

基于节点匹配度的动态网络链路预测方法

李聪, 季新生, 刘树新, 李劲松, 李海涛

(信息工程大学, 河南 郑州 450001)

摘 要: 现实世界存在众多真实网络, 研究真实网络中的动态演化趋势和时序性特征是热点问题。链路预测技术作为网络科学领域重要研究工具可通过挖掘历史连边信息推测网络演化规律, 进而对未来连边进行预测。通过分析动态真实网络中的拓扑结构演化, 发现通过分析网络拓扑中节点间的交互性和匹配度问题能够更充分捕捉网络的动态特征, 提出一种基于节点匹配度的动态网络链路预测方法。该方法对网络节点的属性特征进行分析, 定义基于原生影响力和次生影响力的节点重要性量化方法; 引入时间衰减因子, 刻画不同时刻网络拓扑对连边形成的影响程度; 结合节点重要性和时间衰减因子定义动态节点匹配度 (TMDN, temporal matching degree of nodes) 方法, 用于衡量节点对之间未来形成连边的可能性。在 5 个真实动态网络数据集中的实验结果表明, 相比现有 3 类主流动态网络链路预测方法, 所提方法在 AUC 和 Ranking Score 两种评价标准下均取得更优的预测性能, 预测结果最高提升 42%, 证明了节点间存在着交互匹配优先级, 同时证实了节点原生影响力和次生影响力的有效性。

关键词: 动态网络; 链路预测; 节点匹配度; 节点重要性; 时间衰减因子

中图分类号: TP393

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2022053

Link prediction method for dynamic networks based on matching degree of nodes

LI Cong, JI Xinsheng, LIU shuxin, LI Jinsong, LI Haitao

Information Engineering University, Zhengzhou 450001, China

Abstract: The research of dynamic evolutionary trends and temporal characteristics in real networks which are important part of the real world is a hot research issue nowadays. As a typical research tool in the field of network science, link prediction technique can be used to predict the network evolution by mining the historical edge information and then predict the future edge. The topological evolution of dynamic real networks was analyzed and it found that the interaction and matching between nodes in the network topology can capture the dynamic characteristics of the network more comprehensively. The proposed method analyzed the attribute characteristics of

收稿日期: 2021-01-05; **修回日期:** 2021-05-04

通信作者: 季新生, jixinsheng@ndsc.com.cn

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: 李聪, 季新生, 刘树新, 等. 基于节点匹配度的动态网络链路预测方法[J]. 网络与信息安全学报, 2022, 8(4): 131-143.

Citation Format: LI C, JI X S, LIU S X, et al. Link prediction method for dynamic networks based on matching degree of nodes[J]. Chinese Journal of Network and Information Security, 2022, 8(4): 131-143.

network nodes, and defined a node importance quantification method based on primary and secondary influences. Besides, a time decay factor was introduced to portray the influence of network topology on the formation of connected edges at different moments. Furthermore, the node importance and time decay factor were combined to define the Temporal Matching Degree of Nodes (TMDN), which was used to measure the possibility of future edge formation between node pairs. The experimental results in five real dynamic network datasets showed that the proposed method achieves better prediction performance under both AUC and Ranking Score, with a maximum improvement of 42%. It also proved the existence of interactive matching priority among nodes, and confirmed the effectiveness of both primary and secondary influence of nodes. As the future work, we will add diversified feature information to further deepen the analysis of dynamic real networks and then predict the evolution law more accurately.

Keywords: dynamic networks, link prediction, matching degree of nodes, node importance, time decaying parameter

0 引言

随着社会学^[1]、生物学^[2]、医学^[3]以及互联网技术^[4]的飞速发展,复杂网络^[5]被广泛应用于复杂系统的抽象化表示、形式化分析与深度挖掘等,极大地推动了网络科学的蓬勃发展。复杂系统^[6]中的个体,如人类、细胞、基站、路由器等可以看成实体,在网络拓扑中用节点表示,节点之间的连边用来描述实体之间的关系^[7]。链路预测作为研究复杂科学的重要技术之一,已成为近年来的研究热点问题^[8-11]。链路预测技术旨在利用复杂系统中已观测的结构信息、属性信息等,发现网络拓扑结构中的缺失连边,以及推测未来可能形成的连边^[12]。链路预测为生物学、医学、计算机网络、社会科学等领域的研究提供了强大的技术支持。通过对大量节点之间复杂关系的挖掘可清晰地分析网络结构,发现网络中连边的演化趋势^[13],实现对未来连边的预测。例如,通过分析社交群体近期的社交活动,利用链路预测技术寻找未来可能发生的诈骗、推销等异常行为。

近年来,社交网络分析在各行各业中引起了相当大的关注,被广泛应用于市场营销、业务流程建模及传染病模型。链路预测作为社交网络分析的重要手段能够捕捉社交网络信息,及时发现新的节点和连接。现实中的社交网络由个体及个体间复杂关系组成,是一个高度动态变化的网络。通过链路预测从已知的网络结构中寻找和检测隐藏的连边和节点,可进行异常节点挖掘、重要连边预测等,映射到真实社交网络中可识别诈骗信息。当前,对于静态网络的链路预测技术研究已经相当成熟,文献[14]详细对比分析了各种方法,

