基于高阶路径相似度的复杂网络链路预测方法

顾秋阳[1,2]，吴宝[1,2]，池仁勇[1,2]

（1. 浙江工业大学管理学院，浙江 杭州 310023；2. 浙江工业大学中国中小企业研究院，浙江 杭州 310023）

摘 要：针对目前链路预测方法普遍存在精度不高、效率低等问题，提出了基于高阶路径相似度的复杂网络链路预测方法。首先，利用路径作为判别特征对复杂网络中的缺失链接进行预测，以实现资源的有效分配，并通过惩罚公共近邻对信息泄露进行限制。其次，将高阶路径作为判别特征，对种子节点对间的可用长路径实施惩罚。最后，利用多个真实复杂网络数据集进行数值算例。实验结果表明，与其他基线方法相比，所提方法具有更优的精度与效率。

关键词：高阶路径相似度；复杂网络；链路预测；相似性度量；公共近邻

中图分类号：TP391

文献标识码：A

DOI: 10.11959/j.issn.1000−436x.2021055

Link prediction method based on the similarity of high path

GU Qiuyang[1,2], WU Bao[1,2], CHI Renyong[1,2]

1. School of Management, Zhejiang University of Technology, Hangzhou 310023, China

2. China Institute for Small and Medium Enterprises, Zhejiang University of Technology, Hangzhou 310023, China

Abstract: For the problem that the existing link prediction method has many problems, including low accuracy and low efficiency, a method of high-order path similarity link prediction was proposed. Firstly, the path was used as the judging feature to predict missing links in complex networks, which could make resource allocation more effective and restricts information leakage by punishing public neighbor pairs. Secondly, by using high order paths as judging features, the available long paths between seed nodes would be punished. Finally, several real complex network datasets were used for numerical examples calculation. Experimental results show that the proposed algorithm is more accurate and efficient than other baseline methods.

Keywords: similarity of high path, complex network, link prediction, similarity measurement, common neighbor

1 引言

复杂系统普遍存在于人类社会中，常采用复杂网络对人类行为进行研究。随着近年来信息技术的不断发展，网络信息已成为大众获取信息、参与社交必不可少的媒介工具，而链路预测作为连接复杂网络与信息科学的重要桥梁，主要用于处理缺失信息的还原与预测，相关研究已受到国内外学者的广泛关注。复杂网络可对社会结构、生物系统和信息系统等进行有效建模，其中节点可表示网络中的个体或群体，边表示节点间的关系[1]。复杂网络研究已成为计算机科学、物理学、系统科学等学科的研究热点。现有研究多着重于复杂网络的演化机制、拓扑结构等特征，而链路预测则为其中的基本问题。在生物信息学领域中，链路预测在传染病、蛋白质等研究中都具有重要意义。链路预测在社会网

收稿日期：2020−09−22；**修回日期**：2021−01−27

基金项目：国家自然科学基金资助项目（No.71173194）；国家社科基金资助项目（No.20VYJ073, No.17ZDA088）；浙江省社科规划重点基金资助项目（No.20NDJC10Z）

Foundation Items: The National Natural Science Foundation of China (No.71173194), The National Social Science Foundation of China (No.20VYJ073, No.17ZDA088), The Social Science Planning Key Foundation of Zhejiang Province (No.20NDJC10Z)

络的相关研究中应用于好友预测及推荐。此外，还可通过链路预测实现科研合作推荐、知识产权推荐等[2]。链路预测技术有助于探测网络中个体间存在的潜在关系，可用于恐怖分子间的关系发现，以有效预防和制止犯罪行为[3]。由此可知，复杂网络链路预测技术在不同应用场景中存在广泛应用案例。

现有复杂网络常用图表示，其中节点表示个体或群体，边表示个体或群体间的交互行为。由于真实复杂网络中一般存在不断进入和移除的节点和边，故导致了系统的复杂性[4]。现今应用在复杂网络中的链路预测方法普遍存在精度不高、计算时间过长等问题，故提升链路预测的精度和效率是学术界亟待解决的重要问题。学术界通常将静态链路预测定义为被观测网络中寻找节点间的缺失链接，将动态链路预测定义为基于现有复杂网络是否可预测节点间存在未来链路的可能性。计算简单、效率高的基于相似性的链路预测方法最常见。通常包括基于邻域的链路预测方法（如公共近邻（CN, common neighbor）[5]、Jaccard 法[6]、Adamic/Adar（AA）法[7]、资源分配（RA, resource allocation）法[8]、偏好依附（PA, preferential attachment）法[9]等）和基于路径的链路预测方法（如 Katz 指数[10]等），以探索网络中的全局信息。类局部方法采用的信息量与全局方法相同，计算效率与局部方法相同，故通常认为其为局部方法和全局方法间的权衡。

目前，已有较多文献对复杂网络中的链路预测问题提出了解决方法。基于相似性的链路预测方法中，节点对间相似性得分根据网络的局部/全局拓扑结构特征值计算得到。为提高链路预测方法性能，现有方法常在复杂网络拓扑特征中增加附加信息[11]，这在提升算法精度的同时也会使计算更加复杂。翟东升等[12]认为，从较长路径中提取较短长度的路径更具相关性，而高阶路径数量是由更复杂的线性指标决定的。Wu 等[13]以公共邻居节点聚类系数的形式提取三角结构信息，并利用真实复杂网络数据集进行验证。Rahimi 等[14]指出，根据网络拓扑结构，高连接节点的影响要小于网络中心度较高但连接数量较少的节点。Kumar 等[15]将聚类系数推广到存在高阶路径的复杂网络拓扑结构中以提取局部路径信息，并将扩展聚类系数概念应用于链路预测中。

