

基于社团多特征融合嵌入表示的时序链路预测方法

朱宇航, 吉立新, 李英乐, 李海涛, 刘树新

(信息工程大学, 河南 郑州 450002)

摘要: 时序动态网络在静态网络基础上综合了时间属性的概念, 包含了网络结构的复杂性、动态性等内涵, 是研究复杂网络链路预测问题的较优思维对象, 因在现实世界中具有较高应用价值而备受关注。目前大部分传统方法研究对象仍局限于静态网络, 存在对网络时域演化信息利用不充分、时间复杂度较高等问题。结合社会学理论, 提出一种基于社团多特征融合嵌入表示的时序链路预测方法, 该方法的核心思想是通过分析网络动态演化特性, 在社团范围内学习节点的嵌入表示向量, 融合多特征以衡量节点间连边的生成概率。利用网络集体影响力的方法对节点和连边的权值进行计算, 基于集体影响的连边权值进行社团划分, 将网络划分为若干个社团子图, 得到基于集体影响的相似性指标。在社团范围内, 利用有偏的随机游走, 结合梯度优化的 Skip-gram 方法获取所有节点的嵌入表示向量, 得到基于社团范围游走的相似性指标。融合节点的集体影响、社团范围节点的多个中心性特征和学习到的节点表示向量, 得到多特征融合的相似性指标, 3 种新指标都可以用于衡量节点之间形成连边的概率。对比基于移动平均、嵌入表示、图神经网络等经典时序链路预测方法, 在 6 个真实数据集上的实验结果表明, 所提基于社团多特征融合的方法在 AUC 评价标准下取得更优的预测性能。

关键词: 时序链路预测; 社团特征; 影响力; 多特征; 随机游走; 嵌入表示

中图分类号: TP393

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2023013

Temporal link prediction method based on community multi-features fusion and embedded representation

ZHU Yuhang, JI Lixin, LI Yingle, LI Haitao, LIU Shuxin

Information Engineering University, Zhengzhou 450002, China

Abstract: Dynamic networks integrates time attributes on the basis of static networks, and it contains multiple connotations such as the complexity and dynamics of the network structure. It is a better thinking object for studying complex network link prediction problems in the real world. Its high application value has attracted much

收稿日期: 2022-01-03; 修回日期: 2022-04-20

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)

引用格式: 朱宇航, 吉立新, 李英乐, 等. 基于社团多特征融合嵌入表示的时序链路预测方法[J]. 网络与信息安全学报, 2023, 9(1): 67-82.

Citation Format: ZHU Y H, JI L X, LI Y L, et al. Temporal link prediction method based on community multi- features fusion and embedded representation[J]. Chinese Journal of Network and Information Security, 2023, 9(1): 67-82.

attention in recent years. However, most of the research objects of traditional methods are still limited to static networks, and there are problems such as insufficient utilization of network time-domain evolution information and high time complexity. Combining sociological theory, a novel temporal link prediction method was proposed based on community multi-feature fusion embedding representation. The core idea of this method was to analyze the dynamic evolution characteristics of the network, learn the embedded representation vector of nodes within the community, and effectively fuse multiple features to measure the generation probability of the connection between nodes. The network was divided into several subgraphs by using community detection with collective influence weights and the Similarity index was proposed based on the collective influence. Then, the biased random walk and the Skip-gram were used to get the embedded vectors for every node and the Similarity index was proposed based on the random walk within the community. Integrating the collective influence, multiple central features of the community, and the representation vector learned within the community, the Similarity index was proposed based on the multi-features fusion. Compared with classical temporal link prediction methods, including moving average methods, embedded representation methods, and graph neural network methods, experimental results on six real data sets show that the proposed methods based on the random walk within the community and the multi-features fusion both achieve better prediction performance under the evaluation criteria of AUC.

Keywords: temporal link prediction, community feature, influence, multi-feature, random walk, embedded representation

0 引言

信息作为事物及属性统一在思维中的反映,是个体认识客观世界、指导实践的主要素材,其价值在传递过程中得以体现。人类社会作为众多个体组成的复杂系统,在内部信息交互过程中表现出自组织特性与动态演化特性,即信息的交互导致个体相互的吸引与排斥。学校、工作单位、籍贯、民族等社会属性从不同视角把人们分类为规模各异的社团,并对每个人的人际交往持续产生影响,即属性相似的个体在网络演化过程中自组织为紧密社团的概率更高。基于这一社会学理论,从社团角度研究复杂系统相关问题一直是该领域研究热点。

链路预测技术已经应用于复杂网络的社区发现、异常检测、影响力分析和推荐系统等方面。链路预测旨在利用已观察到的网络拓扑结构、属性和网络时序演化等信息解决一个最基本的科学问题——缺失信息的还原与预测^[1-2],具体包括预测网络中缺失连边、错误连边和未来可能产生的连边^[3-4],包含网络重构和网络演化两个方面。链路预测的相关技术除了具有应用价值以外,还具有重要的理论研究价值,可以为网络演化机制提

供合理解释^[5],挖掘网络动态变化规律,为认识网络内在变化机制提供可靠的理论支撑。真实世界中的复杂网络往往是动态时序网络,网络中节点和连边都随着时间而在不停变化,时序链路预测的方法能更好地挖掘网络变化的历史信息,可以更加有效地实现预测的性能。

链路预测在大规模网络、多维异质网络和动态时序网络中有很多研究成果,通常可以将预测方法分为基于网络拓扑结构、基于似然分析、基于机器学习 3 大类^[6]。经典的基于网络结构相似性的指标包括统计共同邻居数量的 CN (common neighbors) 指标^[7],区分共同邻居对链路预测不同贡献度的 AA (adamic-adar) 指标^[8],使用路径长度为 3 的局部路径进行量化的 LP (local path) 指标^[9]、计算网络全部路径不同影响的 Katz 指标^[10]等;基于似然分析的方法,利用已知的网络结构和属性信息构建概率模型,用于量化评估未存在连边出现的概率,如 Zhao 等^[11]提出一种贝叶斯概率模型,该模型结合了各种有向和无向关系网络中的节点属性,刘树新等^[12]提出基于资源双向传输匹配度的相似性模型,Javaari 等^[13]对稀疏网络的局部和全局属性建立链路标签模型实现预测相关功能, Pan 等^[14]结合聚类机制,提出了

一种闭合路径的条件概率模型实现链路预测, 这一类方法能够较好地利用拓扑的结构信息, 但是方法的计算复杂度较高, 不适合较大规模的网络建模; 基于机器学习的方法是把网络的结构和属性信息输入各种神经网络进行训练, 输出节点或连边的嵌入向量表示, 进而实现分类和预测的功能, 如 Li 等^[15]利用受限的玻尔兹曼机进行优化的深度学习方法实现了动态网络链路预测, Chen 等^[16]利用长短期记忆网络对网络的时间信息进行嵌入学习, 同时利用图神经网络对网络的结构信息进行嵌入学习, 联合两个向量进行的实验大大提高了预测的准确率。基于机器学习的方法通常比基于结构相似性的方法预测性能更高, 比基于结构似然概率模型的方法时间复杂性更低, 但是在真实应用中, 往往需要较苛刻的条件, 对模型最优参数的训练需要消耗较大的资源。

不同类别的预测方法针对不同的网络场景, 基于拓扑结构的方法的计算复杂度较小, 只需要网络连接信息就可以实现, 应用场景广。随着社交媒体的普及, 越来越多的学者利用经典的社会学理论研究社交网络中的链路预测问题^[17,18]。刘书新等^[19]定义了三元组社团一致性指标描述, 三元组的社区属性对链路预测的影响, Jorge 等^[20]提出了基于用户兴趣行为和社团信息相结合的预测方法。Liu 等^[21]基于弱链接和共同邻居的 3 个节点中心性(度、紧密度和介数中心性), 提出一种链路预测方法。这类方法虽然简单利用社交信息提高了预测精度, 但是并没有区分不同社团规模对预测结果的作用, 还可以进一步挖掘不同社团结构对相似性的影响。

已有的社团理论和结构相似性方法在静态网络上实现了链路预测的功能, 但在动态网络中的相关研究较少, 本文针对时序网络的链路预测问题, 挖掘网络拓扑信息, 结合影响力最优化、社团发现、网络嵌入表示等方法进行研究, 进而实现一种基于社团多特征融合嵌入表示的时序链路预测方法, 具体贡献如下。

(1) 利用渗流理论中集体影响的概念, 得到节点最优影响力的量化方法, 并将集体影响力视为节点的有效权值。

(2) 利用 Louvain 算法对动态网络进行社区划分, 然后在社团范围内使用带偏随机游走和 Skip-gram 模型得到每个节点上下文环境的网络嵌入表示向量。

(3) 将集体影响力、网络中心性特征与节点的表示向量进行连接, 利用新联合的节点向量和社区关系权重计算所有连边的得分矩阵, 并实现时序链路预测。

(4) 在 6 个真实动态网络数据集中的实验结果表明, 相比传统经典的动态网络链路预测方法, 本文所提方法在 AUC 评价标准下取得更优的预测性能。

1 相关背景

1.1 时序网络链路预测问题描述

真实世界的复杂网络随着时间变化而动态变化, 无时无刻都有新连边的出现和旧连边的消失, 网络的结构也因此在不断演化更新。有效挖掘网络的动态演化信息对研究预测未来的拓扑结构具有重要意义, 也是链路预测任务的关键因素。