将其在局部和全局两个维度进行归类。对于动态网络的研究也在近几年发展迅速。Ahmed 等^[15]提出一种通过集成网络中的时间信息、社区结构和节点中心性进行动态网络链预测的方法,通过节点在社区中的交互情况构造特征向量来预测节点未来的重要性。Günes 等^[16]提出了一种演进网络的链路预测方法,先计算过去不同时间段的网络结构中不同节点的相似性得分,再通过时间序列预测模型 ARIMA 来预测未来节点的连边情况。Ahmed 等^[17]提出了一种针对时间不确定网络的链路方法,首先将不确定网络中的链路预测问题转化为确定网络的随机游走问题,然后在子图中计算节点与邻居之间的相似性分数。针对异构网络,Sett 等^[18]结合多域拓扑特征和时间维度,提出一种 TMLP 算法,利用二元交互情况和图拓扑特征变化情况,结合时间序列模型进行预测。但上述方法通过转化成静态图或考虑局部特征,容易丢失动态网络中重要拓扑信息,影响预测精度。

本文通过捕捉历史拓扑特征,并将其作为节点相似度的依据,提出一种基于节点匹配度的动态网络链路预测方法。该方法根据节点原生重要性和次生重要性定义节点匹配度,用于衡量节点间未来产生交互的可能性。进而结合网络拓扑的历史变化特征及时间影响力,捕捉网络拓扑的动态信息。实验结果表明,本文所提方法在动态网络链路预测方面相比现有方法取得更优的效果。

1 相关工作

一个动态网络图,可记为 $G=(V,E,T)$, 由节点的集合 $V=\{v_1, v_2, \dots, v_n\}$ 和边的集合 $E=\{e_1, e_2, \dots, e_n\}$ 以及时间 $T=\{t_1, t_2, \dots, t_n\}$ 组成,

每一个时刻对应一个网络子图，即 $G = \{G_1, G_2, \dots, G_n\}$ 。在此网络中，任意一条边对应一个节点的二元组 $e_{ij} = \{v_i, v_j\}$ ，且满足以下4个条件：

- 1) 任意节点不能形成自闭环；
- 2) 任意节点之间最多只能含有一条边；
- 3) 节点间的交互无方向；
- 4) 连边只代表节点之间关系的存在。

对于一个无向无权网络 $G = (V, E, T)$ ，将其中任一节点 v_i 的度记为 k_i ，表征节点 v_i 与邻域内其他节点产生交互的次数。令 $\Gamma^t(i)$ 表示 t 时刻节点 v_i 的邻居集合，有 $k_i = |\Gamma(i)|$ 。将节点 v_i 的影响力记为 I_i ，通常情况下可用节点度 k_i 衡量，将节点 v_i 与节点 v_j 之间连边的重要性表示为 $I_{e_{ij}}$ 。

2 方法介绍

原生能力^[19]与次生能力^[20]在现实生活中无处不在，大到生态系统，小至单一细胞，近些年频发的自然灾害也会造成原生灾难^[21]和次生灾难^[22]。例如，在社交网络中，有些个体天生擅长某些事物，称之为原生能力；个体具备依托已存关系而带动新关系形成的能力，此能力建立在原生能力的基础上，称之为次生能力。复杂网络作为真实网络的抽象，需准确刻画真实网络的网络特性及个体属性。在网络拓扑中描述实体的节点应反映出实体的原生能力和次生能力。度是衡量节点重要性^[23]的指标，是真实网络中实体原生能力在复杂网络中的抽象表现，在网络拓扑中可作为节点原生能力的体现。例如，在社交网络中，节点度高的节点之间更容易形成连接；在食物链网络中，节点度高象征着在食物链中的等级高，这部分节点更倾向于与节点度低的实体形成连边。在网络拓扑中，节点的度是由节点自身性质决定的，而连边是在节点度的基础上产生的，两者是真实系统中原生与次生逻辑关系的抽象。作为节点的次生产物^[24]，连边影响力被多次提起。连边具有带动新连边形成的能力，此能力决定于节点本身，这正是真实网络中次生能力的抽象，在网络拓扑中，称之为节点的次生能力。本文根据节点原生能力和次生能力定义节点原生重要性和次生重要性，并结合两者提出节点匹配度的概念。

2.1 节点重要性

对于节点重要性的衡量可从节点自身和邻域拓扑两个角度加以考察。连边是基于节点性质而形成的，可看作节点的次生产物。在复杂网络中，节点度可用于衡量节点在网络中的重要性程度，其可以解释为节点周围连边贡献的和^[25]。然而，节点度将每条边的贡献程度设为1，无法区分不同边的连接能力^[25]，因此，本文用连边连接能力刻画该贡献程度，定义节点次生重要性。