通过对现有研究成果的梳理发现，复杂网络中的链路预测方法已受到了国内外学者的重视并得到了丰富研究。上述研究成果虽然对本文具有一定的借鉴意义，但也存在一些不足。首先，已有很多学者对链路预测方法开展了一系列研究，但多数集中在基于路径相似和公共近邻的方法（如吴翼腾等[16]），很少有利用高阶路径作为判别特征，并对种子节点对间的可用长路径实施惩罚进行链路预测的文献记录。其次，几乎所有的相关文献都只基于复杂网络图特征对链路预测方法提出了改进（如李永立等[17]），很少有在网络资源分配过程的启发下基于高度路径相似度进行链路预测方法设计的文献记录。最后，现有文献多在链路预测过程中使用了路径相似度指数，很少有在计算过程中基于公共近邻的连接惩罚限制信息泄露，以最大化描述节点间相似度节点对间信息流的文献记录。本文利用路径作为判别特征对复杂网络中的缺失链接进行有效预测，提出了基于高阶路径相似度的链路预测（HPS-LP, high-order path similarity link prediction）算法。然后通过惩罚公共近邻对信息泄露进行限制，以最大化描述节点间相似度节点对间的信息流。最后以高阶路径（定义为六度分隔理论）作为判别特征，对种子节点对间的可用长路径实施惩罚。

2 算法设计

大多数社会网络中会表现出小世界、聚类和无标度等拓扑性质。本文利用不同长度的路径来计算复杂网络中节点间的相似度，基于网络中的资源分配过程向目的地发送信息[8]。现有用于复杂网络环境的链路预测方法普遍存在精度不高、计算时间过长等问题，且还不能对信息泄露等问题进行有效控制。本文基于资源分配过程进行设计，根据目标节点接收到的信息量进行相似性分析。通过公共邻节点间的信息泄露来最大化节点间相似性，以提升复杂网络链路预测精度和效率，对信息泄露进行惩罚以实现有效控制。长度为 2 的路径相似度可作为公共邻度，故可将任意节点对间路径长度为 2 的相似性得分表示为

$$S_{x,y}^{2}=\frac{1}{\sum_{z=1}^{z}k_z} \tag{1}$$

其中，k_z 表示节点 z 的度。而高阶路径相似度每一条长度大于 2 的路径都可分解为长度为 l−1 的路径

和与其相连的一条边。而连接 2 个节点的路径相似度得分为其分解所得长度为 l−1 的路径的相似度得分与构成边的路径相似度得分的乘积，而较长的路径得分可表示为

$$S_{x,y}^{l-1}=\psi\sum_{l=2}^{l-1}f_1f_2\left(l-1\right) \tag{2}$$

其中，f_1和 f_2分别为路径边与上一周期中路径边的相似性得分，最高路径阶数为 6（由六度分隔理论定义），ψ为惩罚较长路径的惩罚参数。路径相似度得分在第一次迭代中由路径长度为 2 的得分计算得到（具体如式(1)所示）。路径长度得分 f_1和 f_2的累积效应可得到长度为 3 的路径相似性得分。在本文所提算法中，使用类似方法迭代计算高阶路径相似度，如算法 1 所示。

算法 1　HPS-LP 算法

输入　复杂网络图 G

输出　得分矩阵

1) 计算路径为 2 的相似度得分

2) 对每个节点对(*i*,*j*)do

3) 将公共近邻赋值为 z

4) 计算路径得分

5) 基于步骤 4)结果计算高阶路径得分

6) 循环计算高阶路径长度

7) 计算邻节点的高阶路径相似度得分

8) 结合上期末的路径相似度得分进行如式(2)所示计算

9) 基于步骤 8)结果循环更新路径得分

10) 进行参数更新

11) 返回网络得分矩阵

算法 1 中的输入为复杂网络图，输出为包含所有节点对得分的矩阵。该算法主要包括初始化、计算和更新 3 个阶段，初始化阶段在网络的每个节点对间为矩阵 Score 分配长度为 2 的相似度得分，计算阶段基于 2 个长度路径得分矩阵迭代计算更高路径长度分数，更新阶段根据前一阶段计算分数迭代更新上述 2 个矩阵。

3　数值算例

3.1　数据说明与评价标准

本文使用国内外共 10 个相关数据集对所提 HPS-LP 法进行验证。本文利用 Python 工具从新浪微博（爬取时间为 2020 年 4 月 6 日至 2020 年 8 月 12 日）、Twitter（爬取时间为 2020 年 3 月 14 日至 2020 年 7 月 29 日）、Facebook（爬取时间为 2020 年 2 月 10 日至 2020 年 8 月 19 日）和 Dolphins（爬取时间为 2020 年 3 月 27 日至 2020 年 7 月 13 日）的 API 端口，以“华为”和“HUAWEI”为关键词选取 30 个节点作为初始节点，爬取真实用户数据集作为实验的基础数据。Netscience[8]是 2006 年编制的网络理论和实验的合著者网络，SmaGri[4]是 HistCite 软件公司 2009 年生产的 Garfield collection 引文网络，为科学网络搜索的结果。hep-ph 与 astro-ph 表示 2004—2009 年科研论文作者合作网络，其中 hep-ph 为物理现象学领域，astro-ph 为天体物理学领域，且使用了至少撰写 3 篇论文及以上的作者作为节点形成的网络[18]。dblp-collab 来自 1999—2003 年 DBLP 计算机科学文献[19]，其中 dblp-collab 为计算机科学作者合作网络。本文使用五重交叉验证程序来评估所提算法的性能，并将所有数据以 CSV 格式保存在 MySQL 数据库中以便进行数据处理。使用 Rapidminer 数据挖掘工具随机选取各用户评级数据的 20%作为测试集，并将剩下的