动态网络根据对时序信息分析粒度的不同, 一般可划分为离散时序网络和连续时序网络两种, 本文针对普适性更优的离散时序网络预测问题开展研究。离散时序网络通常根据时序对网络数据进行等间隔切分, 把动态网络分解为一系列的时间快照图集合, 这样的建模方式相当于对动态网络在时间维度上进行采样, 将动态网络拆分成多个静态网络序列。给定一个动态网络 G , 将其按固定时间间隔划分为 T 个网络快照序列 $\{G_1, G_2, \dots, G_T\}$, 其对应的邻接矩阵序列为 $\{A_1, A_2, \dots, A_T\}$ 。

如图 1 所示, 网络 G 表示动态网络所有节点和连边的集合, $G_t = G(V_t, E_t)$ 表示时刻 t 的网络快照, V_t 和 E_t 分别为时刻 t 的节点集合和连边集合, 每一个节点可以包含多个属性信息, 每一个连边可以包含多个权重信息, 节点和连边在时序网络存在动态增加和消失的情况。利用 T 时刻之前多个快照的拓扑和属性信息作为训练集合进行建模分析, 预测 T 时刻网络的拓扑连接情况, 就是时序链路预测问题。

1.2 基于社团划分的链路预测方法

在社交网络中, 用户常被抽象为节点形式表

示, 用户间的关系常被抽象为连边, 二者共同构成整个网络结构。在这样的网络中, 有的用户之间的连接较为紧密, 有的用户之间的连接关系较为稀疏, 社团分类示意如图2所示, 连接较为紧密的部分可以被看成一个社区, 其内部的节点之间有较为紧密的连接, 而在两个社区间则相对连接较为稀疏, 这就是社团结构。

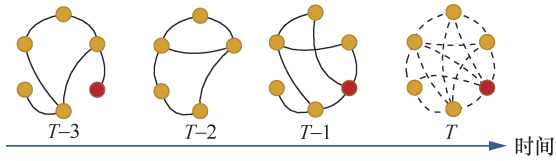


图1 时序链路预测示意
Figure 1 Schematic of temporal link prediction

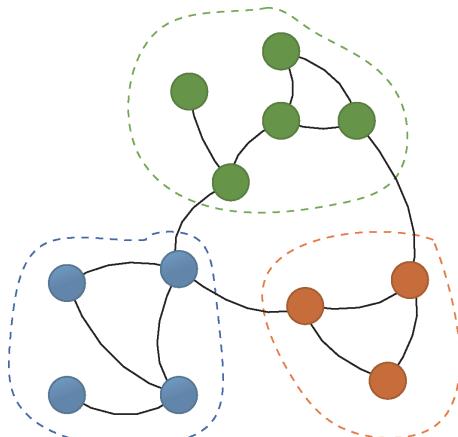


图2 社团分类示意
Figure 2 Schematic of community classification

基于社团划分的链路预测最简单的思路是将节点与社区的关系属性作为相似性增益, 文献[22]中定义了式(1), 用于计算社团相似度的增益, 节点 i 和节点 j 存在连接关系, 如果两个节点都在相同的社团, 则节点对之间的增益为 ρ , 并且 $\rho > 1$, 如果两个节点在不同的社团, 则节点对的增益为1, 表示社团对连边相似性没有增益, 如果两个节点之间不存在连接关系, 则社团增益为0。

$$S(i, j) = \begin{cases} \rho, & i, j \in C_t^k, k=1, 2, \dots \\ 1, & i \in C_t^k, j \in C_t^{k'}, k, k'=1, 2, \dots \\ 0, & \text{其他} \end{cases} \quad (1)$$

将社区结构信息与局部相似性信息进行融合的链路预测研究有很多, 如 Valverde-Rebaza 等[23]量化了社团内共同邻居的集合和社团之间共同邻

居的不同作用场景; Wang 等[24]提出将资源分配 RA 指标和来自社区内共同邻居的信息进行融合实现链路预测的功能; 焦杨[25]提出将社区指标分别与局部相似性的 PA 指标和 RWR 指标进行融合; 该类方法对社团的拓扑信息挖掘越充分, 对网络拓扑预测的提升精度就越大。

1.3 基于时序网络模型的预测方法

时序链路预测是在静态方法的基础上, 叠加时序网络模型实现预测相关功能。时序网络模型通常可以分为基于移动平均的模型、基于自回归的模型、基于指数平滑的模型, 以及基于深度神经网络学习的模型。这些模型通过对历史网络数据的学习, 实现对未来网络拓扑的预测。

基于移动平均的模型利用移动范围内快照集合的均值进行相似性度量[26], 移动的范围可以是全部历史快照、固定长度的移动窗口或者仅仅是预测前一个时刻的快照。基于自回归的模型包括线性回归、逻辑回归、最小二乘回归、岭回归等模型[3], 回归模型的本质是利用前馈网络和反向传播方法对多项式的系数进行训练学习。基于指数平滑的模型包括单指数平滑模型、双指数平滑模型, 以及三指数平滑模型等。单指数平滑模型实现相似性数据的加权处理, 数据的权重随时间变化呈指数形式递减; 双指数平滑模型在单指数平滑模型的基础上, 考虑了时间序列的变化趋势, 即同时考虑历史平均和序列变化趋势, 如 ARIMA (autoregressive integrated moving average model) 模型[27]在考虑历史数据和趋势数据的同时, 利用差分机制对非平稳序列进行转换; 三指数平滑模型在双指数平滑模型基础上增加了周期性因素, 适用于处理包含周期性的时间序列数据, 如 Holter-winter 模型[28]和 SARIMA (seasonal autoregressive integrated moving average) 模型[29]。基于图神经网络 (GNN, graph neural network) 和循环神经网络 (RNN, recurrent neural network) 结合的模型能够较好地利用动态网络的空间和时间信息[30], GNN 通过端到端的方式自动地学习节点特征表示, 可以捕捉节点间的非线性关系, RNN 方法则负责处理网络演化的时序信息, 能够更有效地聚合不同时刻的节点嵌入表示, 从而大幅提升时序预测的准确率。

1.4 评价指标

常用于衡量链路预测算法精确度的评价标准主要包括 AUC (area under curve)、正确率 (accuracy)、精度 (precision)、召回率 (recall) 和 $F1$ 分数^[31]。定义 AUC、正确率、精度和召回率的公式如下。

$$AUC = \frac{n' + 0.5n''}{n} \quad (2)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

其中, AUC 标准从整体上评估链路预测算法的性能, 评价原理为在测试集中随机选取 n 次连边进行独立比较, 相等取 1, 不相等取 0.5, 得到最终的平均得分为 AUC 精度值。TP 表示存在连边并且预测正确的样本数, FN 表示存在连边并且预测错误的样本数, FP 表示不存在连边关系并且预测错误的样本数, TN 表示不存在连边关系并

且预测正确的样本数。Accuracy 反映了正确预测的样本数占总预测样本数的比重; Precision 反映了预测存在连边的正样本占有所有预测为存在连边关系的比重; Recall 反映了被正确判定存在的样本占有所有真实存在连边的比重。

$F1$ 分数是基于精度和召回率的调和平均, $F1$ 分数的定义公式如下。

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (6)$$

AUC、 $F1$ 的值越大, 表示预测的效果越好。

2 算法实现

本文所提算法可划分为如下步骤。首先, 利用集体影响算法量化节点的权值, 结合 Louvain 算法按照模块度最优化的原理进行社团划分; 其次, 在社团的上下文范围内, 对每一个节点进行有偏的随机游走, 使用 Skip-gram 模型进行训练, 得到节点的嵌入表示向量; 最后, 使用节点的集体影响力、度中心性、介数中心性等特征与训练得到的节点表示向量进行连接, 得到所有节点对的得分矩阵, 进而实现时序链路预测功能。具体流程如图 3 所示。

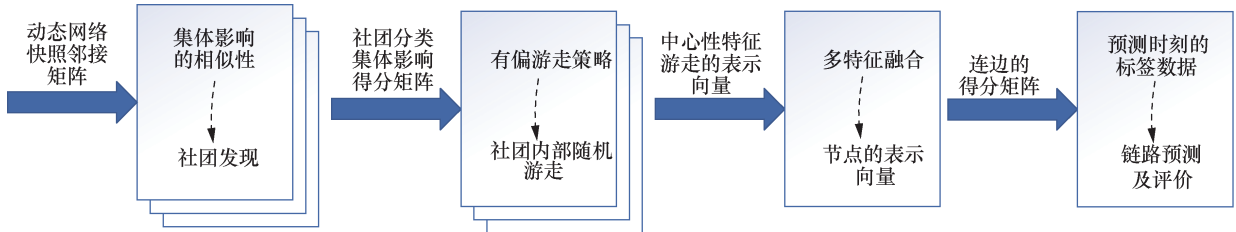


图 3 基于社团多特征融合嵌入表示的时序链路预测流程
Figure 3 The multi-features fusion and embedded representation for temporal link prediction