定义 1 节点原生重要性，即节点对于网络拓扑的重要性，可体现为节点度 k 。节点在拓扑结构中的贡献随着节点度的增加而越发变大，节点之于网络拓扑则越为重要。

定义 2 节点次生重要性，即节点所产生连边对于网络拓扑的重要性。连边的重要性决定其连接能力的大小。对于 t 时刻动态网络中的某一节点 v_i ，设与其直接相连的连边 e_{ij} 的连接能力为 $I_{ij}(t)$ ，则该节点的次生重要性为

$$I_i(t) = \sum_{j \in \Gamma^t(i)} I_{ij}(t) \quad (1)$$

下面讨论连边连接能力的量化问题。通常情况下，复杂网络中连边的形成不是随机的，而是具有特定倾向性。该倾向性在无标度网络中通常可用“优先连接 (PA, preferential attachment)”^[26]机制解释。该机制描述了复杂网络中“富者越富，强者越强”的现象。对于一条连边而言，可利用优先连接机制刻画其连接能力。传统 PA 方法只考虑端节点度的乘积，对网络中三角形结构具有的贡献能力未做区分。

三角形结构^[27]稳定性强，结构不容易发生改变，是复杂网络中最为稳定的局部结构之一。稳定性强则意味着结构不容易发生改变，在动态网络中则意味着连边不会随时间的变化而轻易消失。

图 1、图 2 所示分别为不同时刻的网络拓扑面示意图。在图 1 中，随着时间的演化， v_1 与 v_2 之间的三角形结构数量在增多，4 个时刻的网络结构的稳定性依次增强， e_{12} 作为三角形的公共边，其稳定性随三角形数量的增多而增强。在图 2 中， v_1 与 v_2 之间虽未形成三角形结构，但随着不完整三角形数量的增多，形成三角形结构的趋势更加强烈，随着时间的变化两节点之间产

生交互的概率增加。由上述示例可知，两节点之间形成的三角形结构数量可作为衡量连边稳定性的标准。

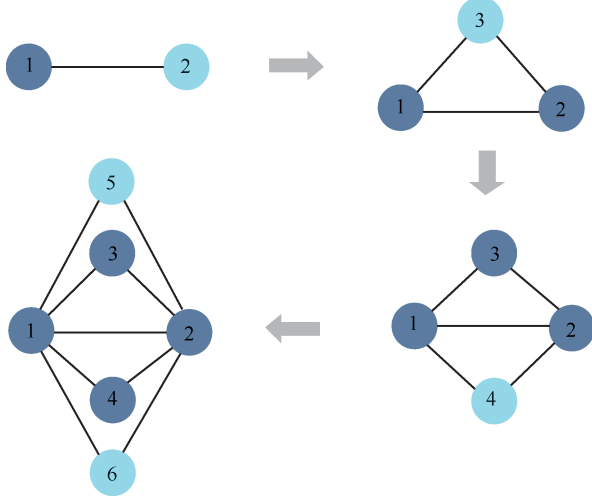


图 1 不同时刻网络拓扑图示例 1
Figure 1 Example of network topology diagram 1 at different times

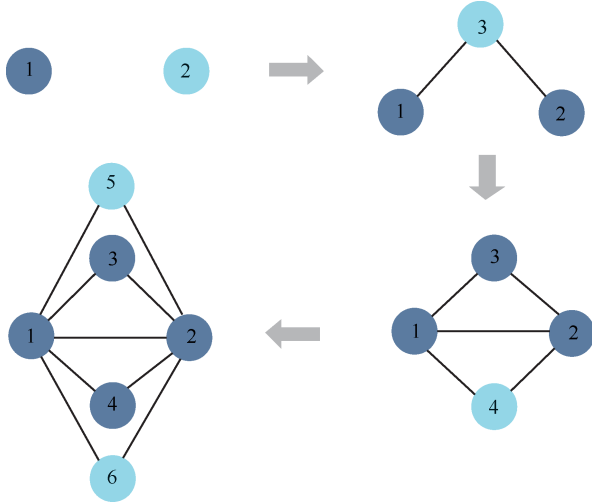


图 2 不同时刻网络拓扑图示例 2
Figure 2 Example of network topology diagram 2 at different times

以多节点的网络拓扑为例，图 3 为动态网络 G_1 第 N 时刻的网络拓扑子图。结合上述分析， e_{56} 因形成三角形结构而比无三角形结构的 e_{67} 重要，而从全局考虑， e_{67} 是连通 $\{v_1, v_2, v_3, v_4, v_5, v_6\}$ 团体和 $\{v_7, v_8, v_9, v_{10}\}$ 团体的唯一路径，由此可知， e_{67} 的重要性高于 e_{56} ，可见，三角形的数量不是衡量连边连通能力的唯一标准，还应考虑其端节点的性质和位置等。基于上述分析，从端节点自身性质出发，定义节点次生重要性。