表 1　数据集统计信息

网络名称	节点个数/个	边数/条	平均路径	平均度	凝聚系数
Facebook	32 849	554 830	3.910	45.432	0.032
Twitter	39 491	673 922	4.398	39.029	0.005
新浪微博	40 393	784 994	5.343	68.934	0.046
Dolphins	31 929	493 809	4.285	40.293	0.001
SmaGri	1 365	5 943	2.911	12.045	0.047
Netscience	2 149	5 390	4.377	3.091	0.027
hep-ph	15 393	239 840	2.301	13.621	0.031
astro-ph	15 393	493 208	3.293	12.389	0.008
dblp-collab	392 801	3 102 930	4.390	9.286	0.001

80%用户数据作为训练集。表 1 为数据集统计信息。

链路预测问题通常被视为二分类任务，故用于二分类任务的大部分评估指标都可用于链路预测评估。在具有 2 个类别的二分类任务的评估混淆矩阵中[13]，真正例（TP, true positive）表示正确预测链接的数量；真负例（TN, true negative）表示正确的未预测链接的数量；假正例（FP, false positive），表示错误预测链接的数量；假负例（FN, false negative）表示错误的未预测链接的数量。基于此，可得真正例率（TPR, true positive rate）、假正例率（FPR, false positive rate）、真负例率（TNR, true negative rate）和精确率（precision）等，其计算式可参考文献[20]。基于以下 2 个指标进行评估，即 ROC 曲线下面积（AUROC, area under the receiver operating characteristics curve）[12]和平均精度（AP, average precision）[15]。ROC 曲线是 Y 轴上的真正例率（敏感性）和 X 轴上的假正例率（1-特异性）间的曲线，曲线下面积值为[0,1]的数据可使用梯形规则累加计算得到。曲线下面积值越高，链路预测方法的性能越好。平均精度是基于不同召回阈值计算的单点汇总值，为区间[0,1]召回值的平均精度值。

3.2 基准算法

为体现本文所提 HPS-LP 法的优越性，将所述方法与多种非监督结构链路预测算法进行对比。

1) CN 法。在给定的复杂网络或图中，给定一对节点 x 和 y 的公共邻居的规模被计算为 2 个节点邻节点的交集大小，如式(3)所示。

$$S_{x,y}^{\mathrm{CN}} = \Gamma(x) \cap \Gamma(y) \tag{3}$$

其中，$\Gamma(x)$ 和 $\Gamma(y)$ 分别为节点 x 和 y 的邻居，在 x 和 y 间存在联系的可能性随着其间共有邻居的数量而增加。

2) AA 法。该方法为一种基于共享特征的相似度计算方法，并将其用于链接预测中，具体计算方法如式(4)所示。

$$S_{x,y}^{\mathrm{AA}} = \sum_{z=1}^{z} \left[\frac{1}{\log(z)} \right] \tag{4}$$

其中，z 表示节点 x 和 y 的公共邻节点，由此可知更多的权重被分配给阶数较小的公共邻节点。

3) Jaccard 系数法[6]。此度量方法与公共近邻法相同，不同之处在于，如果 2 个节点有较多的公共邻点和较少的非公共邻点，则其相似度较大。在将上述相似度分数标准化后，可表示为

$$S_{x,y}^{\mathrm{JC}} = \frac{\Gamma(x) \cap \Gamma(y)}{\Gamma(x) \cup \Gamma(y)} \tag{5}$$

其中，Jaccard 系数被定义为从任意一个节点的所有邻节点中选择成对节点的公共邻节点的概率。

4) 优先连接（PA, preferential attachment）[9]。其将优先依附的思想用于生成一个不断增长的无标度网络中，节点度数是预测新链接的关键属性。度数越高的 2 个节点在未来彼此交互的可能性越大，添加与节点 x 和 y 相关的新链接的可能性与节点度 $k(x)$ 和 $k(y)$ 成正比。节点 x 和 y 间的优先链接分数为

$$S_{x,y}^{\mathrm{PA}} = k(x)k(y) \tag{6}$$

其中，节点特征值上的聚合函数可用于计算链接特征值。

5) CAR 指数[11]。基于两节点共同近邻为本地社区成员，则两节点间的链路更可能存在的假设进行设计，且会随着种子节点对间的公共近邻链路数量而增加，具体如式(7)所示。

$$S_{x,y}^{\mathrm{CAR}} = S_{x,y}^{\mathrm{CN}} \sum_{z=1}^{z} \frac{\gamma(z)}{2} \tag{7}$$

其中，$\gamma(z)$ 为节点邻域子集。

6) Katz 指数[10]。该指标可看作最短路径度量的改进，可直接聚集 x 和 y 间的所有路径，并对较长路径进行指数转储惩罚，具体如式(8)表示。

$$S_{x,y}^{\mathrm{Katz}} = \sum_{t=1}^{\infty} \beta A_{x,y} \tag{8}$$

其中，β 表示控制路径权重，A 表示邻接矩阵。

7) 本地路径指数（LP, local path index）[6]。本地路径指数法为局部与全局链路预测方法在精度和计算复杂度间的良好折中。如 x 和 y 间没有直接联系，表示 x 和 y 间长度为 n-1 的不同路径，该指数也可扩展为如式(9)所示的形式。

$$S_{x,y}^{\mathrm{LP}} = \sum_{n=1}^{\infty} \left(\varepsilon^{n-1} A^{n+1} \right) \tag{9}$$

