。各类社会网络[2]、通信网络[3]、生物网络[4]等均可抽象为复杂网络。链路预测[2,5-7]是在上述网络下利用已知数据来预测未知连接关系的研究方法,对理解网络演化及网络动力学具有较大研究意义,在推荐系统[8]等方面已有相关应用。

相似性是链路预测的重要假设,即认为具有相似地位或拓扑结构的节点间存在连边的可能性较高[9]。目前无向网络链路预测以所需信息可分类为局部、准局部及全局相似性指标[7]。局部相似性基于共同邻居越多,则相似性就越大的假设,包括CN(Common Neighbors)指标、AA(Adamic-Adar)指标和RA(Resource Allocation)指标等,计算简便,适用于多类大规模网络;全局相似性指标包括Katz指标、MFI(Matrix-Forest Index)指标、Cosplus指标等,需获取全局网络拓扑,一般精度最好,但计算代价极大;准局部指标的精度及复杂度介于局部及全局相似性指标之间,如LP及ERA指标在三跳路径的距离上计算相似度,相比局部指标,其预测精度更好。

然而,现实世界中的网络多为有向网络,需对预测方法作必要的改进。Zhang等[10]通过区分连入连出邻居$\Gamma_{in}(i)$,$\Gamma_{out}(i)$,定义有向共同邻居为$z\in\Gamma_{out}(i)\bigcap\Gamma_{in}(j)$,提出了扩展的DCN(Directed Common Neighbors),DAA(Directed Adamic-Adar)和DRA(Directed Resource Allocation)等指标。Shang等[11]研究了有向网络中的互惠边作用,发现通过互惠边连接到同一邻居的节点联系更强。Zhang等[12]提出有向

到稿日期:2021-01-14　返修日期:2021-04-07
基金项目:国家自然科学基金(61803384)
This work was supported by the National Natural Science Foundation of China(61803384).
通信作者:刘树新(liushuxin11@126.com)

网络势能理论，认为势能沿有向边依次降低，并使用包含一条反向边的 Bifan 模体设计了高效精准的 Bifan 预测器。Li 等[13]通过零模型分析了有向网络中互惠边构成模体的显著性。Bütün 等[14]研究了网络中三元组的闭合指数，提出了一种基于监督学习的有向网络预测框架。Pech 等[15]则将邻居节点的间接连边贡献置为未知量，以线性规划方式求解该邻居节点的最优贡献。Li 等[16]参考网络层级信息并通过信息优化方式来设计预测方法。

上述方法的共同之处在于通过有向边量化邻居节点的不同信息贡献，而网络中的路径同样是促进连边生成的重要动力，节点间存在的连通路径越多，将来它们之间建立直接联系的可能性就越大[17]。Lü 等[17]基于该假设考虑节点间三阶，以内路径数量计算相似度。Wang 等[18-19]深入研究了无向路径上的拓扑有效性及资源承载度等。Liu 等[20-21]分析了路径资源传输能力及匹配度等影响，在无向网络中进一步提高了路径相似度的预测能力，而该类方法在有向网络中无法对不同路径连接强度进行区分，以指标定义出发，LP 指标计算中的$\{A^2\}_{ij}$在有向网络中，则特指正向路径 $i\to k\to j$。同理，$\{A^n\}_{ij}$特指 n 阶正向路径，未考虑包含反向边、互惠边的混合路径与正向路径连接强度的区别。

鉴于上述分析，本文设计边权矩阵，使用自适应参数来区分有向边的类型，量化节点间多阶有向路径连接强度的差异，提出适用于有向网络的路径相似度指标 DMP（Directed Multi-Path）。该指标的复杂度介于准局部与全局相似性指标之间，通过边权矩阵计算实现的过程简便，在数个真实公开网络上均有较高的准确性及鲁棒性。

2 问题描述及相关工作

2.1 链路预测问题描述

假设一个给出的网络 $G(\boldsymbol{V},\boldsymbol{E})$，$\boldsymbol{V}=\{v_1,v_2,\cdots,v_N\}$，$\boldsymbol{E}=\{e_1,e_2,\cdots,e_M\}$分别表示节点和连边集合，节点数为 N，连边数为 M。以 $N\times N$ 的矩阵 $\boldsymbol{A}$ 表示节点的邻接关系，$e_{ij}\in\boldsymbol{E}$ 时，$a_{ij}=1$。显然，在无向网络中 $\boldsymbol{A}$ 为一个对称矩阵，反之在有向网络中 $\boldsymbol{A}$ 为非对称矩阵。预测算法通过已知的节点及拓扑信息为当前未连接的节点计算一个连边相似分值，相似分值越大，该连边存在的可能性就越大。

在预测目标中，通常依据不同评价需求设定相似分阈值，相似分高于阈值的预测边将作为推荐结果；或依据相似分值为节点对排序，按需取前列 m 条预测边作为预测结果。预测连边进一步可应用于好友推荐系统或在工程实验中作为指导依据等。

2.2 相似性指标

链路预测在无向网络中得到了广泛且深入的研究，部分预测方法已扩展到有向网络。以 $\Gamma_{\text{out}}(i)$，$\Gamma_{\text{out}}(j)$分别表示节点 i 的出度邻居、节点 j 的入度邻居。下文将介绍当前有向网络中的通用预测指标。

（1）DCN。以节点间的共同邻居数作为相似分值，定义为：

$$s_{ij}^{DCN}=|\Gamma_{\text{out}}(i)\bigcap\Gamma_{\text{in}}(j)|=\sum_{z\in\Gamma_{\text{out}}(i)\cap\Gamma_{\text{in}}(j)}1 \tag{1}$$

（2）DAA。以共同邻居节点度对数的倒数作为该邻居权重，定义为：

$$s_{ij}^{DAA}=\sum_{z\in\Gamma_{\text{out}}(i)\cap\Gamma_{\text{in}}(j)}\frac{1}{\log(k_{\text{out}}(z))} \tag{2}$$