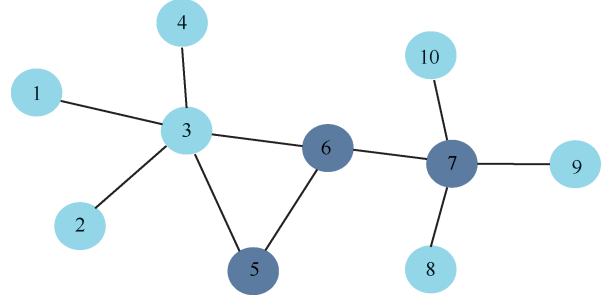


图 3 G_1 第 N 时刻网络拓扑子图
Figure 3 Sub-graph of network topology G_1 at time N

基于上述分析，本文考虑三角形结构与非三角形结构的贡献区别，基于 PA 指标定义连边的连接能力，如下：

$$I_{ij}(t) = (k'_i + P^t)(k'_j + P^t) \quad (2)$$

其中， P 是 t 时刻两节点之间三角形结构的数量， k'_i 、 k'_j 分别是 t 时刻节点 i 和节点 j 的度。

图 3 中通过上述方法可计算得出

$$I_6 = I_{67} + I_{36} + I_{56} = 48 \quad (3)$$

$$I_5 = I_{35} + I_{56} = 26 \quad (4)$$

由此可知， $I_6 > I_5$ 。从全局因素考虑， v_6 明显比 v_5 重要。经以上论述可得：在小规模网络拓扑中，节点次生重要性可以作为节点重要性的衡量标准。

图 4 是更为复杂的动态网络 G_2 在 T 时刻的网络拓扑子图。

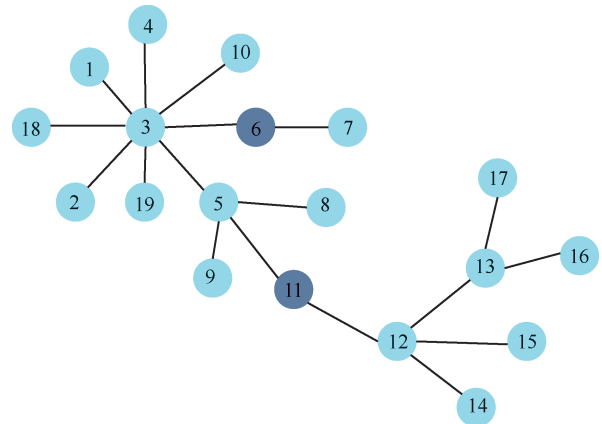


图 4 G_2 第 T 时刻网络拓扑子图
Figure 4 Sub-graph of network topology G_2 at time T

在图 4 中，有

$$I_6 = I_{36} + I_{67} = 18 \quad (5)$$

$$I_{11} = I_{511} + I_{1112} = 16 \quad (6)$$

在基于节点次生重要性的基础上可计算出 v_6 比 v_{11} 重要性大, 从全局考虑, v_{11} 是连接 v_3 团体和 v_{12} 团体的唯一桥梁, 其重要性高于 v_6 。显然仅考虑节点次生影响力而忽视节点自身性质作为衡量节点重要性的标准难以应用于大规模网络。基于上述现象, 从节点自身出发, 利用节点度刻画节点原生重要性, 定义节点重要性方法如下。

定义 3 节点重要性。对于网络中某一节点 v_i , 设与其直接相连的连边 e_{ij} 的连接能力为 $I_{ij}(t)$, 端节点度为 k_i^t , 考虑节点原生重要性和次生重要性后, 该节点的重要性指标为

$$I_i(t) = \sum_{j \in \Gamma(i)} I_{ij}(t) \cdot \frac{k_i^t}{k_i^t + k_j^t} \quad (7)$$

上述定义考虑了节点原生重要性和节点次生重要性, 并对节点间单个节点重要性进行量化, 因而对节点重要性的刻画更加合理。在图 4 所示的拓扑子图中可以看到, 通过此方法计算得到的节点重要性数值为

$$I_6 = I_{e_{36}} + I_{e_{67}} = \frac{68}{15} \quad (8)$$

$$I_{11} = I_{e_{511}} + I_{e_{1112}} = \frac{16}{3} \quad (9)$$

可知, $I_{11} > I_6$, 符合上述结构特征的分析。

2.2 节点匹配度

节点重要性决定节点的影响力, 在实际网络中, 影响力大的个体捕获资源和信息的能力更强。Wang 等^[28]提出, 在电网中, 每个电站都具有有限负载和有限容量, 而节点的承载能力决定了电站的负载和容量, 重要的节点会承载更多的负载, 与其他节点的匹配能力^[29]会更强。综合以上情形, 给出节点匹配度的定义, 用于刻画拓扑结构中节点间产生连边的概率。

定义 4 节点匹配度。基于动态网络中任意两节点 v_i 和 v_j 随时交互的特性, 将交互可能性称为节点匹配度。节点匹配度在动态网络结构中可用于量化两节点之间的连接可能性, 具体公式如下。

$$S'_{ij} = I_i(t) \cdot I_j(t) \quad (10)$$

其中, $I_i(t)$ 和 $I_j(t)$ 分别表示 t 时刻 v_i 和 v_j 的重要性指标, s'_{ij} 表示 v_i 与 v_j 之间的匹配度。当 $E_{ij}(t)=0$ 时, s'_{ij} 越高说明 v_i 与 v_j 之间产生连边的可能性越大; 当 $E_{ij}(t)=1$ 时, s'_{ij} 越高说明 v_i 与 v_j 之间的连边越稳定, 下一时刻连边越不容易消失。因此, 匹配度可作为链路预测的依据。