其中，n 为最大阶数。

8) Node2vec（N2V）[16]是一种节点嵌入技术，学习图中节点的低维连续表示，目的是为了保持邻域结构，并将有偏随机游动作为抽样策略。其中共存在 4 个参数，即行走次数（即为每个节点生成的随机游动数）、游动长度（即每次随机游动中的节

点数）、返回超参数 p 和输入输出超参数 q。

9) 扩展资源分配（ERA, extended resource allocation）[17]是一种 2 个节点间通过本地路径传输的潜在资源，基于节点间的资源交换，提出了一种扩展的资源分配指标，具体如式(10)所示。

$$S_{xy}^{\mathrm{ERA}}=\sum_{z\in C_{xy}}\frac{2+\sigma(n_{zy}+n_{zx})}{k_z}+\sum_{C_{xy}^{'}}\frac{\sigma(n_{zy}+n_{zx})}{k_z} \tag{10}$$

10) 资源传输容量（PIC, potential information capacity）[21]在考虑信息通道和容量的情况下，提出了信息容量的定义，用以量化节点间的信息传输能力，具体如式(11)所示。

$$S_{xy}^{\mathrm{PIC}}=\sum_{Z_y\in\Gamma(y)}\left(a_{xz_y}+\frac{n_{xz_y}}{k_{z_{xz_y}}^{'\max}-1}\right)^{\beta}+\sum_{Z_x\in\Gamma(x)}\left(a_{yz_x}+\frac{n_{yz_x}}{k_{z_{yz_x}}^{'\max}-1}\right)^{\beta} \tag{11}$$

11) Propflow 指数[22]。基于流随机游走的链路算法通过将粒子的随机游走限制在有限步以内，当游走时遇到已游走过的节点或者回到原始节点则停止游走。

3.3　实验结果

本文分别在 Python 环境中使用 Scipy[11]、Numpy[6]和 LPmade 工具包[13]执行上述算法和本文所提 HPS-LP 法。然后利用上述评价指标，对不同的真实网络数据集进行评估，并对不同参数进行了敏感性分析。由于实验中 HPS-LP 法预测列表在每次运行时的结果都有可能不同，故设置评估结果为迭代 500 次运行后的平均值，运行的平均标准差为 1.286。表 2 为各数据集在不同方法中的曲线下面积值（其他算法的曲线下面积值都差于表中算法，故在此不再赘述）。从表 2 可以看出，本文所提 HPS-LP 法在社交网络数据集中都具有较优的实验结果。在科研协作网络中虽然结果也较好，但不能保证最优。hep-ph 科研论文作者合作网络的聚类系数较低，故具有长度 3 的路径较少，故本文所提方法在上述 2 个数据集中的实验结果相对较差，而在其他聚类系数较大的网络中有较好表现。除本文所提 HPS-LP 法外，LP 法与 PIC 法次之，且分别在 hep-ph 数据集和 Twitter 数据集中具有较好表现，这是由于 LP 法作为一种类局部链路预测法在非大型数据集中具有较优效果，PIC 法作为一种重视信息传输的链路预测方法在信息资源爆炸的社交网络数据集中具有较优效果。而基于 CN 的链路预测算法普遍效果较差，这是由于在现今信息爆炸的状况下，受各种拓扑特征和系统动力学要素的影响，只考虑 CN 并不足以涵盖预测所需的信息。

表 3 为各数据集在不同方法中的平均精度值

表 2　　**各数据集在不同方法中的曲线下面积值**

方法名称	Facebook	Twitter	新浪微博	Dolphins	SmaGri	Netscience	hep-ph	astro-ph	dblp-collab
LP	0.944	0.932	0.935	0.728	0.849	0.905	**0.977**	0.936	0.811
N2V	0.864	0.873	0.918	0.791	0.759	0.875	0.956	0.830	0.719
ERA	0.836	0.892	9.835	0.823	0.817	0.795	0.932	0.912	0.802
PIC	0.874	0.951	0.856	0.846	0.832	0.846	0.913	0.868	0.831
Propflow	0.814	0.823	0.864	0.873	0.866	0.834	0.9431	0.895	0.854
HPS-LP	**0.969**	**0.964**	**0.956**	**0.844**	**0.938**	**0.970**	0.954	**0.954**	**0.889**

注：粗体显示的值均表明其所对应的模型性能良好。

表 3　　**各数据集在不同方法中的平均精度值**

方法名称	Facebook	Twitter	新浪微博	Dolphins	SmaGri	Netscience	hep-ph	astro-ph	dblp-collab
LP	0.349	0.297	0.082	0.074	0.074	0.473	**0.539**	0.075	0.385
N2V	0.146	0.192	0.075	0.059	0.255	0.110	0.122	0.128	0.218
ERA	0.228	0.251	0.134	0.175	0.318	0.393	0.435	0.214	0.283
PIC	0.212	0.202	0.201	0.162	0.285	0.264	0.334	0.248	0.229
Propflow	0.241	0.284	0.139	0.192	0.263	0.255	0.268	0.252	0.134
HPS-LP	**0.364**	**0.419**	**0.539**	**0.368**	**0.474**	**0.469**	0.472	**0.694**	**0.519**