（3）DRA。类似 DAA 指标，以节点度的倒数作为该邻居权重，定义为：

$$s_{ij}^{DRA}=\sum_{z\in\Gamma_{\text{out}}(i)\cap\Gamma_{\text{in}}(j)}\frac{1}{k_{\text{out}}(z)} \tag{3}$$

（4）DPA。启发于节点的偏好连接现象，认为大度节点间更易产生连接，定义为：

$$s_{ij}^{DPA}=k_{\text{out}}(z)\cdot k_{\text{in}}(z) \tag{4}$$

（5）LP。在 CN 指标基础上引入三阶路径的影响，以参数 α 调节三阶路径的权重，定义为：

$$\boldsymbol{S}^{LP}=(\boldsymbol{A}^2)_{ij}+\alpha\cdot(\boldsymbol{A}^3)_{ij} \tag{5}$$

（6）Bifan。基于势能理论，认为某条连边建立若能生成更多“Bifan”模体，则该连边存在的可能性越大，“Bifan”模体预测器的定义为：

$$\boldsymbol{S}_{ij}^{Bifan}=(\boldsymbol{A}\times\boldsymbol{A}^{\mathrm{T}}\times\boldsymbol{A})_{ij} \tag{6}$$

（7）矩阵森林指数（MFI）。将节点 i 和节点 j 的相似性理解为节点 i 和节点 j 同属于节点 i 为根节点的森林的比例，以 $\boldsymbol{L}$ 表示网络拉普拉斯矩阵，MFI 定义为：

$$\boldsymbol{S}^{MFI}=(\boldsymbol{I}+\boldsymbol{L})^{-1} \tag{7}$$

（8）Katz。综合考虑两节点间所有路径的共同影响，表示为：

$$\begin{aligned}s_{ij}^{Katz}&=\sum_{l=1}^{\infty}\alpha^l\cdot|paths_{i,j}^{\langle l\rangle}|\\&=\alpha\boldsymbol{A}+a^2(\boldsymbol{A}^2)_{ij}+\alpha^3\cdot(\boldsymbol{A}^3)_{ij}+\cdots\end{aligned} \tag{8}$$

若令 $\alpha<\lambda_{\max}^{-1}$使得该级数收敛，$\lambda_{\max}$为 $\boldsymbol{A}$ 的最大特征值，可简化为：

$$\boldsymbol{S}=(\boldsymbol{I}-\alpha\cdot\boldsymbol{A})^{-1}-\boldsymbol{I} \tag{9}$$

3 基于路径连接强度的链路预测方法

3.1 有向边强度分析及量化

有向网络与无向网络的本质区别在于节点间包括正向边、反向边及互惠边这 3 种不同类型。如图 1 所示，正向边由始节点指向终节点，即 $e_{ij}\in\boldsymbol{E}$ 且 $e_{ji}\notin\boldsymbol{E}$，反向边由终节点指向始节点，即 $e_{ik}\notin\boldsymbol{E}$ 且 $e_{ki}\in\boldsymbol{E}$，互惠边则表示二者同时指向对方，即 $e_{il}\in\boldsymbol{E}$ 且 $e_{li}\in\boldsymbol{E}$。注意到，有向边类型是相对始节点描述的，如图 1（a）中 e_{ij} 表示节点 i 的正向边，而 e_{ji} 则表示节点 j 的反向边。

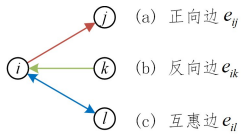


图 1 3 种有向连边示意图

Fig. 1 Three kinds of directed connected diagrams

多数有向网络中，正向边是最常见的连接关系，通常代表节点的主动行为，图 1(a)可表示微博用户 i 关注用户 j，文献 i 引用文献 j，物种 i 捕食物种 j 等，对预测节点 i 的连边有积极作用；而反向边通常代表节点被动行为，对预测节点 i 连边连接强度较小甚至抑制新生连边；互惠边则是人类社会中互惠性的体现，在社交网络中的占比较大，表示两节点间具有更紧密的联系，因此对预测连边的连接强度最大[11]。不同连边对节点的连接强度存在区别，结合上述分析，分别为 3 类连边

定义不同权重，以量化其对节点相似性的影响。

定义 1 给定一有向网络 $D(\boldsymbol{V},\boldsymbol{E})$，以邻接矩阵 $\boldsymbol{A}$ 表示，其邻接边权矩阵 $\boldsymbol{W}$ 表示为：

$$\boldsymbol{W}_{ij}=\begin{cases}1, & a_{ij}=1 \text{ 且 } a_{ji}\neq 1\\ \lambda, & a_{ij}\neq 1 \text{ 且 } a_{ji}=1\\ 1+\lambda, & a_{ij}\neq 1 \text{ 且 } a_{ji}=1\end{cases} \tag{10}$$

以图 1 为例，对应的邻接矩阵及边权矩阵的对比如图 2 所示。邻接矩阵 $\boldsymbol{A}$ 仅表示出了节点之间的正向连接关系，元素值 1 对应节点间的正向边，反向边无法在邻接矩阵中表示，且所有连边在拓扑中视为相同地位，导致无法区别以不同有向边连接的节点间的强度。而边权矩阵 $\boldsymbol{W}$ 将邻接矩阵中的元素值 1 进一步细化为不同的边权值，可同时表示 3 类有向边，并区分有向边方向对始点和终点的不同强度。值得注意的是，λ 为待定参数，表示反向边相对于正向边的连接强度，在不同类型网络中取值通常不同。进一步，在矩阵运算中，邻接矩阵累乘表示节点间的路径数目，而边权矩阵累乘则表示节点间路径的连接强度。

$$\begin{array}{c|cccc} & i & j & k & l\\ i & 0 & 1 & 0 & 1\\ j & 0 & 0 & 0 & 0\\ k & 1 & 0 & 0 & 0\\ l & 1 & 0 & 0 & 0\end{array} \qquad \begin{array}{c|cccc} & i & j & k & l\\ i & 0 & 1 & \lambda & 1+\lambda\\ j & \lambda & 0 & 0 & 0\\ k & 1 & 0 & 0 & 0\\ l & 1+\lambda & 0 & 0 & 0\end{array}$$