2.3 时间衰减因子

在动态网络中, 网络的拓扑结构是受时间影响的, 节点和连边是随着时间的流逝而变化的, 在这种情况下需要考虑时间的影响力, Ozcan 等^[30]将这种影响力刻画在静态网络中, 即用静态图表示每个时刻的拓扑结构, 这种方法会弱化最新连边对未来连边的影响力, 从而忽略网络结构本身的动态特征。本文方法考虑了网络动态特性, 把时间影响力刻画成一个累加的过程, 距离 T 时刻近的时间点对 G_T 影响力会更大, 同时每个时间点对于 T 时刻的影响程度是随时间间隔的增大而减少的, 以此定义一个时间衰减因子 α , 它的取值范围为 $0 \sim 1$ 。

结合衰减因子, 节点匹配度的定义如下。

$$s'_{ij} = \sum_{\tau=t-T}^t \frac{I_i(\tau) \cdot E_{ij}(\tau)}{t-\tau} \cdot \frac{I_j(\tau) \cdot E_{ij}(\tau)}{t-\tau} \cdot \alpha^{t-\tau} \quad (11)$$

式(11)计算得到的节点匹配度 s'_{ij} 可作为时刻 T 下节点 i 和 j 形成连边的可能性, 实现动态网络链路预测。将上述所提方法命名为动态节点匹配度 (TMDN, temporal matching degree of node)。

式(11)的定义将 v_i 与 v_j 之间连边的出现频率和时间考虑在内。连边的出现与消失反映的是真实网络中个体间关系的变化。例如, 在邮件网络中, 邮件往来的密集程度反映了个体间关系的密切与否, 高频率的往来抽象于复杂网络中意味着节点间连接更稳定, 将来时刻消失的可能性更低。同时, 连接出现的时间节点越近, 则为其分配的权重越大, 其影响力也越大。此方法相较静态图方法更加可靠, 考虑到新连接比旧连接有更大的影响力, 考虑因素更为全面。

3 衡量标准及数据集

3.1 对比方法介绍

(1) 传统方法

共同邻居 (CN, common nodes) 方法^[31]: 以两节点之间共同邻居数量为衡量标准, 数学公式表示为

$$S_{xy} = |\Gamma(x) \cap \Gamma(y)| \quad (12)$$

其中, $\Gamma(x)$ 表示 v_x 的邻居数量, $\Gamma(y)$ 表示 v_y 的邻居数量。

Adamic-Adar (AA) 方法^[32]: 此方法是建立在共同邻居的基础上的, 即两节点的共同邻居度越大, 则两节点之间的相似性分数越高, 数学公式表示为

$$S_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log k_z} \quad (13)$$

其中, 节点 z 是 v_x 和 v_y 的共同邻居, k_z 表示共同邻居的度。

资源分配 (RA, resource allocation) 方法^[33]: RA 与 AA 有异曲同工之妙, 两者的思想相同, 只是赋予节点权重的方式不同, 数学表达式为

$$S_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k_z} \quad (14)$$

其中, k_z 表示共同邻居的度。

局部路径 (LP, local path) 方法^[34]: 将三阶路径考虑进来, 数学表达式为

$$S_{xy} = A^2 + \alpha \cdot A^3 \quad (15)$$

其中, α 是可调参数, A 是网络的邻接矩阵, $(A^3)_{xy}$ 是节点之间路径为 3 的数目。

(2) 拓展方法

① 基于时间序列的拓展方法

首先将在 t 时刻观察到的网络 G 划分成多个时间分段的快照, 然后定义一个预测窗口 w , 将这些时间快照按照窗口长度进行划分。

$$S = \{[G_1, G_2, \dots, G_w], [G_{w+1}, G_{w+2}, \dots, G_{2w}], \dots, [G_{(n-1)w+1}, G_{(n-1)w+2}, \dots, G_{nw}]\} \quad (16)$$

利用相似性指标计算每个子图上节点对的相

似性分数, 构造时间序列, 利用平均预测模型, 得到相似性分数。预测模型为

$$\hat{Z}_t = \frac{Z_{t-1} + Z_{t-2} + \dots + Z_{t-n}}{T} \quad (17)$$

此类拓展方法用 TCN、TAA、TRA、TLP^[35] 表示。

② 基于时间因子的拓展方法

Güneş 等^[16]将时间因子应用于相似性方法内, 结合时间因子, 网络的邻接矩阵表示为

$$M_{t_0, T} = \sum_{t=0}^{T-1} A_{t_0+T}(i, j) \cdot \alpha^{T-t} \quad (18)$$

相应的拓展方法为

$$TF_CN = \sum_{z \in \Gamma(x) \cap \Gamma(y)} M_{xz}^T \cdot M_{yz}^T \quad (19)$$

$$TF_AA = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\sum_{c \in \Gamma(z)} M_{zc}^T} \quad (20)$$

$$TF_RA = \sum_{z \in \Gamma(x)} M_{xz}^T \cdot \sum_{z \in \Gamma(y)} M_{yz}^T \quad (21)$$

$$TF_LP = M_{t_0, T}^2 + \alpha M_{t_0, T}^3 \quad (22)$$

3.2 算法评价标准

与静态网络链路预测相同, 动态网络链路预测也要进行数据集的划分, 本文的划分方式是对每个时间窗口内的数据集按照 90% 和 10% 的比例划分训练集 E^T 和测试集 E^P , 其中 $E = E^T \cup E^P$ 且 $E^T \cap E^P = \emptyset$ 。本文采取 AUC 和 RS (ranking score) 两种主流评价指标。两种指标考虑的着重点不一样, 其中, AUC 是最常用的一种评价指标, 它能从整体衡量算法的精确度; 而 RS 则更多考虑所预测的边的排序。