注：粗体显示的值均表明其所对应的模型性能良好。

（其他算法的精度都差于表中算法，故在此不再赘述）。结果显示，本文所提 HPS-LP 法在复杂网络数据集中具有较高的平均精度值，而在聚类系数较低的科研协作网络稀疏数据集中的平均精确度较低。除本文所提 HPS-LP 法外，LP 与 ERA 法的精度次之，都在科研合作网站中具有较优效果，其原因是 LP 法作为一种类局部链路预测法在非大型数据集中具有较优效果；ERA 法作为一种基于节点间本地路径传输潜在资源的链路预测方法，故能有效对科研合作网络中的潜在合作链路实现预测。基于 CN 的链路预测算法普遍效果较差，这是由于在信息爆炸的现状下，受各种拓扑特征和系统动力学要素的影响，CN 并不足以涵盖预测所需的信息。N2V 法作为一种节点嵌入技术，为保持邻域结构牺牲了部分预测精度，故在大型数据集中性能较差。

3.4　敏感性分析

针对参数值ψ的影响，本文将不同长度的路径作为特征属性来计算网络中两节点间的相似度。本文就参数值ψ在区间[0.0,1.0]范围内对曲线下面积和精度值的影响进行了分析。图 1 为不同复杂网络中基于曲线下面积值的惩罚参数敏感性分析（由于其他算法的曲线下面积敏感性参数都远差于本文所提算法，故在此不再赘述）。由图 1 可知，本文所提 HPS-LP 法在社交网络和科研协作网络中受参数值ψ影响不显著，而 LP 法受参数值ψ的稳定性仅次于 HPS-LP 法。相比其他网络，链路预测方法普遍在社交网络数据集中具有更优表现。PIC 法受参数值ψ的稳定性较差，这是由于其需考虑信息通道与信息容量。