(a)邻接矩阵 $\boldsymbol{A}$ (b)边权矩阵 $\boldsymbol{W}$

图 2 邻接矩阵与边权矩阵的对应关系

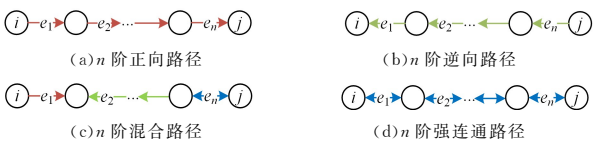
Fig. 2 Adjacency matrix and edge weight matrix

3.2 有向路径连接强度分析及量化

有向路径是端节点通过中间节点及有向边连接形成的链式结构，其定义如下。

定义 2 有向网络拓扑中，始节点与终节点间通过 $n(n\geqslant 2)$ 条有向边及 $n-1$ 个中间节点交替连接，且不包含自环及重复边形成的链式结构，称为有向路径，记为 L_{ij}^{n}。

如图 3 所示，有向路径以始节点参照可分为 4 类：正向路径、逆向路径、混合路径及强连通路径。



(a)n 阶正向路径 (b)n 阶逆向路径

(c)n 阶混合路径 (d)n 阶强连通路径

图 3 4 种有向路径示意图

Fig. 3 Four directed path schematics

正向路径表示由始点 i 依次仅由正向边连接到下一节点的路径。逆向路径表示由始点 i 依次仅由反向边连接到下一节点的路径。强连通路径表示各节点均由互惠边相连接的路径。混合路径表示节点间同时存在正向边、反向边及互惠边的路径。

不同的有向路径在网络中的作用不同，结合 3.1 节有向边连接强度的定义，以路径中各跳连边强度之积表示路径总强度，则给定一个 n 阶路径的连接强度可表示为：

$$w(Path_{ij}^{n})=\prod_{e_i\in Path_{ij}^{n}} w_{e_i}=1^{n-l-m}\cdot\lambda^{l}\cdot(1+\lambda)^{m} \tag{11}$$

其中，$|\vec{e}|=n-l-m$，$|\overleftarrow{e}|=l$，$|\overleftrightarrow{e}|=m$ 分别表示路径中正向边、反向边及互惠边的数量。正向路径、逆向路径、强连通路径可视为混合路径的 3 种特例，其连接强度可特殊表示为：

$$w(Path_{ij}^{n})=\begin{cases}1, & |\vec{e}|=n\\ \lambda^{n}, & |\overleftarrow{e}|=n\\ (1+\lambda)^{n}, & |\overleftrightarrow{e}|=n\end{cases} \tag{12}$$

以图 4 为例，节点 1 和节点 2 之间存在的有向路径及其连接强度如表 1 所列。得益于自适应参数 λ 的调节作用，考虑了反向边对路径强度的影响，使相同长度的路径得到了不同的路径权重值。相比无向网络路径仅数目统计的方式，该方法能够进一步对路径的强度进行差异量化，更适合用于有向网络。

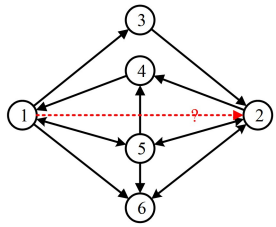


图 4 节点间不同连接强度的有向路径示意

Fig. 4 Directed paths of different connection strengths

表 1 有向网络多阶路径连接强度差异的权值量化

Table 1 Weighted quantization of multiorder path strength differences in directed networks

编号	有向路径	类型	连接强度
1	1→3→2	正向路径	1
2	1←4←2	逆向路径	λ^2
3	1↔5↔2	强连通路径	$(1+\lambda)^2$
4	1→6↔2	混合路径	$(1+\lambda)$
5	1←4←5↔2	混合路径	$\lambda^2\cdot(1+\lambda)$
6	1↔5→4←2	混合路径	$(1+\lambda)\cdot\lambda$
7	1↔5→6↔2	混合路径	$(1+\lambda)^2$
8	1→6←5↔2	混合路径	$\lambda\cdot(1+\lambda)$
9	1←4←5→6↔2	混合路径	$\lambda^2\cdot(1+\lambda)$
10	1→6←5→4←2	混合路径	λ^2

3.3 路径连接强度相似性预测指标

设网络中两节点 $i,j\in\boldsymbol{V}$ 未直接相连，而节点间存在多条有向路径，各路径对连边 e_{ij} 的建立具有一定的贡献，则节点 i 和节点 j 的相似分值可由节点间所有连接路径对连边连接强度的加权和刻画。同时，为了区分不同阶路径的贡献，相比短路径，路径越长其影响越微弱，通过一衰减参数 $\alpha>0$ 逐渐减弱长路径带来的影响，将路径连接强度的相似性指标定义为：

$$\begin{aligned}\boldsymbol{S}_{ij}^{DMP\text{-}L}&=\boldsymbol{W}^{2}+\alpha\cdot\boldsymbol{W}^{3}+\cdots+\alpha^{L-2}\cdot\boldsymbol{W}^{L}\\&=\sum_{l=2}^{L}\alpha^{L-2}\sum w(Path_{ij}^{L})\end{aligned} \tag{13}$$

其中，$\boldsymbol{W}=\boldsymbol{A}+\lambda\cdot\boldsymbol{A}^{\mathrm{T}}$，当 $\lambda=0$ 时，实际上忽略了网络中反向边的作用，仅计算节点间正向可达路径的数量。而当 $n\to\infty$ 时，同样令 $\alpha<\lambda_{\max}^{-1}$，$\lambda_{\max}$ 为矩阵 $\boldsymbol{W}$ 的最大特征值，使得该指标简写为：

$$\begin{aligned}\boldsymbol{S}_{ij}&=\sum_{l=2}^{n\to\infty}\alpha^{n-2}\sum w(Path_{ij}^{n})\\&=\frac{1}{\alpha}(\boldsymbol{I}-\alpha\cdot\boldsymbol{W})^{-1}-\frac{1}{\alpha}\boldsymbol{I}\end{aligned} \tag{14}$$