在静态网络中, AUC 评价指标可以理解为在测试集 E^P 中随机选择一条边的分数值比未连接边的分数值大的概率。与静态网络不同的是, 动态网络的拓扑结构是变化的, 这意味着在下一个时刻整个网络的拓扑结构都是未知的, 所以在动态网络中 AUC 评价指标为: 在 $T+1$ 时刻之前的网络中随机选择一条边比 $T+1$ 时刻未连接边的分数值大的概率。独立比较 n 次, 其中测试集中边的分数大于未连接边的分数则加 1 分, 测试集中边的分数等于未连接边的分数则加 0.5 分。分别用 n' 和 n'' 记录上述两种情况的个数, 则 AUC 表示为

$$AUC = \frac{n' + 0.5n}{n} \quad (23)$$

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

$$RS_e = \frac{r_e}{|H|} \quad (24)$$

整个网络的平均 RS 可通过遍历求和得到, 计算方式如下。

$$RS = \frac{1}{|E^p|} \sum_{e \in E^p} RS_e = \frac{1}{|E^p|} \sum_{e \in E^p} \frac{r_e}{|H|} \quad (25)$$

3.3 数据集

通过调研近两年文献, 发现多数文献选择邮件数据集进行算法性能对比。为验证所提方法的性能, 使其更具有说服力, 本文选用 5 个来自不同领域的真实动态网络数据集进行实验。5 个真实数据集的基本特征参数如表 1 所示。以下详细介绍各数据集的具体信息。

(1) DNC emails (DNC) [36]: 由 2016 年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, 其节点表示邮件账号, 有向边表示某账号发往另一账号的邮件。

(2) Manufacturing emails (MAN) [36]: 一个制造业工厂内部员工的邮件往来网络。属于动态网络, 其中节点表示员工账号, 有向边表示在某一时刻由员工之间发送的邮件。

(3) LevantMonths (LEM) [37]: 一个由堪萨斯事件数据系统基于包含 WEIS 编码事件的文件夹收集的交互网络, 其中包含 8 个国家之间的交

互信息。原始数据集涵盖了 1979 年 4 月至 2004 年 6 月的事件。

(4) Email-Eu-core-temporal (EEC) [38]: 一个由欧洲某大型研究机构内部邮件往来数据构成的有向网络。节点表示邮件账号, 有向边表示某一账号在某时刻发往另一账号的邮件。根据数据来源, 原始数据集包含若干子数据集, 代表研究机构内不同部门的邮件往来。本文选用其中的 email-Eu-core-temporal-Dept1 (EEC1) 和 email-Eu-core-temporal-Dept2 (EEC2)。由于原始数据集的时间戳起始于 0, 确切起始日期无法确定。

4 实验结果与分析

为验证所提 TMDN 在真实数据集中的预测效果, 本文通过对比实验分析了 TMDN 在不同参数、时刻下 AUC 和 RS 结果, 具体分析如下。

4.1 平均性能对比

4.1.1 AUC 评价方法

首先研究所提 TMDN 方法在 AUC 衡量标准下的性能。图 5 为 TMDN 和现有方法在真实数据集中的 AUC 对比结果。可以看出, 与其他指标相比, TMDN 指标在 5 个网络中均取得了较高的预测精度。CN、TCN、TF_CN 仅考虑了二阶邻居数目, 在 AUC 评价标准下结果普遍一般, 但在 MAN 中, CN 明显高于 AA、RA 和全局指标 LP, 预测效果仅次于本文的指标, TCN 也优于 TAA、TRA、TLP; 而加入时间因素的拓展指标性能没有展示出优势, 性能表现整体不如传统指标, 造成此现象的原因是 MAN 数据集动态特性不明显, 考虑时间因素的拓展方法无法发挥性能优势。在 MAN 数据集中, CN 指标优势明显, 说明此数据集局部特征明显, 局部指标可发挥出性

表 1 5 个真实数据集的基本特征参数
Table 1 Basic characteristics of five realistic datasets

数据集名称	节点数	连边数	平均度	起始日期	结束日期	总时长	网络快照时长	网络快照数	窗口长度
DNC	2 029	39 264	19.35	2016/4/25	2016/5/23	29 天	1 天	29	10
MAN	167	82 927	496.57	2010/1/3	2010/9/27	268 天	7 天	39	10
LEM	485	196 364	809.75	1979/4/1	2004/6/1	302 月	6 月	50	10
EEC1	309	34 041	60546	1970/1/1	1971/6/9	525 天	7 天	75	10
EEC2	162	46 272	554.17	1970/3/17	1971/6/8	524 天	7 天	75	10