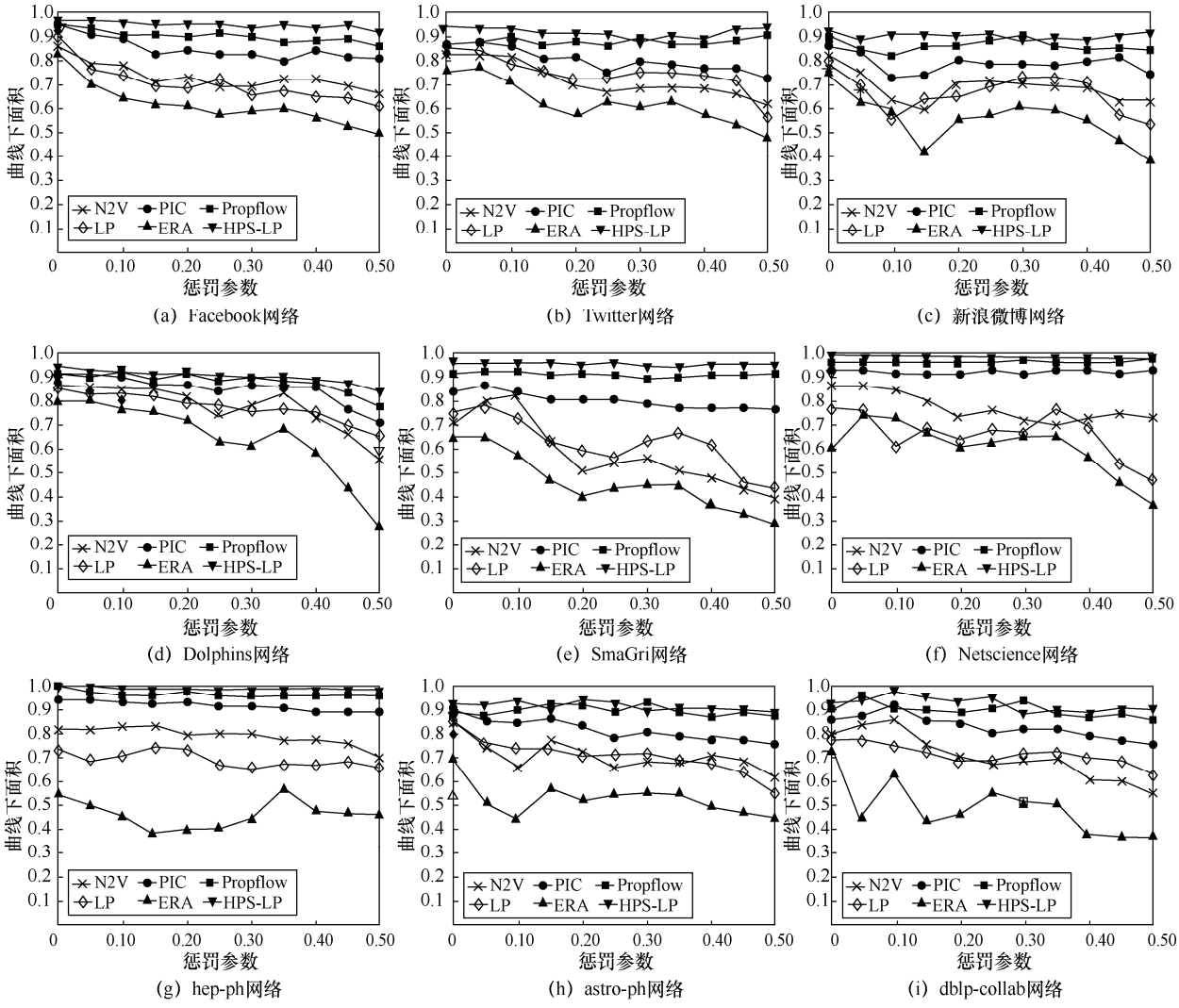

图 1　不同复杂网络中曲线下面积值的惩罚参数敏感性分析

图 2 为不同复杂网络中精度值的惩罚参数敏感性分析（由于其他算法的精度敏感性参数都远差于本文所提算法，故在此不再赘述）。由图 2 可知，本文所提 HPS-LP 法精度最高，LP 法其次。Katz 法随惩罚参数值的变化较大，这本质上是由于其为一种最短路径度量方法，故预测结果存在不稳定性。相对社交网络，惩罚参数值对科研协作网络的影响较大，这是由于科研合作网络具有较低度值，故算法稳定性更高。

本文将路径视为特征参数之一，认为在大多数复杂网络中任何两节点间的路径平均长度为 6（即六度分隔理论），故路径长度可达 6。表 4 为短路径 ($l \leqslant 3$) 与长路径 ($l > 3$) 下的链路预测精度。由表 4 可知，社交网络中短路径和长路径的曲线下面积值值没有显著差异；科研协作网络短路径和长路径的曲线下面积值存在显著差异。当考虑较长路径时，

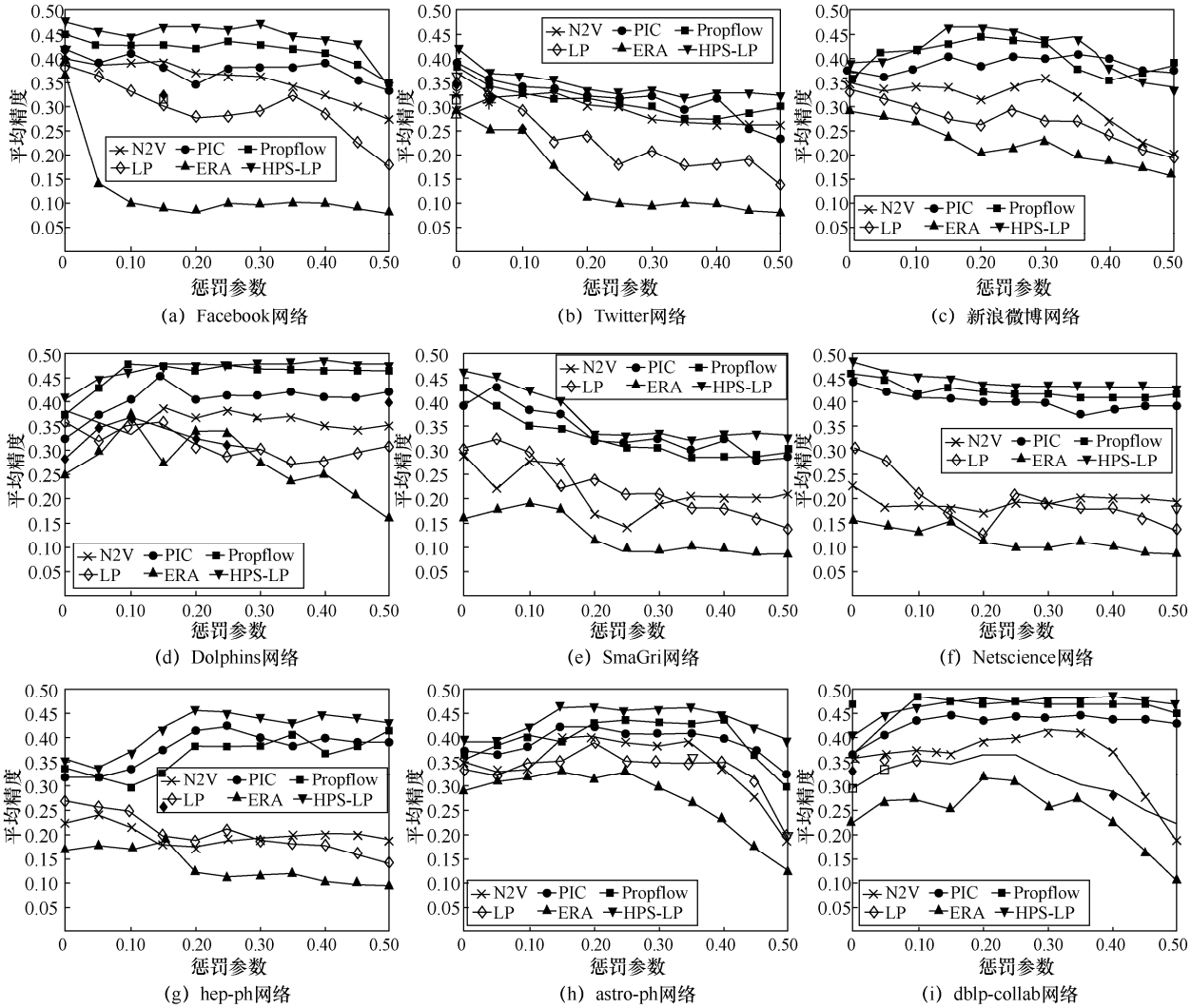

图 2　不同复杂网络中精度值的惩罚参数敏感性分析

表 4　　路径长度对链路预测精度的影响

指标名称	链路数量	Facebook	Twitter	新浪微博	Dolphins	SmaGri	Netscience	hep-ph	astro-ph	dblp-collab
ROC 曲线下面积	短路径	0.954	0.949	0.930	0.825	0.757	0.980	0.983	0.948	0.839
	长路径	0.977	0.961	0.984	0.914	0.779	0.989	0.993	0.959	0.880
平均精度	短路径	0.343	0.176	0.130	0.154	0.041	0.079	0.387	0.040	0.034
	长路径	0.305	0.154	0.119	0.190	0.043	0.076	0.390	0.038	0.045