区别于无向网络仅统计节点间连边数量的 Katz 指标，上述指标以自适应参数 λ 调节了不同有向路径对节点对的连接强度贡献。

以图 5 为例，节点 1、节点 2 之间存在 3 条独立路径，根据

3.2 节中路径连接强度的定义，节点 1 作为始点时，3 条路径的连接强度分别计算为 $w_1=\lambda^2$，$w_2=1^3$，$w_3=1^2\cdot\lambda\cdot(1+\lambda)$，设定在 $\alpha=0.1$，$\lambda=0.2$ 的调节参数取值下，由节点 1 指向节点 2 的连边相似性分值将被具体计算为 $S_{12}=w_1+\alpha\cdot w_2+\alpha^2\cdot w_3=0.1424$，而以节点 2 作为始点时，3 条路径的连接强度分别计算为 $w_1=1^2$，$w_2=\lambda^3$，$w_3=1\cdot\lambda^2\cdot(1+\lambda)$，此时节点 2 指向节点 1 的连边相似性分值将被计算为 $S_{12}=w_1+\alpha\cdot w_2+\alpha^2\cdot w_3=1.0013$，因此节点 1、节点 2 之间产生节点 2 指向节点 1 的连边的可能性更高。因此，该方法在有向网络中能够量化不同路径的连接强度，从而区分有向网络中各条路径的差异，同时预测结果还可给出节点对连边的指向。

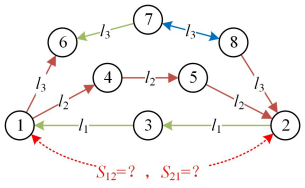


图 5 路径连接强度计算示例

Fig. 5 Example of path connection strength calculation

4 衡量标准及实验设置

预测算法的有效性需在真实网络数据集上进行验证，通常将网络 $G(\boldsymbol{V},\boldsymbol{E})$ 的连边集合 $\boldsymbol{E}$ 按一定比例 f 划分为训练集 $\boldsymbol{E}^P$ 和测试集 $\boldsymbol{E}^T$，$\boldsymbol{E}^P:\boldsymbol{E}^T=f:(1-f)$，将训练集 $\boldsymbol{E}^P$ 作为已知信息为未连边节点计算分值，有效的算法应当为测试集连边赋予更高的相似分值，而对非测试集的连边赋予较低的相似分值。

4.1 评价方式

预测算法的主流衡量标准包括 AUC(Area Under ROC Curve)[22] 及 Precision[23]。前者衡量算法区分未知对象能力；后者侧重精确度，关注预测前列结果命中的比例。二者的定义及计算过程如下。

AUC 指二分类预测器的 ROC 曲线下面积，在衡量链路预测算法性能时，随机抽取 n 组连边(一条来自测试集，一条来自不存在边集)进行对比计算，若测试集连边预测分值高于不存在边集的分值，记算法的有效次数 n' 加 1；反之则无效，不计有效次数；二者相等时，记相等次数 n'' 加 1，最终计算有效及相等次数在总次数中的比例，表示为：

$$AUC=\frac{n'+0.5n''}{n} \tag{15}$$

理想情况下，算法为测试集连边赋予的分值应高于不存在边集，$AUC=1$；随机预测时有 $AUC\approx0.5$，因此 AUC 越高算法的预测能力越好。

Precision 以另一种视角来衡量算法的精确度，如在推荐系统中，人们更在意推荐的前 L 个少数目标是否可信。设前 L 个预测结果中有 m 条连边属于测试集，则该算法的 Precision 的计算式为：

$$Precision=\frac{m}{L} \tag{16}$$

4.2 实验数据集

实验选取涵盖 4 种类别的 9 个真实数据集，这些数据集均是来自 Konect 网站[24] 的公开数据集。

(1) Highschool(HIG)[24]：美国伊利诺伊州一个高中的 70 名男生之间的友谊关系网络，节点表示男生，有向边表示该男生投票给自己认为的好友。

(2) ResidenceHall(RES)[24]：澳大利亚国立大学园区 217 名居民之间的友谊关系网络。

(3) Adolesent(ADO)[24]：1994—1995 年据某项调研建立的学生友谊关系网络。

(4) C. elegans(CEL)[24]：线虫神经元连接网络，节点表示该生物的某神经元，有向边表示神经元之间的连接。

(5) Yeast(YEA)[24]：酵母菌蛋白质相互作用网络。

(6) USAir(USA)[24]：2010 年美国机场间的航班网络，节点表示各机场，有向边表示机场之间的航线。

(7) Openflights(OPF)[24]：全球 2 939 个大型机场间的航班网络。

(8) SmaGri(SMA)[24]：有关“Small & Griffithand”主题论文引用网络，节点表示一篇论文，有向边表示该论文引用了其他论文。

(9) SciMet(SCI)[24]：“科学计量学”相关主题的引论文用网络。

表 2 列出了上述 9 个网络的相关统计特征，$|\boldsymbol{V}|$ 和 $|\boldsymbol{E}|$ 分别表示节点与连边的总数，$k_{\text{out}}^{\max}$ 和 $k_{\text{in}}^{\max}$ 分别表示节点的最大出度、入度，$\langle k\rangle$ 表示网络平均度，d 和 D 分别表示网络平均最短路径和网络直径，ρ 表示网络中互惠边所占的比例。