2.1 基于集体影响的社团划分

渗流模型一般用于解决复杂网络中连通性的临界问题。渗流模型最优化就是找出能够摧毁网络最大连通分量的最小节点集, 即在网络全局连接中发挥至关重要作用的一组节点。为解决该问题, Morone 等^[32]提出了集体影响 (CI, collective influence) 算法用于量化节点的影响力, 一个节点的集体影响力使用球形边界上的节点进行刻画, 而与球内部路径上的其他节点没有关系, 使用集体影响力可以更有效地量化节点在局部范围内的拓扑信息。如图 4 所示, 节点 i 在半径 3 的范围集体影响力可以使用 j_1 到 j_9 这 9 个边界节点和节点 i 共同作用进行刻画。

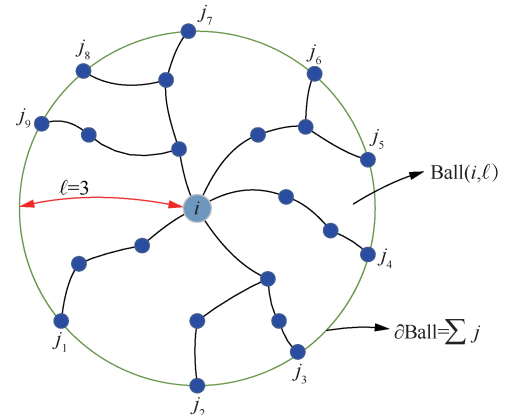


图 4 节点的集体影响力示意
Figure 4 Diagram of the collective influence of nodes

具体定义如式(7), ℓ 为设置的球半径(球上节点到中心节点的路径), $\text{Ball}(i, \ell)$ 为以节点 i 为中心、以 ℓ 为半径的球内的节点集合, $\partial\text{Ball} = \sum_j$ 表示该球的边界, 则节点 i 从 ℓ 层节点所得到的集体影响力强度, 即 CI 值为

$$\text{CI}_\ell(i) = (k_i - 1) \sum_{j \in \partial\text{Ball}(i, \ell)} (k_j - 1) \quad (7)$$

其中, k_i 表示节点 i 的度, ℓ 取值为非负整数且不超过有限网络的网络直径。使用 CI 算法能更有效地利用局部拓扑信息, 本文设置 ℓ 取值为 3。

动态网络可以按照固定时间间隔划分为多个快照集合, 如图 5 所示, 对某一时刻的网络拓扑进行预测时, 需要 p 个时间窗口作为训练集。为了充分利用网络时间演化因素, 利用指数函数拟合网络的时序演化作用, 得到训练集每个节点的集体影响力权值, 如式(8)所示。

$$\text{CI_train}(i) = \sum_{t=T-p}^T \alpha^{t-T} \cdot \text{CI}(i, t) \quad (8)$$

其中, α 表示时间衰减参数, 取大于 1 的数值, 值越大, 离预测时刻相对较远的快照起的影响越小, $\text{CI}(i, t)$ 表示节点 i 在 t 时刻的集体影响力强度。

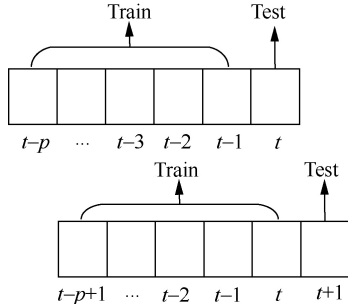


图 5 动态网络数据集划分示意

Figure 5 Schematic diagram of dynamic network data set division

在得到所有节点的集体影响力后, 基于社交网络连接偏好的特性, 定义节点对 (i, j) 集体影响力的相似性指标 S^{CI} 为式(9), 因为 CI 一般是大于 1 的数值, 分子为两个节点的 CI 之积, 用于刻画两侧节点对局部拓扑结构的集聚影响效果, 分母为两个 CI 之和, 用于量化两侧节点对周边连边的平均影响力。如果两个节点间只有这一条连边, 则定义该节点对的集体影响力为 1。

$$S^{\text{CI}}(i, j) = \begin{cases} \frac{\text{CI_train}(i) \cdot \text{CI_train}(j)}{\text{CI_train}(i) + \text{CI_train}(j)}, & k_i > 1 \text{ 且 } k_j > 1 \\ 1, & k_i = 1 \text{ 且 } k_j = 1 \end{cases} \quad (9)$$

对于带权子图的构建, 较弱的关联性往往会影响节点间关联程度的表达, 为保证对大规模网络进行社团划分时不产生特别不合理的大社团, 将连边的 CI 权重矩阵中最小的 5% 的值替换为 0, 如式(10)所示。

$$\text{CI}(i, j) = \begin{cases} S^{\text{CI}}(i, j), & S^{\text{CI}}(i, j) > \text{rank}(5\%) \\ 0, & \text{其他} \end{cases} \quad (10)$$

为确保边的权重在网络上实际存在, 将 CI 权重矩阵与邻接矩阵对应相乘, 得到所有连边集体影响力的权值矩阵。此时, 邻居矩阵 A 为训练集时间步长 p 范围内的累积无权矩阵。

$$M = \text{CI}(i, j) \cdot A(i, j)$$

$$A(i, j) = \sum_t A_t(i, j), \quad a_{ij} = 1 \text{ 且 } a_{ij} > 0 \quad (11)$$

进一步, 利用 Louvain 社团检测算法^[33]把网络划分为若干个子网, 输入训练集的网络拓扑结构信息和连边权值矩阵, 按照模块度最大化的方法迭代计算, 模块度的计算方法如下。

$$Q = \frac{1}{2m} \sum_{i,j,i \neq j} (M_{ij} - \frac{k_i k_j}{2m}) \quad (12)$$

其中, M_{ij} 表示节点 i 和节点 j 连接边的实际权重, $k_i = \sum_j M_{ij}$ 表示所有和节点 i 连接的边的权重之和, $\frac{k_j}{2m}$ 表示节点和整个图上任意一个节点连接的概率, $k_i \cdot \frac{k_j}{2m}$ 表示节点 i 和节点 j 连接边的期望权重。

值, 两者之差代表最终的增益, $m = \frac{1}{2} \sum_{i,j} M_{ij}$ 表示网络所有边的权值之和。

计算集体影响力权值矩阵和进行社团划分的伪代码流程如算法 1 所示。

算法 1 CreateComWithCI(g, T, α) /*按集体影响构造社团子图*/

输入 时间步长 τ ;

时间衰减参数 α ;

网络快照 $g_t = \{G^t, G^{t-1}, \dots, G^{t-\tau+1}\}$

```

1) for nodei in G' do /*选取当前G'中的节点 nodei*/
2)   for j=1 to τ do /* 处理τ 内的G'*/
3)     按式(7)建立所有 nodei 集体影响权值
4)   end for
5)   按式(8)构造训练集迭代集体影响值
6) end for
7) 按式(9)构造连边的集体影响力权值矩阵
8) 按式(10)、式(11)去掉弱连接的连边权值矩阵
9) G* = Louvain(G', edge_weight) 得到各个社团子图

```

输出 G*, CI^v, CI^e /*返回社团子图列表 G*, 节点和连边的集体影响力权值矩阵*/

$$P(j|i) = \begin{cases} \frac{CI(i,j)}{\sum_z CI(i,z)} & , i, j \in C^k, k=1, 2, \dots \\ 0, & \text{其他} \end{cases} \quad (13)$$

其中, z 表示节点 i 所在社团的全部邻接节点, C^k 表示第 k 个社团。网络中两个节点同属于一个社团, 并且它们之间的关联程度越大, 进行有偏的随机游走的概率越大。

2.2 基于子图范围游走的嵌入表示学习

受 word2vec^[34]和 node2vec^[35]的启示, 词嵌入表示学习的基本元素是单词, 对应网络的节点表示的元素是节点; 词嵌入是对构成一个句子中单词序列进行分析, 因此网络嵌入表示中节点构成的序列就是随机游走。node2vec 方法通过设置参数 p 和 q 来控制游走的方向、广度和深度, 本文的改进思路是把节点游走的上下文范围限定在节点所属的社团子图内部, 同时在节点选择下一跳时, 根据式(9)得到的所有连边集体影响力进行概率判断, 使得权值大的连边在随机游走的过程中优先选择, 式(13)表示从节点 i 游走到邻居节点 j 的概率。在节点 i 到节点 j 进行随机游走时, 当存在共同邻居时, 优先选择游走到共同邻居节点, 如不存在共同邻居时, 再按照参数 p 和 q 进行游走控制。

在得到每个节点的随机游走序列后, 可以将序列信息输入经典的 Skip-gram 模型学习得到节点的表示向量。Skip-gram 模型是用于最大化出现

在窗口内的单词间共现概率的语言模型, 可在已知当前节点的情况下预测其“上下文”节点, 即输入一个节点的游走信息, 输出多个节点的表示向量。本节利用 Skip-gram 模型结合划分好的各个社团子图, 设置以下目标函数来增大节点共同出现的概率:

$$\zeta = \max_{\theta} \sum_{i \in V} \log P(\Gamma(i)|i; \theta, C) \quad (14)$$