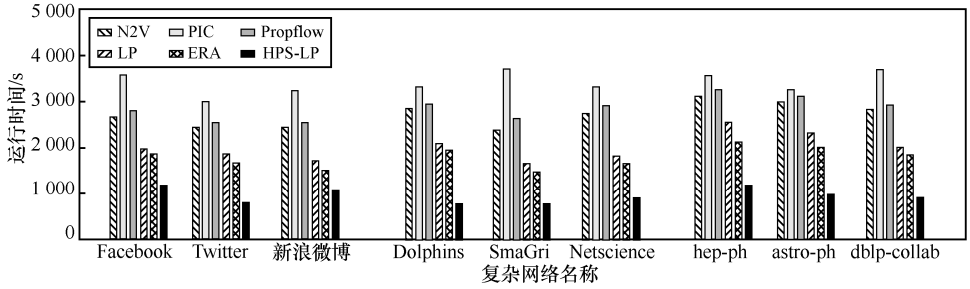

图 3 时间复杂度的算法比较结果

曲线下面积值在大多数科研协作网络中甚至会降低，链路预测方法的精度没有显著提高。

本文基于以下假设对时间复杂度进行了讨论，即大多数复杂网络都为稀疏的，并将每个节点的平均边数设为 n（网络中节点数的阶数）。最高阶路径 $l_{\max}$ 是节点对间路径最大长度，超过最高阶路径则定义认为相似性概率对高阶路径不产生影响。本文所提 HPS-LP 法的关键是计算 2 个及以上的路径长度分数。算法 1 会循环迭代至 $O(n^2)$ 次，普通节点对的 CN 值的成本为 $O(n^2K)$。对于阶数更高的路径，时间复杂度由 2 个循环使用时间 $O(n^2K)$ 和 $O(n\lg n)$ 组成，具有较高路径的总时间复杂度为 $O(n^2K)+O(n^2)$，故本文所提 HPS-LP 法的总时间复杂度为 $O(n^2K)$。基准算法的时间复杂度如文献[14]所示，在此不再赘述。时间复杂度的算法比较结果如图 3 所示。

4 结束语

复杂网络中的拓扑结构和演化会受到各种拓扑特征和系统动力学要素的制约。现有基于公共近邻的链路预测方法普遍存在效率较低、维数灾难等问题。而当今信息爆发的现状对复杂网络链路预测算法的运行效率和精度提出了更高的要求。社交关系的重要性则要求必须在链路预测算法中考虑边权重的影响。由于有关研究说明用户间的弱关系具有极高的商业价值，故对高阶路径中的链路预测及其隐私保护提出了更高的要求。故在本文所提 HPS-LP 方法中，利用路径作为判别特征来预测复杂网络中的缺失链接；并以资源分配过程为目标，通过基于公共近邻的连接惩罚限制信息泄露，以最大化描述节点间相似度的节点对间的信息流。高阶路径（基于六度分隔理论）也被用作判别特征，并对其应用惩罚函数。本文在多个真实复杂网络中进行了仿真实验，结果表明所提 HPS-LP 法优于基准方法，且考虑高阶路径相似度能有效提高其对复杂网络的链路预测精度，并显著影响计算时间复杂度。

尽管本文已提出了上述具有重要意义的发现，但还具有一定局限性，其中一些可能会为未来的进一步研究指明方向。首先，本文所提 HPS-LP 法利用路径作为判别特征进行链路预测，故在后续研究中可结合更多具有路径复杂性的复杂网络数据集来对本文所提链路预测方法进行验证。其次，由于本文所提 HPS-LP 法应用了公共近邻思想，故未来可尝试利用神经网络或探索异构动态网络嵌入技术对链路预测方法进行进一步优化，以消除可能存在的维数灾难等问题。最后，由于本文所提方法以资源分配过程为目标，故可结合社会化调查法和计算实验等研究框架，分别在有/无向图中对所提链路预测方法进行进一步佐证。

参考文献：

[1] 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.
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 & Information Technology, 2020, 42(3): 653-660.

[2] 胡钢, 高浩, 徐翔, 等. 基于重要度传输矩阵的复杂网络节点重要性辨识方法[J]. 电子学报, 2020, 48(12): 2402-2408.
HU G, GAO H, XU X, et al. Importance identification method of complex network nodes based on importance transfer matrix[J]. Acta Electronica Sinica, 2020, 48(12): 2402-2408.

[3] IMTIAZ Z B, MANZOOR A, ISLAM S U, et al. Discovering communities from disjoint complex networks using Multi-Layer Ant Colony Optimization[J]. Future Generation Computer Systems, 2021, 115: 659-670.

[4] AN C, O'MALLEY A J, ROCKMORE D N. Towards intelligent complex networks: the space and prediction of information walks[J]. Applied Network Science, 2019, 4(1): 35.

[5] 舒坚, 张学佩, 刘琳岚, 等. 基于深度卷积神经网络的多节点间链路预测方法[J]. 电子学报, 2018, 46(12): 2970-2977.