表 2 数据集统计特征

Table 2 Statistical characteristics of datasets

Data	$\|\boldsymbol{V}\|$	$\|\boldsymbol{E}\|$	$k_{\text{out}}^{\max}$	$k_{\text{in}}^{\max}$	$\langle k\rangle$	d	D	ρ
HIG	70	366	18	12	5.23	2.64	6	0.50
RES	217	2672	34	51	12.31	2.39	4	0.62
ADO	2539	12969	27	10	5.11	4.56	10	0.39
CEL	297	2345	134	39	7.90	2.46	5	0.17
YEA	332	2126	77	99	6.40	2.74	6	0
USA	2375	11693	103	99	4.92	5.10	15	0.08
OPF	2939	30501	236	237	10.38	4.10	14	0.97
SMA	1059	4918	89	232	4.64	2.98	6	0
SCI	3084	10412	121	104	3.38	4.18	12	0

5 实验结果与分析

5.1 参数影响分析

首先，为验证有向网络中不同类型连边是否对预测具有不同贡献，实验设置 $L\leqslant3$，路径衰减参数 $\alpha=0.01$，令 $\lambda\in[0,1]$，观察 DMP 指标预测效果随可调参数 λ 的变化曲线。

图 6 给出了 AUC-λ 的变化曲线，当 $\lambda=0$ 时，表示节点的反向边赋权为 0，此时等价于仅考虑正向路径的 LP 指标。当 λ 逐渐增大时，各数据集上 AUC 值均有显著提升，中间较为平稳，在某个取值处达到最高，表明 λ 取值对该指标整体预测能力的影响波动较小。图 6 中标注点给出了最优参数 λ_{opt} 及当前 AUC 值，如社交网络 HIG 和 RES 分别在 λ 为 0.43 和 0.44 时取得最好的 AUC 值，表示该类网络中反向边带来的连边贡献约为正向边的一半，而 ADO 网络中 $\lambda_{\text{opt}}=0.95$，OPF 网络中 $\lambda_{\text{opt}}=1$，表示将反向边的权重调节至与正向边大小相同时，预测效果最好，说明该网络中反向边与正向边对路径产生的影响一致。OPF 网络互惠系数为 0.97，说明网络中绝大部分连边为互惠边，当网络中全部为双向连接的互惠边时，

正向边和反向边区别的影响极小，或是使得正反连边权重相当的因素。同样在 CEL 网络中 $\lambda_{opt}=0.13$，USA 网络中 $\lambda_{opt}=0.14$，而 YEA，SMA，SCI 网络分别为 0.02，0.01，0.01，且这 3 个网络中互惠系数为 0，表明反向边的连边贡献极小但不可忽略，若忽略反向边或为反向边赋权过高均会导致 AUC 值的下降。

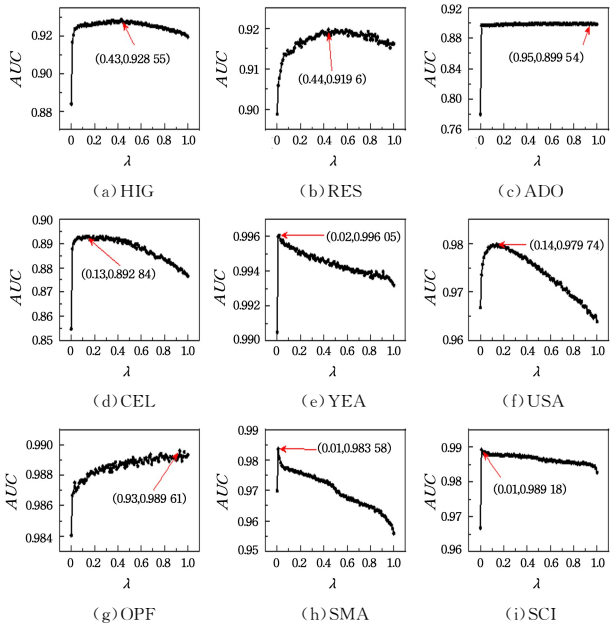


图 6 DMP-3 指标 AUC 值随参数 λ 的变化曲线

Fig. 6 AUC value of DMP-3 index varies with the parameter λ

图 7 给出了 Precision-λ 的变化曲线，与 AUC-λ 稍有不同，此时多个网络的 Precision 均展现出先上升后下降的趋势，波动幅度较大，表明 λ 取值与前列预测结果的精度相关联，此时 λ_{opt} 设置对 Precision 至关重要。λ_{opt} 在部分网络上接近于网络互惠系数 ρ，从图 7 所示结果来看，HIG 及 CEL 在 λ=0.4 附近精度较高，OPF 在 λ=0.99 处精度最好，而其他网络中可为 λ 赋值为 0.3。

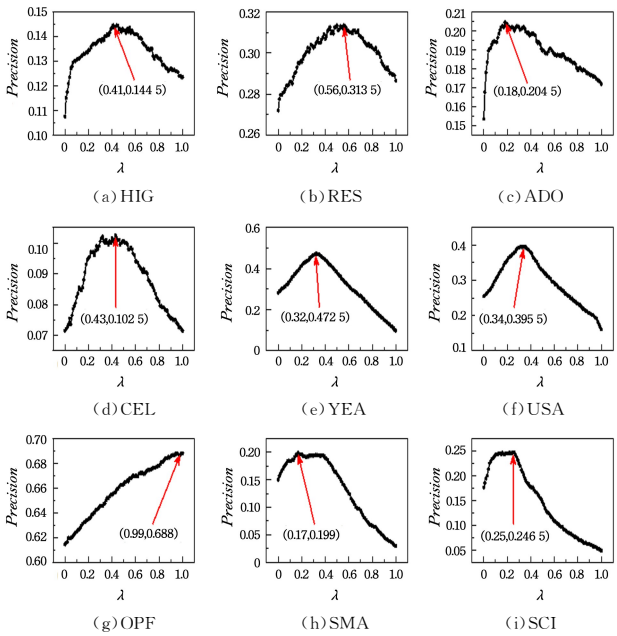


图 7 DMP-3 指标 Precision 值随参数 λ 的变化曲线

Fig. 7 Precision value of DMP-3 index varies with the parameter λ

整体而言，有向网络中的反向边对路径权重的影响并不完全为 0，通过自适应参数 λ 调节有向网络中不同类型连边的权重，能有效提高预测指标在不同衡量标准下的性能，在实际预测时，可依据网络类型及预测需求分别设置不同的 λ_{opt}。