其中, i 表示输入的节点信息, $\Gamma(i)$ 表示输出的邻居节点信息, C 表示社团信息, θ 表示模型参数, θ 由两个矩阵 u 和 v 构成, u 为上下文矩阵, v 为节点特征矩阵, 通过迭代训练最终得到最优的 u 和 v 。式(14)中, $P(\Gamma(i)|i; \theta, C)$ 可以表示为

$$P(\Gamma(i)|i; \theta, C) = \prod_{j \in \Gamma(i)} p(j|i; \theta, C) \quad (15)$$

其中, $p(j|i; \theta, C)$ 表示邻居节点 z 在游走的序列中存在的概率。最后, 利用 softmax 函数对输出结果进行优化, 使

$$p(j|i; \theta, C) = \frac{e^{u_j \cdot v_i}}{\sum_{z \in V} e^{u_z \cdot v_i}} \quad (16)$$

其中, u_j 表示矩阵 u 第 j 行的向量, v_i 则表示矩阵 v 的第 i 行向量, 分别可以看作邻居节点 j 的上下文向量和种子节点 i 的表示向量, z 表示除了当前上下文节点之外的其他社区节点, 对这两个向量的内积做 softmax 运算可以得到式(15)的条件概率结果。

为了提高算法的计算效率, 采用负采样法对模型进行优化。节点被选作负样本的概率可根据节点的度分布得到。由文献[29]可知, 节点被选作负样本的概率可表示为

$$P(i) = \frac{f(i)^{\frac{3}{4}}}{\sum_{j=1}^K f(j)^{\frac{3}{4}}} \quad (17)$$

其中, K 是负样本数, $f(i)$ 表示节点 i 的集体影响力权值。因此, 基于改进的 Skip-gram 模型和负采样法, 算法的目标函数可以更新为

$$O(X) = \log \delta(u_j \cdot v_i) - \sum_{k=1}^{\text{neg}} \delta(u_k \cdot v_i) \quad (18)$$

其中, $\delta(x) = 1 / (1 + \exp(-x))$ 是 sigmoid 函数, u_j 表示社团内采样的邻居节点向量信息, u_k 表示社团内采样的负样本信息。

两个节点向量之间越相似, 点乘的结果越大, 从而归一化之后得到的概率值越大, 可以利用点乘归一化后的概率值表示连边的相似性程度, 基于社团有偏嵌入表示的相似性指标 S^{Com} 定义如式(19), X^{emb} 表示经过社团划分、有偏游走和 Skip-gram 学习后的所有节点嵌入表示向量。

$$S^{Com} = \begin{cases} X^{emb} \cdot (X^{emb})^T, & i \neq j \\ 0, & \text{其他} \end{cases} \quad (19)$$

计算基于社团有偏游走的嵌入表示学习过程伪代码流程如算法 2 所示。

算法 2 ComRWSkipgram($g, w, n, d, l, CI^v, CI^e, G^*$)

输入 时间步长 τ ; 窗口大小 p ; 游走长度 l ; 节点数量 n ; 游走次数 w ; 向量维度 d ; 网络快照 $g_t = \{G^t, G^{t-1}, \dots, G^{t-\tau+1}\}$

输出: 节点的向量表示矩阵 $X \in \mathbf{R}^{V \times d}$; 基于社团有偏嵌入表示的相似性指标 S^{Com}

```

1) 初始化向量矩阵  $X$ 
2)  $G^*, CI^v, CI^e = \text{CreateComWithCI}(g, T, \alpha)$ 
/*计算集体影响和社团划分*/
3) for com to  $G^*$  do /*遍历所有社团*/
4)   for  $i=1$  to  $w$  do
5)     for  $v_i \in \text{com}$  do /*遍历社团中的
所有节点*/
6)       按式(13)计算节点游走的概率
转移矩阵  $P$ 
7)       walks=BiasedRandomwalk(com,
 $v_i, l, P$ ) /*在 com 中进行带偏置的随机游走获取节
点序列 walks*/
8)        $X^{emb} = \text{Skipgram}(X, p, \text{walks})$  /*
将节点序列 walks 输入 Skip-gram 模型学习节点的
表示向量  $X^*$ */
9)     end for
10)   end for
11) end for
12) 按式(19)计算得到  $S^{Com}$ 
```

2.3 基于社团多特征融合表示的时序链路预测

节点的中心性指标体现了节点在社交网络中的重要性。文献[22]利用了网络的结构信息, 提出了融合社团和中心性指标的相似性指标, 文献[36]中提出了多个局部相似性指标与权值相融合的子图相似性特征序列, 本文考虑将式(8)得到的节点集体影响力权值、节点的子图中心性特征, 与式(19)中的 X^{emb} 表示向量融合得到新的节点表示向量 X^{new_emb} , 之后利用余弦相似性计算得到多特征向量融合的相似性指标 S^{MFF} 。

只在节点所属的社团范围内对节点的度中心性特征进行计算, 如式(20)所示, α 表示指数衰减参数, k_i 表示与节点 i 的度数, 即与之相连的边数之和, n 表示网络所有的节点数量, C 表示节点 i 所属的社团。

$$DC(i) = \sum_{t=T-p}^T \alpha^{t-T} \cdot \frac{k_i}{n-1} \quad | \quad k_i = \Gamma(i), i \in C \quad (20)$$

节点的介数中心性特征计算如式(21), α 表示指数衰减参数, $\sigma_{st}(i)$ 表示经过节点 i 的 $s \rightarrow t$ 的最短路径条数, σ_{st} 表示 $s \rightarrow t$ 的最短路径条数, C 表示节点 s, i, t 所属的社团。

$$BC(i) = \sum_{t=T-p}^T \alpha^{t-T} \cdot \sum_{s \neq t, i \in V} \frac{\sigma_{st}(i)}{\sigma_{st}} \quad | \quad s, i, t \in C \quad (21)$$

节点的接近中心性特征计算如式(22)所示, α 表示指数衰减参数, n 表示网络所有的节点数量, $d(i, j)$ 是节点 i 和节点 j 之间的平均距离, C 表示节点 i, j 所属的社团。

$$CC(i) = \sum_{t=T-p}^T \alpha^{t-T} \cdot \frac{n-1}{\sum_{j=1}^{n-1} d(i, j)} \quad | \quad i, j \in C \quad (22)$$

多特征融合后得到新的节点表示向量, 如式(23)所示, 特征直接相互连接形成新的向量。

$$X^{new_emb} = [DC : BC : CC : CI, X^{emb}] \quad (23)$$

余弦相似度通过计算两个向量之间夹角的余弦值来衡量两个向量的相似度, 对于 d 维向量 A 和 B 。

$$\text{cosine}(A, B) = \frac{\sum_1^n (A_i \cdot B_i)}{\sqrt{\sum_1^n A_i^2} \cdot \sqrt{\sum_1^n B_i^2}} \quad (24)$$

$$S^{MFF} = \text{cosine}(X^{new_emb}, X^{new_emb})$$

通过式(24)得到多特征融合的相似性指标 S^{MFF} 。余弦相似度的范围在 $[-1,1]$ ，值越大表示两个向量的夹角越小，相似度越高；值越小表示两个向量的夹角越大，相似度越低。

基于社团嵌入表示的时序链路方法的处理数据流如图 6 所示，首先，黄色线的数据流实现了节点和连边的集体影响力权值计算，并在此基础上进行了社团划分，得到式(9)的相似性指标 S^{CI} ；然后，蓝色线的数据流实现了有偏的随机游走，并结合社团子图范围进行了 Skip-gram 训练，得到节点的嵌入表示向量 X^{emb} 和式(19)的相似性指标 S^{Com} ；最后，绿色线的数据流实现了节点的 3 个中心性指标计算，并融合 CI 特征和 X^{emb} 形成新的表示向量 X^{new_emb} ，利用余弦相似度计算得到相似性指标 S^{MFF} 。

基于多特征融合嵌入表示的时序链路预测方

法的伪代码流程如算法 3 所示。

算法 3 MultiFeaturesFusion($g, T, \dots$)

输入 动态网络 $G = \{G^1, G^2 \dots G^T\}$ ；时间窗口 n ；向量维度 d ；负样本数 N ；时间步长 τ ；随机游走次数 w ；随机游走长度 l

输出 多特征融合的相似性指标 S^{CCC}

- 1) 初始化向量矩阵 X^{new_emb}
- 2) $G^*, CI^v, CI^e = \text{CreateComWithCI}(g, T, \text{node})$
- 3) $X^{emb} = \text{ComRWSkipgram}(g, w, n, d, l, CI^v, CI^e, G^*)$
- 4) 按式(20)~式(22)得 DC、BC、CC
- 5) 按式(23)融合多特征 $X^{new_emb} = [DC:BC:CC:CI, X^{emb}]$
- 6) 按式(24)计算所有连边的余弦相似度作为最终预测的得分矩阵