SHU J, ZHANG X P, LIU L L, et al. Multi-nodes link prediction method based on deep convolution neural networks[J]. Acta Electronica Sinica, 2018, 46(12): 2970-2977.

[6] JACCARD P. Distribution de la Flore Alpine dans le Bassin des Dranses et dans quelques régions voisines[J]. Bulletin De La Societe Vaudoise Des Sciences Naturelles, 1901, 37(140): 241-272.

[7] 郭丽媛, 王智强, 梁吉业. 基于边重要度的矩阵分解链路预测算法[J]. 模式识别与人工智能, 2018, 31(2): 150-157.
GUO L Y, WANG Z Q, LIANG J Y. Link prediction algorithm by matrix factorization based on importance of edges[J]. Pattern Recognition and Artificial Intelligence, 2018, 31(2): 150-157.

[8] ZHOU T, LU L, ZHANG Y C. Predicting missing links via local information[J]. European Physical Journal B, 2009, 71(4): 623-630.

[9] 王智强, 梁吉业, 李茹. 基于信息融合的概率矩阵分解链路预测方法[J]. 计算机研究与发展, 2019, 56(2): 306-318.
WANG Z Q, LIANG J Y, LI R. Probability matrix factorization for link prediction based on information fusion[J]. Journal of Computer Research and Development, 2019, 56(2): 306-318.

[10] IMTIAZ Z B, MANZOOR A, ISLAM S U, et al. Discovering communities from disjoint complex networks using Multi-Layer Ant Colony Optimization[J]. Future Generation Computer Systems, 2021, 115: 659-670.

[11] 刘树新, 李星, 陈鸿昶, 等. 基于资源传输匹配度的复杂网络链路预测方法[J]. 通信学报, 2020, 41(6): 70-79.
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.

[12] 翟东升, 刘鹤, 张杰, 等. 一种基于链路预测的技术机会挖掘方法[J]. 情报学报, 2016, 35(10): 1090-1100.
ZHAI D S, LIU H, ZHANG J, et al. Approach to mining technology opportunity based on link prediction[J]. Journal of the China Society for Scientific and Technical Information, 2016, 35(10): 1090-1100.

[13] WU Z H, LIN Y F, WANG J, et al. Link prediction with node clustering coefficient[J]. Physica A: Statistical Mechanics and Its Applications, 2016, 452: 1-8.

[14] RAHIMI F, REZAEI H. An event-triggered recursive state estimation approach for time-varying nonlinear complex networks with quantization effects[J]. Neurocomputing, 2021, 426: 104-113.

[15] KUMAR A, SINGH S S, SINGH K, et al. Level-2 node clustering coefficient-based link prediction[J]. Applied Intelligence, 2019, 49(7): 2762-2779.

[16] 吴翼腾, 于洪涛, 黄瑞阳, 等. 采用组合方法进行链路预测的理论极限研究[J]. 通信学报, 2020, 41(6): 34-50.
WU Y T, YU H T, HUANG R Y, et al. Theoretical limit of link prediction using a combination method[J]. Journal on Communications, 2020, 41(6): 34-50.

[17] 李永立, 罗鹏, 张书瑞. 基于决策分析的社交网络链路预测方法[J]. 管理科学学报, 2017, 20(1): 64-74.
LI Y L, LUO P, ZHANG S R. Link prediction in social networks based on decision analysis[J]. Journal of Management Sciences in China, 2017, 20(1): 64-74.

[18] 孟绪颖, 张琦佳, 张瀚文, 等. 社交网络链路预测的个性化隐私保护方法[J]. 计算机研究与发展, 2019, 56(6): 1244-1251.
MENG X Y, ZHANG Q J, ZHANG H W, et al. Personalized privacy preserving link prediction in social networks[J]. Journal of Computer Research and Development, 2019, 56(6): 1244-1251.

[19] WANG Z Q, LIANG J Y, LI R. A fusion probability matrix factorization framework for link prediction[J]. Knowledge-Based Systems, 2018, 159: 72-85.

[20] 刘留, 王煜尧, 倪琦瑄, 等. 一种基于博弈论的时序网络链路预测方法[J]. 计算机研究与发展, 2019, 56(9): 1953-1964.
LIU L, WANG Y Y, NI Q X, et al. A link prediction approach in temporal networks based on game theory[J]. Journal of Computer Research and Development, 2019, 56(9): 1953-1964.

[21] 林原, 王凯巧, 刘海峰, 等. 网络表示学习在学者科研合作预测中的应用研究[J]. 情报学报, 2020, 39(4): 367-373.
LIN Y, WANG K Q, LIU H F, et al. Application of network representation learning in the prediction of scholar academic cooperation[J]. Journal of the China Society for Scientific and Technical Information, 2020, 39(4): 367-373.

[22] LI X, LIU S X, CHEN H C, et al. A potential information capacity index for link prediction of complex networks based on the cannikin law[J]. Entropy, 2019, 21(9): 863.

[作者简介]

顾秋阳（1995− ），男，浙江杭州人，浙江工业大学博士生，主要研究方向为智能信息处理、数据挖掘、中小企业高质量发展等。

吴宝（1979− ），男，浙江金华人，博士，浙江工业大学研究员、博士生导师，主要研究方向为复杂网络链路预测、金融信用风险控制与中小企业发展。

池仁勇（1959− ），男，浙江温州人，博士，浙江工业大学教授、博士生导师，主要研究方向为复杂网络链路预测、中小企业智能信息管理与创新创业。