5.2 对比结果分析

在不同有向网络上获得适应参数 λ_{opt} 后，进而比较 DMP-L 指标与其他预测指标的预测性能，分别选取 4 种局部、1 种准局部及 3 种全局指标与不同路径长度下的 DMP-L 进行对比，考虑到各个网络的平均最短路径为 2～6，取路径长度为 $2\leqslant L\leqslant 6$ 及 $L\to\infty$ 时的 DMP 指标来分析路径长度的影响，DMP-N 表示在 $L\to\infty$ 时的预测结果。

表 3 列出了分别在 9 个数据集上使用各个预测指标得到的 AUC 结果，明显观察到全局指标普遍优于局部、准局部指标的现象。DMP-N 考虑所有阶混合路径的贡献总和，在 6 个网络上取得了最高值，仅在 SMA 网络略低于全局 Katz 指标，反而在互惠系数接近于 1 的 OPF 网络上极差。考虑到 OPF 网络绝大多数为互惠边，此时有向连边的贡献几乎无差异，因此各指标预测结果近似于在该网络中对应无向拓扑的预测结果。其次，DMP 指标的收敛性或与网络的平均最短路径相关联，并观察到在路径长度由 2 增加到 3 时，DMP-3 的 AUC 值有较大幅度的提升，随后在 4～6 阶路径计算下 AUC 趋于稳定，在前 3 个社交网络中尤为明显，DMP 分别在 $L=3$(HIG)，$L=3$(RES)，$L=4$(ADO)时不在变化，而这 3 个网络对应平均最短路径分别为 2.64，2.39，4.56，故 L 大于等于平均最短路径时，DMP 的 AUC 值将趋于稳定。

表 3 不同数据集上各种预测方法的 AUC 结果对比

Table 3 Comparison of AUC results of each method in different datasets

Net	HIG	RES	ADO	CEL	YEA	USA	OPF	SMA	SCI
DCN	0.853	0.887	0.713	0.805	0.945	0.962	0.968	0.908	0.895
DAA	0.857	0.892	0.713	0.811	0.937	0.961	0.970	0.898	0.877
DRA	0.856	0.891	0.714	0.813	0.946	0.972	0.971	0.908	0.895
DPA	0.624	0.653	0.677	0.813	0.951	0.951	0.924	0.965	0.957
LP	0.889	0.895	0.778	0.864	0.990	0.965	0.984	0.970	0.967
Bifan	0.824	0.857	0.767	0.888	0.980	0.959	0.965	0.958	0.941
MFI	0.847	0.788	0.870	0.837	0.994	0.880	0.976	0.978	0.995
Katz	0.890	0.893	0.876	0.877	**0.998**	0.966	0.981	**0.990**	0.993
DMP-2	0.911	0.915	0.804	0.883	0.966	0.979	0.976	0.944	0.926
DMP-3	0.923	0.916	0.898	0.899	0.996	0.979	**0.989**	0.983	0.990
DMP-4	0.923	0.916	0.924	0.898	0.997	0.979	0.987	0.985	0.995
DMP-5	0.923	0.916	0.927	0.897	**0.998**	0.979	0.983	0.985	0.995
DMP-6	0.922	0.915	0.927	**0.899**	**0.998**	0.979	0.981	0.986	0.996
DMP-N	**0.925**	**0.941**	**0.928**	0.898	**0.998**	**0.986**	0.588	0.989	**0.998**

AUC 提升方面，尤以社交网络突出，HIG，RES 网络上多数指标的 AUC 为 0.85～0.89，ADO 网络只有 0.67～0.87，而 DMP-3 可在 3 个网络中达到 0.9，全局条件下的 DMP-N 可将 AUC 提升至 0.92 以上，相比较优的 Katz 提升幅度为 3.9%～6.0%，表明了社交关系的有向边作用确有差异，区分该差异的确可提高预测 AUC 值。其次 YEA 及引文网络 SMA 和 SCI 较为特殊，网络中互惠系数 $\rho=0$，因此预测指标不存在互惠边强联系的影响，各指标均有不错的预测效果，如局部指标 DCN 等在 SMA 网络接近 0.9，准局部指标 LP 和 Bifan 约为 0.96，而全局指标 MFI，Katz 已达 0.99，DMP-L 指标通过区分正向、逆向及混合路径不同连边贡献，当 $L=2$

时，AUC 可达 0.92，当 $L=3$ 时，可将 AUC 提升至 0.99，相比局部指标提升 0.9%～12.0%，相比准局部指标提升 0.9%～5.7%。而航空网络 USA 及 OPF 的平均最短路径较小，各机场通过枢纽存在直接或间接航班，共同邻居或路径拓扑等信息均可作为有效相似性参考，各指标 AUC 普遍在 0.95 之上，相比较优的全局指标 Katz 提升幅度为 0.8%～2.0%。

表 4 列出了在 9 个数据集上使用各个预测指标得到的 Precision 结果，DMP 指标在不同路径下的表现有别于 AUC。首先在社交网络 HIG，RES，ADO 中均是在 $L\to\infty$ 阶条件下取得最高精度，表明社交网络存在“小世界特性”，即潜在的间接人际关系对连边建立存在影响。其次是在其余网络中，准局部的表现更好，如 YEA 及 SMA 网络下 Bifan 的表现远优于其他指标，表明网络中的局部模体更适用于精确预测符合模体构成的少量连边，而 CEL，USA，SMA，SCI 网络中 DMP 在 2 阶局部条件下取得了最高的 Precision 值，反而在高阶路径场景下精度下降，说明多余的高阶路径对局部的预测精准性有负面作用。航空网络 OPF 同样较为特殊，DMP-N 有最差的 AUC 值，即0.588，却取得了最高的 Precision，即 0.92，而 USA 网络则不存在该现象，可观察到两个航空网络中互惠系数存在差异，USA 为 0.08，而 OPF 为 0.97，说明有向网络互惠系数接近 1 时，有向特性影响反而逐渐降低，导致部分指标出现异常。