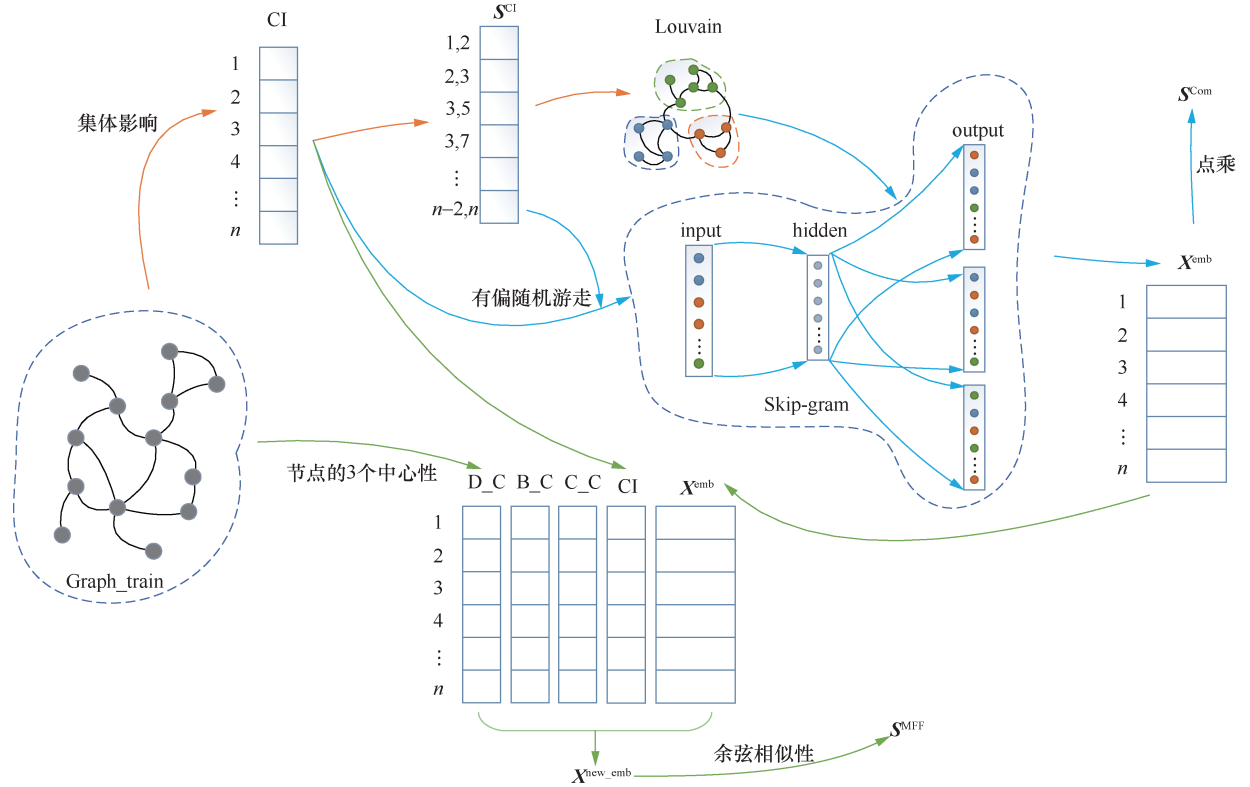


图 6 基于社团嵌入表示的时序链路方法的处理数据流

Figure 6 Logic flow chart of temporal link prediction method based on community multi-features fusion

3 仿真实验

3.1 实验数据集

本文采用 6 个通信网络数据集（数据集基本

特征参数如表 1 所示）对算法性能进行评估。E-mail 数据集^[37]是某大型欧洲研究机构的电子邮件数据生成的；Enron 数据集^[38]是由 Enron 公司雇员之间发送的电子邮件数据组成的；DNC 数据

表 1 数据集基本特征参数
Table 1 Basic characteristic parameters of datasets

基本特征参数	E-mail	Enron	DNC	MAN	UCI	LEM
节点数	1 005	87 273	2 029	167	1 899	485
连边数	332 334	1 048 576	39 264	82 927	59 835	196 364
起始日期	2003-10	2001-1-1	2016-4-23	2010-01	2004-4-15	1979-04
结束日期	2005-5	2002-3-31	2016-5-25	2010-10	2004-10-26	2004-06
总时长	526 天	454 天	33 天	268 天	195 天	303 月
时序周期	周	周	天	周	周	半年
快照图数	76	66	33	39	28	51

集^[39]是美国民主党全国委员会邮件泄露事件中采集的邮件往来网络；MAN 数据集^[39]是某制造业工厂内部员工的邮件记录；UCI 数据集^[40]由加州大学欧文分校的学生之间传递的短信构成的在线社交网络；LEM 数据集^[41]由堪萨斯事件数据系统基于包含 WEIS 编码事件的文件收集的交互网络，涵盖了 1979 年 4 月至 2004 年 6 月内的事件。

3.2 实验对比算法

本文对比的算法包括基于移动平均的方法^[26]、基于节点嵌入表示的方法（node2vec）^[35]、基于一阶和二阶相似度信息进行嵌入表示的方法（LINE）^[42]、基于图注意力的方法（DySAT）^[43]、基于动态图卷积的方法（EvolveGCN）^[44]。具体介绍如下。

（1）基于移动平均的方法

该类方法利用移动范围内相关快照的相似性均值进行量化分析，如式(25)所示，记作 Average；如果移动范围仅为预测之前的 1 个窗口，则模型可作为最近时刻算法，如式(26)所示，记作 Last；如果把移动范围 m 个时刻整体视为一个加权静态窗口，则模型可作为迭代周期的最近邻算法，如式(27)所示，记作 Reduce。

$$\hat{S}^{\text{Average}}(T) = \frac{1}{m} \sum_{t=T-m+1}^{T-1} f_2(A(t)) \quad (25)$$

$$\hat{S}^{\text{Last}}(T) = f_2(A(T-1)) \quad (26)$$

$$\hat{S}^{\text{Reduce}}(T) = f_2\left(\sum_{t=T-p+1}^{T-1} A(t)\right) \quad (27)$$

（2）基于节点嵌入表示的方法（node2vec）

该方法借鉴词嵌入的思路，输入网络拓扑结构，输出每个节点的表示向量；式(28)为该方法的目标函数， $f(u)$ 为节点 u 的映射函数， $N_s(u)$ 为

通过采样策略 S 采样出的节点 u 的特征节点集合。本文实验设置游走参数 $p=1$ ， $q=2$ ，优先选择广度游走策略。

$$\max \sum_{u \in V} [-\log Z_u + \sum_{n_i \in N_s(u)} f(n_i) \cdot f(u)] \quad (28)$$

（3）基于一阶和二阶相似度信息进行嵌入表示的方法（LINE）

该方法利用节点之间直接联系的一阶信息和共同邻居的二阶信息共同设计目标函数，如式(29)所示， O_1 和 O_2 为一阶和二阶的节点 i 的目标函数， w_{ij} 为连边关系的权重。本文实验采用二阶相似度方式进行计算。

$$\begin{aligned} O_1 &= - \sum_{(i,j) \in E} w_{ij} \log p_1(v_i, v_j) \\ O_2 &= - \sum_{(i,j) \in E} w_{ij} \log p_2(v_j | v_i) \end{aligned} \quad (29)$$

（4）基于图注意力的方法（DySAT）

该方法结合网络的结构自注意力层特征和时间自注意力层特征，并使用多头注意力机制捕捉网络快照之间的演化特性，利用得到的图嵌入表示向量实现链路预测。式(30)为预测模型的损失函数。

$$\begin{aligned} L &= \sum_{t=1}^T \sum_{v \in V'} \left(\sum_{u \in N_{\text{walk}}^t(v)} -\log(\sigma(\langle e_u^t, e_v^t \rangle)) - w_n \cdot \right. \\ &\quad \left. \sum_{u' \in P_n^t(v)} \log(1 - \sigma(\langle e_u^t, e_{v'}^t \rangle)) \right) \end{aligned} \quad (30)$$

（5）基于动态图卷积的方法（EvolveGCN）

该方法使用门控循环单元（GRU，gated recurrent unit）或者长短期记忆网络（LSTM，long short-term memory network）在每个时间步长调节图卷积网络（GCN，graph convolutional network）参数，实现对图序列的动态性特征的捕获，进而

实现预测功能。本文使用 LSTM 方式实现对网络参数的动态调整。

$$\begin{aligned} H_t^{(l+1)} &= \text{GCONV}(A_t, H_t^{(l)}, W_t^{(l)}) \\ W_t^{(l)} &= \text{GRU/LSTM}(W_{t-1}^{(l)}) \end{aligned} \quad (31)$$

3.3 实验结果

实验 1 AUC 精度对比

在实验中，设置集体影响力的球半径 $\ell = 3$ ，即考虑半径为 3 的球半径上的点作为边界进行计算；设置训练集的时间窗口 $T = 7$ ，即利用预测时刻之前的 7 个时间步长的网络数据进行训练，设置集体影响力的时序衰减系数 $\alpha = 0.9$ ，利用式(9)的相似性指标进行链路预测，记作 S_ci；设置节点的向量维度为 128，负样本数为 5，应用到随机游走和 Skip-gram 模型的算法中，设置随机游走长度为 80，游走次数为 10 次，窗口大小为 10，利用式(19)的相似性得分进行链路预测，记作 S_com；拼接节点的集体影响、3 个中心性影响和游走的嵌入向量，利用式(24)的相似性分数进行链路预测，记作 S_mff。AUC 精度对比结果如表 2 所示，S_mff 方法在所有数据集中均取得了最好的预测精度值。