能优势,而考虑共同邻居的 AA、RA 不及 CN,说明该数据集共同邻居特征不明显。AA、RA 及其拓展指标考虑了共同邻居节点度,也表现出了不错的预测效果,在 DNC、LEM、EEC1、EEC2 这 4 个网络中预测效果明显优于 CN。而考虑全局信息的 LP 及其拓展方法 TLP、TF_LP,在大部分网络中表现出了较高的预测精度,甚至在 EEC2 网络中,LP、TLP 的 AUC 得分强于 TMDN,加入时间因子,将网络的动态信息考虑进来,全局方法 TF_LP 明显高于局部方法 TF_CN、TF_AA 和 TF_RA。整体而言,拓展方法表现优于传统方法,具体到数据集中表现却参差不齐,这是由于动态特性只是动态网络的一个特征,还要考虑其局部特征和全局特征等多个因素,在一些动态特性较高的网络中,会出现拓展方法低于传统方法的现象,所以,算法的普适性也是衡量指标性能的重要标准。

在考虑节点匹配度后,TMDN 预测效果明显领先,说明现实网络连边构建中均存在一定程度的匹配倾向性。同样考虑共同邻居间存在的社团关系,TMDN 在 5 个网络中的预测效果均优于局部相似性方法及拓展方法 TCN、TF_CN、TAA、TF_AA、TRA、TF_RA,且提升精度在 1%~33%。相比全局方法 LP、TLP、TF_LP,在 DNC、MAN、LEM 中同样有较高的提升,提升比例为 1%~42%,但在 EEC2 中,本文的方法效果不如 LP 和 TLP。总体上,相比其他方法,TMDN 方法的平均提升比率为 19.3%,最高提升比例为 42%。

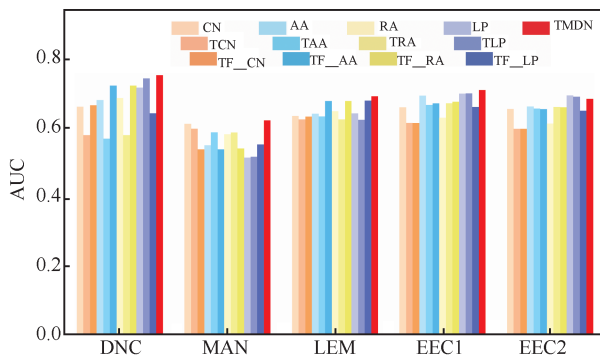


图 5 不同方法的平均 AUC 性能

Figure 5 Average AUC performance of different methods

4.1.2 RS 评价方法

本节在 RS 衡量标准下对 TMDN 的性能进行

对比分析。图 6 为 TMDN 与其他方法在 5 个真实网络中的 RS 对比结果。

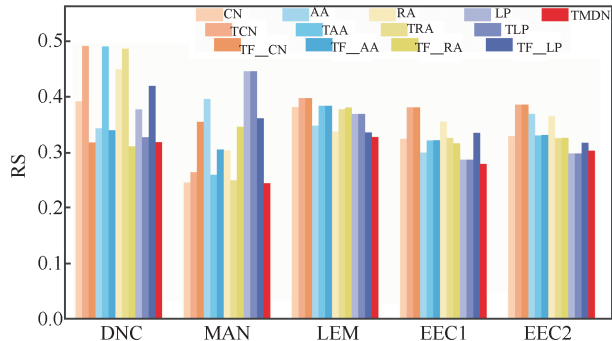


图 6 不同方法的平均 RS 性能对比

Figure 6 Comparison of average RS performance of different methods

相比其他相似性方法,本文所提 TMDN 在 5 个网络中均有较优的性能。与 AUC 结果不同,局部相似性方法及拓展方法在不同网络中表现各异,某些网络中 CN 表现较好,而在另一些网络中则表现较差。考虑共同邻居节点度的 AA、RA 方法,在 AUC 衡量标准下展现出了一定的优势,而在 RS 衡量标准下却表现出了较差的性能,仅在 LEM 中预测性能高于 CN。在大部分网络中,结合时间因素的局部拓展方法性能较传统方法有所提升,这说明在动态网络中,传统方法可能会忽略网络结构的动态特性,导致预测结果不佳。而这些局部方法性能表现各异,说明共同邻居节点本身信息的利用对于刻画动态网络结构特征的贡献难以界定,像 AA、RA 这些考虑了邻居节点度的额外因素并非一定会起到正面促进作用。全局指标 LP、TLP、TF_LP 在一些网络中表现出了较为稳定的预测性能,但在 MAN 中 LP 和 TLP 的预测效果却最差,说明基于全局信息的方法难以概述动态网络的全部属性,导致预测性能存在未知性。

TMDN 方法在 5 个网络中均表现出较好的性能,RS 得分普遍低于其他指标。在 DNC 和 MAN 网络中性能提升明显,相比其他指标,TMDN 的预测性能最高提升了 49.6%,在 DNC 中,TF_RA 展现出了不俗的预测精度,RS 得分甚至低于 TMDN,但也仅仅低 0.01。在其他 3 个网络中,TMDN 提升幅度不大,但 RS 得分低于其他任何指标,性能提升幅度在 1%~27%。总体上,TMDN