表 4 不同数据集上各种预测方法的 Precision 结果对比

Table 4 Comparison of Precision results of each method in different datasets

Net	HIG	RES	ADO	CEL	YEA	USA	OPF	SMA	SCI
DCN	0.122	0.284	0.203	0.093	0.427	0.345	0.579	0.222	0.255
DAA	0.117	0.315	0.176	0.057	0.458	0.403	0.485	0.199	0.210
DRA	0.112	0.310	0.043	0.077	0.471	0.369	0.378	0.151	0.144
DPA	0.017	0.060	0.000	0.036	0.282	0.330	0.310	0.113	0.078
LP	0.103	0.280	0.175	0.086	0.473	0.318	0.627	0.197	0.240
Bifan	0.090	0.180	0.140	0.103	**0.791**	0.408	0.675	**0.307**	0.221
MFI	0.068	0.090	0.042	0.077	0.145	0.025	0.012	0.031	0.075
Katz	0.101	0.279	0.174	0.086	0.469	0.313	0.633	0.197	0.240
DMP-2	0.143	0.327	0.213	**0.107**	0.437	**0.409**	0.567	0.239	**0.296**
DMP-3	0.144	0.315	0.203	0.102	0.472	0.395	0.688	0.199	0.247
DMP-4	0.145	0.311	0.203	0.103	0.481	0.394	0.667	0.199	0.246
DMP-5	0.144	0.310	0.202	0.103	0.481	0.395	0.651	0.199	0.246
DMP-6	0.144	0.309	0.202	0.103	0.481	0.395	0.645	0.199	0.246
DMP-N	**0.148**	**0.415**	**0.247**	0.094	0.483	0.184	**0.920**	0.199	0.246

5.3 DMP 指标的鲁棒性分析

为验证 DMP 指标在不同训练集比例时的鲁棒性，令训练集比例 $f\in[0.65,0.9]$，相关指标对应的 AUC 变化趋势如图 8 所示。首先，各个网络上 DMP-N 及 DMP-3 的 AUC 曲线整体出现在其他指标上方，具有明显优势且保持稳定。随训练集比例逐步降低，网络的可用信息逐渐减少，大部分相似指标在多个网络上呈现下降趋势，而 DMP 指标在仅用原始网络 65%的信息时仍保持较高的 AUC。其他指标中，全局指标 MFI，Katz 的 AUC 波动不大，但低于 DMP 指标。类比来看，Katz 与 DMP-N 的 AUC 曲线走势接近，表明基于路径的预测指标一般鲁棒，但 Katz 无法区分有向网络混合路径的差异，其 AUC 曲线始终在 DMP 下方。

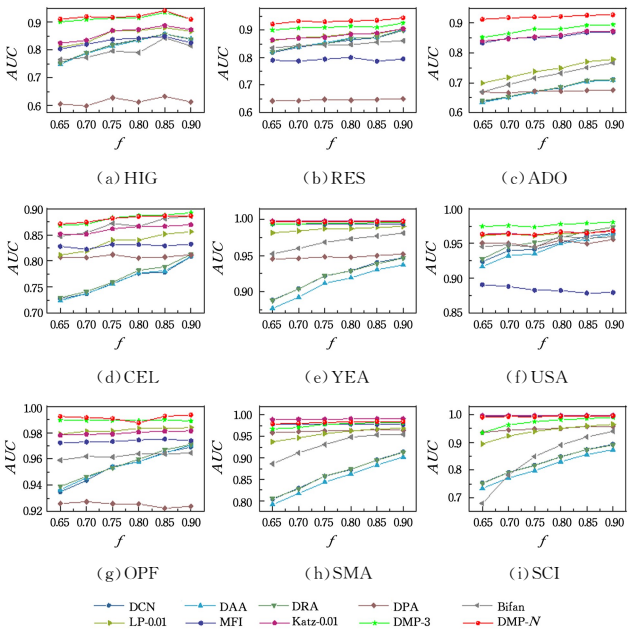


图 8 各指标在不同训练集比例 f 下的 AUC 曲线

Fig. 8 AUC curve in different training set proportions f

图 9 给出了各指标在不同训练集比例 f 下的 Precision 变化曲线，9 个网络上 DMP-3 整体优于其他对比指标，而 DMP-N 在 CEL，YEA，USA，SCI 网络上的精度较差。类似的是，局部相似指标普标精度较高，而全局指标，如 MFI 在 RES，ADO，USA 等网络精度上不及局部指标，说明全局信息在某些网络不适用于精准预测，反而引入冗余信息降低 Precision。另外，与 AUC 变化曲线相反的是，随训练集比例增大，测试集比例降低，可被命中的连边总数减少，因此多数指标 Precision 反而呈现下降趋势，整体来看 DMP 在 3 阶和 N 阶时的表现均较为稳定。总体而言，在两种评价标准下，DMP 指标在局部及全局条件下均表现出了较好的鲁棒性。