首先，通过将本文所提方法与 3 种基于共同邻居指标 CN 进行移动平均时序处理的方法 (Last、Average 和 Reduce) 进行对比可看出，Last 方法的预测性能最差，Average 方法的效果比 Last 好，比 Reduce 方法差，说明使用更多拓扑信息可提高预测的 AUC 精度；而 S_ci 在除 LEM 数据集的其余 5 个数据集的性能均优于 Last，但低于 Average 和 Reduce 方法，说明依赖边界节点影响对刻画中心节点的拓扑变化的效果一般。S_com 和 S_mff 相比 3 种方法的效果都有较大的提高，较 Last 平均提高约 30%，较 Average 平均提高约 15%，较 Reduce 平均提高约 12%。

然后，与基于网络表示学习的 node2vec 和 LINE 方法进行对比，由于这两种方法不是端到端的学习方法，节点向量的学习过程与随机游走的路径有很大关系，因此许多情况下无法精确捕捉到局部拓扑结构中有用的信息，预测的 AUC 精度会上下浮动 1% 左右。S_ci 在 Enron 和 DNC 数据集中表现较差，S_com 和 S_mff 的效果相对提高 2%~8%；在其他 4 个数据集中，S_com 和 S_mff

的精度比 node2vec 平均提高 28% 左右，比 LINE 平均提高约 21%。

表 2 AUC 精度对比结果
Table 2 Results of AUC accuracy

方法	Email	Enron	DNC	MAN	UCI	LEM
Last	0.753 5	0.605 6	0.701 1	0.841 5	0.515 2	0.902 1
Average	0.879 3	0.728 2	0.815 6	0.896 4	0.572 4	0.952 0
Reduce	0.895 4	0.773 2	0.825 2	0.902 5	0.615 7	0.953 0
node2vec	0.838 4	0.887 7	0.842 0	0.602 8	0.522 7	0.744 4
LINE	0.825 3	0.878 4	0.862 4	0.606 8	0.650 7	0.771 3
DySAT	0.860 3	0.788 5	0.837 6	0.902 0	0.718 7	0.835 2
EvolveGCN	0.750 3	0.634 7	0.878 9	0.801 6	0.681 0	0.896 9
S_ci	0.852 4	0.693 8	0.780 9	0.906 1	0.604 0	0.842 9
S_com	0.950 6	0.899 3	0.911 7	0.901 9	0.846 8	0.951 4
S_mff	0.951 8	0.907 7	0.912 0	0.909 2	0.847 1	0.956 3

最后，与基于图神经网络的 DySAT 和 EvolveGCN 方法进行对比，其精度均优于 S_ci，但预测性能不及 S_com 和 S_mff。该结果说明基于图神经网络方法虽利用更丰富数据维度进行训练，且建立更多隐藏层网络表示拓扑，但实际预测中效果提升有限。对比 DySAT 方法，S_mff 的精度在 MAN 数据集有较小的提升，在其他 5 个数据集精度提高 8%~17%；对比 EvolveGCN 方法，S_mff 的精度平均提高 19% 左右。

实验 2 球半径参数敏感性测试

如图 7 所示，设置集体影响方法的球半径为 1 到 6，观察球半径对 3 种指标的影响变化情况；文献[27]在连通性最优化的实验中，推荐设置半径大小为 3，但是在链路预测的应用中集体影响的效果一般，如图 7 (a) 所示，S_ci 在半径为 1 时效果最好，说明利用最近邻居的拓扑信息对结构的时序预测效果最好，在半径大于 2 时，对 AUC 精度的影响趋向相同；如图 7 (b)、图 7 (c) 所示，球半径长度对这两种指标的影响较小，精度的波动幅度在 1% 左右。总体上，集体影响的半径较大时，对节点局部拓扑信息的刻画具有一定的局限性，仅利用集体影响的信息对链路预测的提升效果有限，根据实验效果，推荐设置球半径参数为 1。

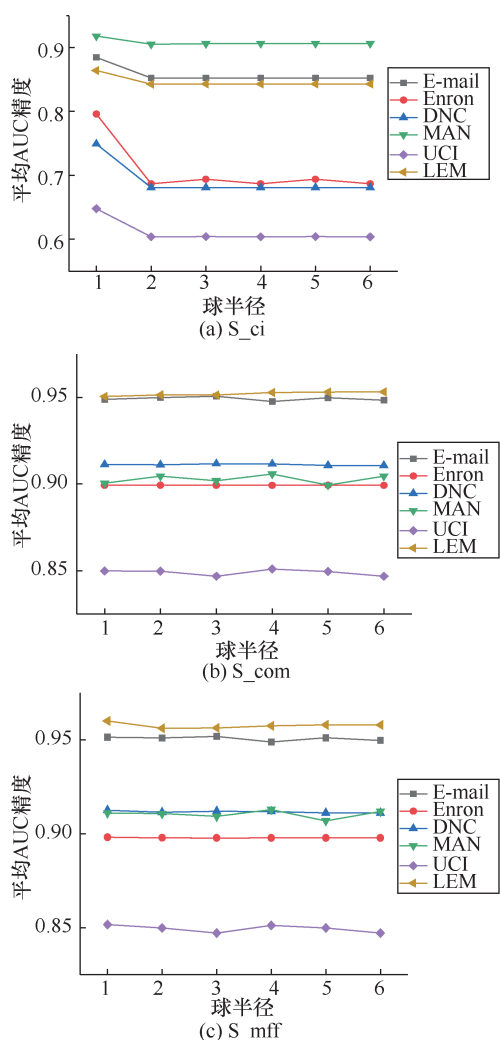


图7 集体影响方法的球半径对比结果
Figure 7 Results of different collective influence radius

实验3 随机游走参数敏感性测试

对 Skip-gram 的游走参数进行敏感性测试, 通过分别改变向量维度大小、随机游走步长、节点游走次数和游走窗口大小等参数, 观察 AUC 分数值的变化情况。

如图8所示, 各个参数的实验结果表示对 S_{mff} 相似性指标中随机游走参数进行不同的设置后, 利用节点表示向量实现链路预测的得分变化情况。图8(a)展示了向量维度参数的测试情况, 随着维度参数的增大, AUC 分数值在小范围内波动变化, 分别在维度为32、64或128时取得最优值, 但维度参数的变化对 AUC 分数值的影响较小, 得分的波动范围在1%左右; 图8(b)展示了游走步长参数的测试情况, 随着该参数增长 AUC 分数值也在小范围内逐渐增长, 在 E-mail

和 MAN 数据集上影响较大, 步长80比步长20增长了3%左右, 其他4个数据集变化不超过1%; 图8(c)、8(d)展示了随机游走次数参数和节点游走窗口变化的测试情况, 随着该参数增长, AUC 分数值有小幅变化, 得分的波动范围在1%以内。考虑计算复杂度, 根据实验平均效果进行评估, 推荐设置向量维数大于64, 游走步长大于40, 游走次数5次, 游走窗口大小为10。

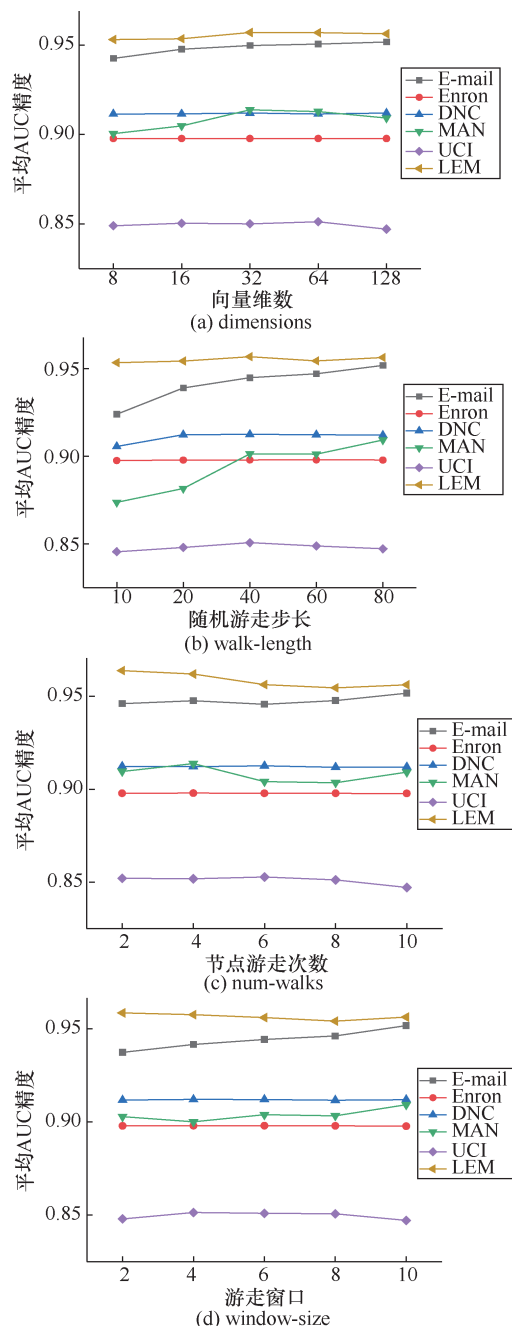


图8 不同游走参数的预测精度结果
Figure 8 Results of different walking parameters

实验4 训练时间窗口敏感性测试

对不同的数据集，分别设置训练集的窗口大小为3、5、7、9，观察对不同情况的预测效果，实验结果如图9所示， S_{ci} 随着窗口的变大AUC分数值也在变大； S_{com} 和 S_{mff} 在Enron和UCI数据集上也是随窗口线性变化的，在其他4个数据集中训练窗口大小为7时取得最优值，说明初始时刻距预测时刻较远，该时刻快照的拓扑信息反而会对链路预测产生一定负面影响。总体而言，大多数情况下训练窗口越大，AUC精度越好，但存在随窗口变化产生波动的情况，因此，针对不同的真实网络应用需要选择合适的窗口大小。