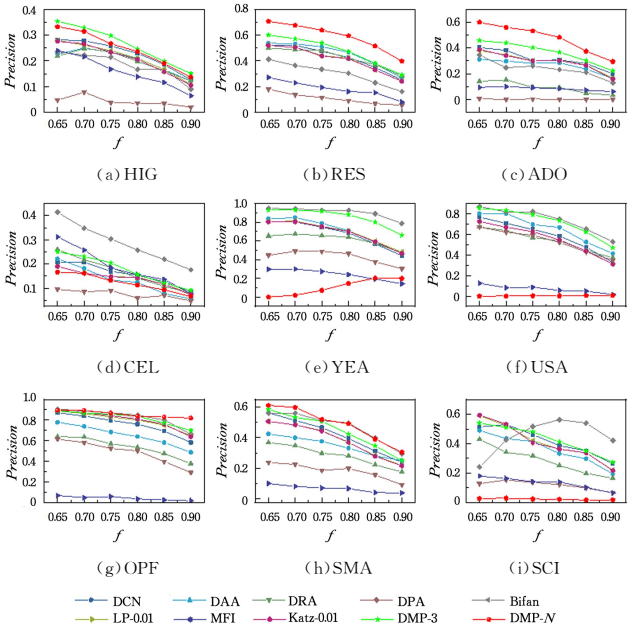


图 9 各指标在不同训练集比例 f 下的 Precision 曲线

Fig. 9 Precision curve in different training set proportions f

结束语 网络中的路径是节点建立连边的重要参考，目前

有向网络路径中的连边方向混合多变，需对多种不同的有向路径进行差异量化。本文通过有向边权为多阶路径计算各自的权重，提出了一种有向路径相似度预测指标 DMP-*L*，并依照网络平均路径等特征分析了不同路径长度下的预测结果。多个类型的真实网络数据集验证表明，在考虑不同路径的有向边差异后，算法的预测效果有了明显提升，全局条件的 DMP-*N* 普遍适用于 AUC 标准，而局部条件的 DMP-3 更适用于 Precision 需求，二者均呈现较好的鲁棒性。借助自适应权值参数 λ，该方法能够适用不同互惠边比例的网络，并有助于探索有向网络反向边与正向边的差异。未来，该工作将聚焦于混合路径上的随机游走过程，探究有向路径差异对随机游走过程的影响，并进一步提出更优的预测方法。

参 考 文 献

[1] TAN Y, WU J, ZHONG Q. Complex network [J]. Journal of Physics: Conference Series, 2020, 1601: 032011.

[2] FAN T, XIONG S, ZHAO W, et al. Information spread link prediction through multi-layer of social network based on trusted central nodes [J]. Peer-to-Peer Networking and Applications, 2019, 12(5): 1028-1040.

[3] DZAFERAGIC M, KAMINSKI N, MCBRIDE N, et al. A Functional Complexity Framework for the Analysis of Telecommunication Networks[J]. Journal of Complex Networks, 2018, 6(6): 971-988.

[4] CANNISTRACI C V, ALANIS-LOBATO G, RAVASI T. Erratum: From Link-Prediction in Brain Connectomes and Protein Interactomes to the Local-Community-Paradigm in Complex Networks: 1[J]. Scientific Reports, 2015, 5(1): 9794.

[5] WANG H, LE Z C, GONG X, et al. Review of Link Prediction Methods Based on Feature Classification[J]. Computer Science, 2020, 47(8): 302-312.

[6] LÜ L, ZHOU T. Link Prediction in Complex Networks: A Survey[J]. Physica A: Statistical Mechanics and Its Applications, 2011, 390(6): 1150-1170.

[7] 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.

[8] ZARE H, NIKOOIE M A, MORADI P. Enhanced recommender system using predictive network approach[J]. Physica A Statistical Mechanics & Its Applications, 2019, 520: 332-337.

[9] LESKOVEC J, HORVITZ E. Planetary-scale views on a large instant-messaging network[C]// Proceedings of the 17th International Conference on World Wide Web. 2008: 915-924.

[10] ZHANG X, ZHAO C, WANG X, et al. Identifying missing and spurious interactions in directed networks[C]// Proceedings of the 9th International Conference on Wireless Algorithms, Systems, and Applications. 2014: 470-481.

[11] SHANG K, SMALL M, YAN W. Link direction for link prediction[J]. Physica A: Statistical Mechanics and its Applications, 2017, 469: 767-776.

[12] ZHANG Q M, LÜ L, WANG W Q, et al. Potential theory for directed networks[J]. PloS One, 2013, 8(2): e55437.

[13] LI J, PENG J, LIU S, et al. Link Prediction in Directed Networks Utilizing the Role of Reciprocal Links[J]. IEEE Access, 2020, 8: 28668-28680.

[14] BÜTÜN E, KAYA M, ALHAJJ R. Extension of Neighbor-Based Link Prediction Methods for Directed, Weighted and Temporal Social Networks[J]. Information Sciences, 2018, 463/464: 152-165.

[15] PECH R, HAO D, LEE Y L, et al. Link Prediction via Linear Optimization[J]. Physica A: Statistical Mechanics and Its Applications, 2019, 528: 121319.

[16] LI X, LI P, ZHU Q. Directed Network Representation Method Based on Hierarchical Structure Information [J]. Computer Science, 2021, 48(2): 100-104.

[17] LÜ L, JIN C H, ZHOU T. Similarity index based on local paths for link prediction of complex networks[J]. Physical Review E, 2009, 80(2): 046122.

[18] 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.

[19] WANG K, LI X, LAN J L, et al. A New Link Prediction Method for Complex Networks Based on Topological Effectiveness of Resource Transmission Paths[J]. Journal of Electronics and Information Technology, 2020, 42(3): 653-660.

[20] LIU S, JI X, LIU C, et al. Extended Resource Allocation Index for Link Prediction of Complex Network[J]. Physica A: Statistical Mechanics and Its Applications, 2017, 479: 174-183.

[21] LIU S X, LI X, CHEN H C, et al. Link prediction method based on matching degree of resource transmission for complex network[J]. Journal on Communications, 2020, 41(6): 70-79.

[22] HANLEY J A, MCNEIL B J. The meaning and use of the area under a receiver operating characteristic (ROC) curve[J]. Radiology, 1982, 143(1): 29-36.

[23] LAWERA M. Predictive Inference: An Introduction[J]. Technometrics, 1995, 37(1): 121.

[24] KUNEGIS J. KONECT: the Koblenz network collection[C]// International Conference on World Wide Web Companion. ACM, 2013: 1343-1350.

ZHAO Xue-lei, born in 1996, postgraduate. His main research interests include complex network and link prediction.

LIU Shu-xin, born in 1987, Ph. D. His main research interests include network evolution and network behavior analysis.

（责任编辑：喻藜）