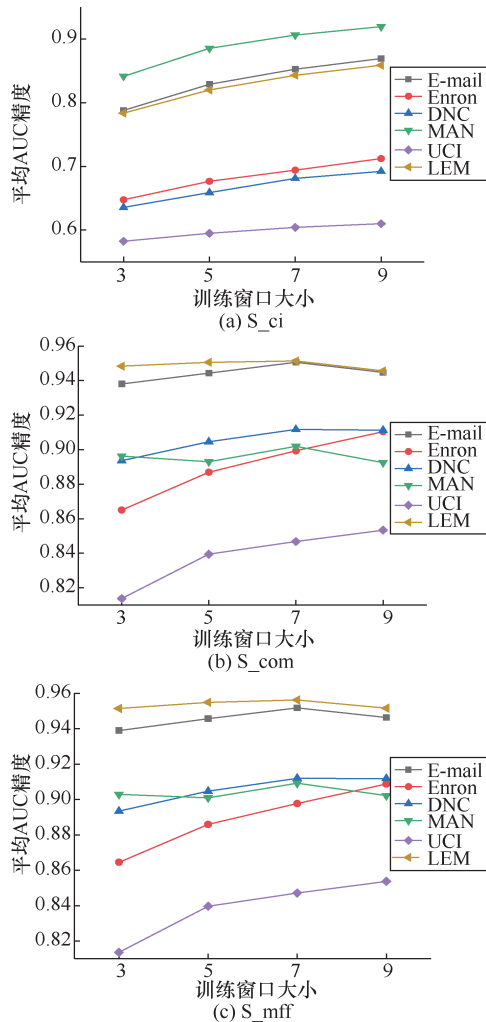


图9 不同方法随时间变化的AUC精度结果
Figure 9 Results of different train window size

实验5 算法时序鲁棒性测试

对所有6个数据集，从第8个快照窗口开始，

计算到数据集快照集合的第28个快照窗口，观察对不同时时刻的预测效果。由图10可知， S_{mff} 和 S_{com} 在6个数据集中基本在每一个时刻都获取相对较优的效果，而且没有发生较大幅度的波动，说明时序网络数据的波动对这两种指标影响较小。 S_{ci} 在Enron、MAN数据集中效果较好，但是AUC精度曲线有较大的波动，在其他5个数据集中的预测效果并不理想。总体而言， S_{mff} 指标和 S_{com} 指标具有较好的时序鲁棒性。

综上所述，本文在社团子图分析的基础上，分3个步骤提出了 S_{ci} 、 S_{com} 和 S_{mff} 这3种时序链路预测方法，可以有效利用网络子图信息和动态网络的时序信息，通过对比移动平均、网络表示学习、图神经网络的几种时序链路预测方法，实验结果表明 S_{ci} 在小规模网络表现较好，在较大规模网络中的效果一般， S_{com} 和 S_{mff} 在6个数据集上均取得较好的AUC精度，也具有较好的时序鲁棒性。

4 结束语

时序链路预测问题是近几年来复杂系统科学领域方面的热门研究问题。本文结合社会学理论，利用集体影响力、度中心性、介数中心性、接近中心性和动态网络拓扑在社团作用下的表示学习特征向量，按照算法的设计流程对预测过程进行了详细分析，分别提出了基于集体影响的相似度指标 S^{ci} ，基于社团有偏游走的相似度指标 S^{com} 和基于社团多特征融合嵌入表示的指标 S^{mff} 。首先，利用集体影响的概念，将球边界上所有节点的有效属性视为中心节点的影响力，并基于节点属性构造边的权重矩阵，并定义了 S^{ci} 指标。然后，利用基于集体影响的权重矩阵和弱连接优化算法，利用Louvain算法将动态网络划分为社区，并设计了社区内节点随机游走的机制，同时，设计了一种新的下一跳策略，优先考虑连通边的存在，并使用改进的Skip-gram模型获得节点表示向量，利用节点向量进行点乘操作得到 S^{com} 指标。最后，将集体影响、网络中心性影响和节点的表示学习向量连接起来，对新向量进行余弦相似性计算得到 S^{mff} 指标。在6个真实网络数据集的实验结果表明 S^{mff} 均取得最好的预测效果，

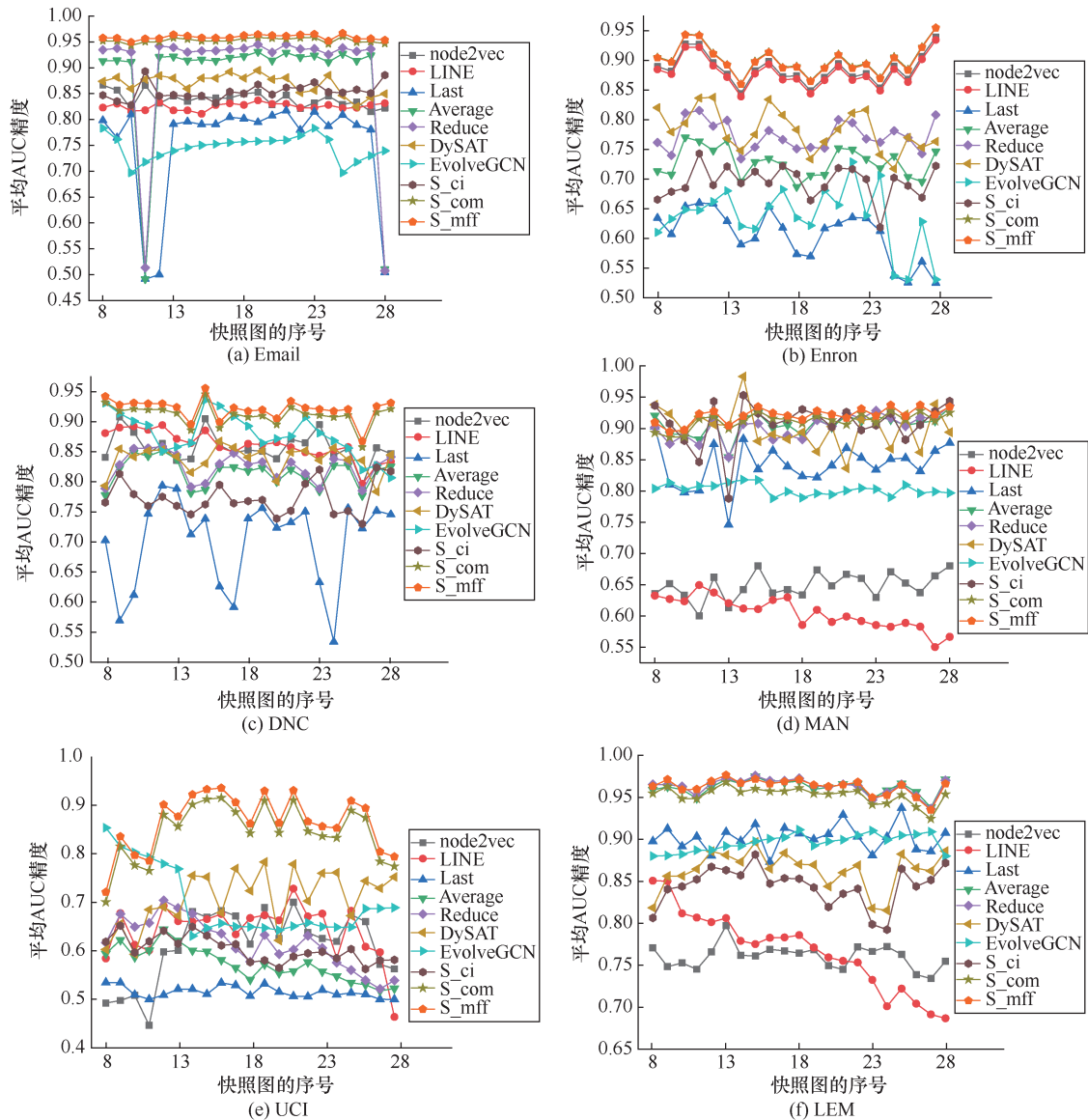


图10 不同数据集中各个算法的时序鲁棒性
Figure 10 Temporal robustness results of each algorithm in different data sets

并具有良好的时序鲁棒性。本文没有分析不同子图之间的相互作用对预测结果的影响，如何进一步刻画子图内部节点不同的作用和子图之间的影响是需要后续考虑的一个研究问题。

参考文献:

- [1] 吕琳媛, 任晓龙, 周涛. 网络链路预测: 概念与前沿[J]. 中国计算机学会通讯, 2016, 012(4):12-19.
LYU L Y, REN X L, ZHOU T. Network link prediction: concepts and frontiers[J]. Communications of China Computer Society, 2016, 12(4):12-19
- [2] WANG M, QIU L, WANG X. A survey on knowledge graph embeddings for link prediction[J]. Symmetry, 2021, 13(3): 485.
- [3] WU X, WU J, LI Y, et al. Link prediction of time-evolving network based on node ranking[J]. Knowledge-Based Systems, 2020, 195: 105740.
- [4] DIVAKARAN A, MOHAN A. Temporal link prediction: A survey[J]. New Generation Computing, 2019, 38(3):1-46.
- [5] BARABÁSI A L, ALBERT R. Emergence of scaling in random networks[J]. Science, 1999, 286(5439): 509-512.
- [6] MARTINEZ V, BERZAL F, CUBERO J C. A survey of link prediction in complex networks[J]. ACM Computing Surveys, 2017, 49(4):69.1-69.33.
- [7] LORRAIN F, WHITE H C. Structural equivalence of individuals in social networks[J]. Social Networks, 1977, 1(1):67-98.
- [8] ZHOU T, LÜ, ZHANG Y C. Predicting missing links via local information[J]. European Physical Journal B, 2009, 71(4):623-630.
- [9] LÜ L Y, 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.
- [10] KATZ L. A new status index derived from sociometric analysis[J]. *Psychometrika*, 1953, 18(1): 39-43.
 - [11] ZHAO H, DU L, BUNTINE W. Leveraging node attributes for incomplete relational data[C]//International Conference on Machine Learning. Sydney, Australia, 2017: 4072-4081.
 - [12] 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(006): 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.
 - [13] JAVARI, A, QIU, H, BARZEGARAN, E, JALILI, et al. Statistical link label modeling for sign prediction: Smoothing sparsity by joining local and global information[C]. *Proceedings of the 2017 IEEE International Conference on Data Mining (ICDM)*, Orleans, LA, USA, 2017: 1039-1044.
 - [14] PAN L, ZHOU T, LÜ L, et al. Predicting missing links and identifying spurious links via likelihood analysis[J]. *Scientific Reports*, 2016, 6(1): 1-10.
 - [15] LI T, WANG B, JIANG Y, et al. Restricted Boltzmann machine-based approaches for link prediction in dynamic networks[J]. *IEEE Access*, 2018, 6: 29940-29951.
 - [16] CHEN J, XU X, WU Y, et al. GC-LSTM: graph convolution embedded LSTM for dynamic link prediction[J]. *Applied Intelligence*, 2021, 52: 1-16.
 - [17] DAUD N N, AB HAMID S H, SAADOON M, et al. Applications of link prediction in social networks: A review[J]. *Journal of Network and Computer Applications*, 2020, 166: 102716.
 - [18] WANG P, XU B W, WU Y R, et al. Link prediction in social networks: The state-of-the-art[J]. *Science China Information Sciences*, 2015, 58(1): 1-38.
 - [19] 刘书新, 刘群, 杜凡. 基于模体演化与社区一致性的时序链路预测方法[J]. *计算机应用研究*, 2019, 36(12): 3674-3678, 3684.
LIU S X, LIU Q, DU F. Time series link prediction method based on motif evolution and community consistency[J]. *Application Research of Computers*, 2019, 36(12): 3674-3678, 3684.
 - [20] JORGE, VALVERDE-REBAZA, ALNEU, et al. Exploiting behaviors of communities of twitter users for link prediction[J]. *Social Network Analysis & Mining*, 2013, 3(4): 1063-1074.
 - [21] LIU H, HU Z, HADDADI H, et al. Hidden link prediction based on node centrality and weak ties[J]. *EPL (Europhysics Letters)*, 2013, 101: 18004.
 - [22] IBRAHIM N M A, CHEN L. Link prediction in dynamic social networks by integrating different types of information[J]. *Applied Intelligence*, 2015, 42(4): 738-750.
 - [23] VALVERDE-REBAZA J C, Lopes A. Link prediction in complex networks based on cluster information[C]//Advances in Artificial Intelligence-SBIA 2012, Berlin: Springer-Verlag, 2012: 92-101.
 - [24] WANG J, MA Y, LIU M, et al. A vertex similarity index using community information to improve link prediction accuracy[J]. *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2017: 158-163.
 - [25] 焦杨. 复杂网络社区检测、链路预测及应用[D]. 西安: 西安电子科技大学, 2018.
JIAO Y. Community detection, link prediction and applications in complex networks[D]. Xi'an: Xidian University, 2018.
 - [26] GÜNEŞ İ, GÜNDÜZ-ÖĞÜDÜCÜ Ş, ÇATALTEPE Z. Link prediction using time series of neighborhood-based node similarity scores[J]. *Data Mining & Knowledge Discovery*, 2016, 30(1): 147-180.
 - [27] OZCAN A, OGUDUCU S G. Supervised temporal link prediction using time series of similarity measures[C]//International Conference on Ubiquitous & Future Networks, Milan, Italy, 2017: 519-521.
 - [28] LIU SHUXIN, LIU QUN, Du fan. Time series link prediction method based on motif evolution and community consistency[J]. *Computer application research*, 2019, 36 (12): 3674-3678.
 - [29] HU W, TONG S, MENERSEN K, et al. Weather variability and the incidence of cryptosporidiosis: comparison of time series poisson regression and SARIMA models[J]. *Annals of Epidemiology*, 2007, 17(9): 679-688.
 - [30] SKARDING J, GABRYS B, MUSIAL K. Foundations and modeling of dynamic networks using dynamic graph neural networks: A survey[J]. *IEEE Access*, 2021, 9: 79143-79168.
 - [31] YANG Y, LICHTENWALTER R N, CHAWLA N V. (2014). Evaluating link prediction methods[J]. *Knowledge and Information Systems*, 45(3), 751-782.
 - [32] MORONE F, MAKSE H A. Influence maximization in complex networks through optimal percolation[J]. *Nature*, 2015, 524(7563): 65-68.
 - [33] BLONDEL V D, GUILLAUME J L, LAMBIOTTE R, et al. Fast unfolding of communities in large networks[J]. *Journal Statistical Mechanics: Theory and Experiment*, 2008, 10: 1008.
 - [34] GOLDBERG Y, LEVY O. Word2vec explained: deriving Mikolov et al.'s negative-sampling word-embedding method[J]. *arXiv*, 2014, arxiv:1402.3722.
 - [35] GROVER A, LESKOVEC J. Node2vec: Scalable feature learning for networks[C]. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 13-17 August 2016: 855-864.
 - [36] SELVARAJAH K, RAGUNATHAN K, KOBTI Z, et al. Dynamic network link prediction by learning effective subgraphs using CNN-LSTM[C]. *Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN)*, Glasgow, UK, IEEE: Glasgow, UK, 2020: 1-8.
 - [37] LESKOVEC J, KLEINBERG J, FALOUTSOS C. Graph evolution: densification and shrinking diameters[J]. *ACM transactions on Knowledge Discovery from Data (TKDD)*, 2007, 1(1).
 - [38] LESKOVEC J, LANG K J, DASGUPTA A, et al. Community structure in large networks: natural cluster sizes and the absence of large well-defined clusters[J]. *Internet Mathematics*, 2009, 6(1): 29-123.
 - [39] KUNEGIS J. Konect: The koblenz network collection[C]. *Proceedings of the 22nd International Conference on World Wide Web*. Rio de Janeiro, Brazil, 2013: 1343-1350.
 - [40] PANZARASA P, OPSAHL T, CARLEY K M. Patterns and dynamics of users' behavior and interaction: network analysis of an online community[J]. *Journal of the American Society for Information Science and Technology*, 2009, 60(5): 911-932.
 - [41] LESKOVEC J, KREVL A. SNAP Datasets: Stanford large network dataset collection[EB]. 2014.
 - [42] TANG J, QU M, WANG M, et al. Line: large-scale information network embedding[C]//*Proceedings of the 24th International Conference on World Wide Web*. Florence, Italy, 2015: 1067-1077.
 - [43] SANKAR A, WU Y, GOU L, et al. Dysat: Deep neural representa-

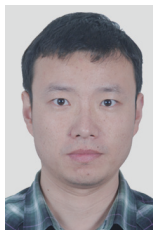
tion learning on dynamic graphs via self-attention networks[C]. Proceedings of the 13th International Conference on Web Search and Data Mining. Houston, TX, USA, 2020: 519-527.

- [44] PAREJA A, DOMENICONI G, Chen J, et al. Evolvegn: evolving graph convolutional networks for dynamic graphs[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York, NY, USA, 2020, 34(4): 5363-5370.

[作者简介]



朱宇航（1982- ），男，江苏徐州人，信息工程大学副教授，主要研究方向为链路预测、网络行为分析。



李英乐（1985- ），男，河北衡水人，信息工程大学副研究员，主要研究方向为社会网络分析、大数据分析。



李海涛（1982- ），男，山东泰安人，信息工程大学副研究员，主要研究方向为通信网络安全、嵌入式设计。



吉立新（1969- ），男，江苏淮安人，信息工程大学研究员、博士生导师，主要研究方向为通信网络安全、社会网络分析。



刘树新（1987- ），男，山东潍坊人，信息工程大学助理研究员，主要研究方向为复杂网络演化、链路预测。