指标表现出了更加稳定的预测性能,RS 结果平均提升幅度为 16.5%, 最高提升幅度为 49.6%。

4.2 参数 α 对性能的影响分析

图 7 是 5 个网络中拓展指标关于参数 α 的性能曲线,在性能方面,除了在 EEC2 中的预测精度没有全局指标 TLP 高, TMDN 指标在其他 4 个网络中的预测精度均为最高,充分说明 TMDN 方法性能的优秀。从图中可以看出时间衰减因子 α 对 AUC 的影响,随着 α 的增大,5 个网络中指标的预测精度变化各异,但最终是呈衰减状态的,证实时间衰减因子是合理的,同时,对实验结果产生的影响跟预想的一样。

多数网络如 DNC、MAN、EEC1、EEC2 中,时间衰减因子 α 越靠近 1, AUC 的衰减速度越快。在 DNC 中, α 从 0.6 增加至 0.7 的过程,局部相似方法 TAA 和 TF_AA 的 AUC 会有一个明显的减小变化,而 TMDN 在 α 由 0.5 至 0.7 的变化过程中,呈先增大后减小的变化状态,并在 α 到达 0.6 时,预测效果最好, AUC=0.755。其他方法在此变化过程中则呈现比较平稳的递减状态。在 MAN 中,当 α 取值为 1 时,意味着前面时刻对于当前时刻的影响是相同的,会看到 TMDN、TF_AA、TF_CN、TF_RA、TF_LP 有一个明显的下降变化。当 $\alpha=0.7$ 时, TMDN 方法的 AUC 值

达到最大 0.619。而在 LEM 中,不管是局部方法、全局方法还是 TMDN,在 α 的变化过程中,都不同程度地出现了先增大后减小的变化状态, TF_AA、TF_CN、TF_RA 的 AUC 变化相同且变化幅度最小, TRA、TCN 变化一般,而 TF_LP 则发生了大幅度的变化,尤其是在 α 由 0.5 至 1 的变化过程中。当 $\alpha=0.3$ 时, TMDN 方法的 AUC 值达到最大值 0.688。

整体而言,当 $0.1 \leq \alpha \leq 0.9$ 时, TMDN 方法在一些网络中运行的结果差异不大,是因为在这些社交网络中群体之间通联关系的差异不大,所以对 α 不敏感。而在 LEM 中可以看出,算法的准确度对于时间衰减因子的选取较敏感。

4.3 不同时刻 t 的性能对比

图 8 是 AUC 评价标准下各方法随时间 t 变化的性能曲线;图 9 是 RS 评价标准下各方法随时间 t 变化的性能曲线。

在 AUC 的衡量标准下,从各个子图中可以看出 TMDN 方法曲线的性能优势。在图 8(a)中,将 DNC 网络划分成 28 个快照,整个预测过程分为 18 个时刻,只有在 T_4 、 T_5 、 T_{11} 、 T_{12} 这 4 个时刻 TMDN 的 AUC 略小于全局拓展指标 TLP,而相对其他方法, TMDN 则是全方位获胜,尤其是相比局部方法,提升十分明显。在图 8(b)中,数

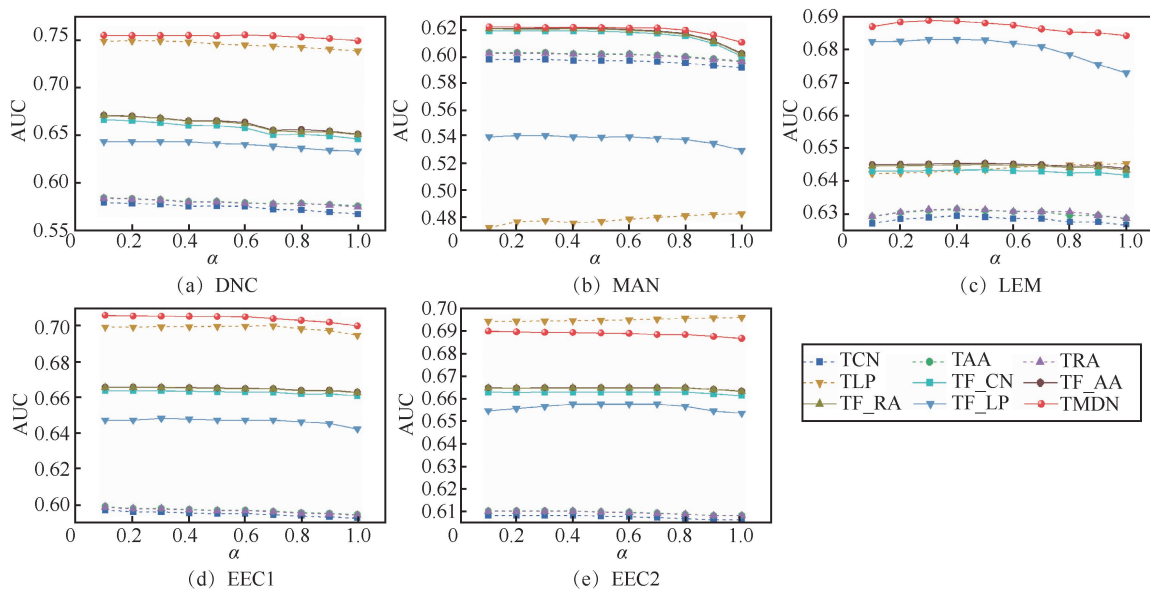
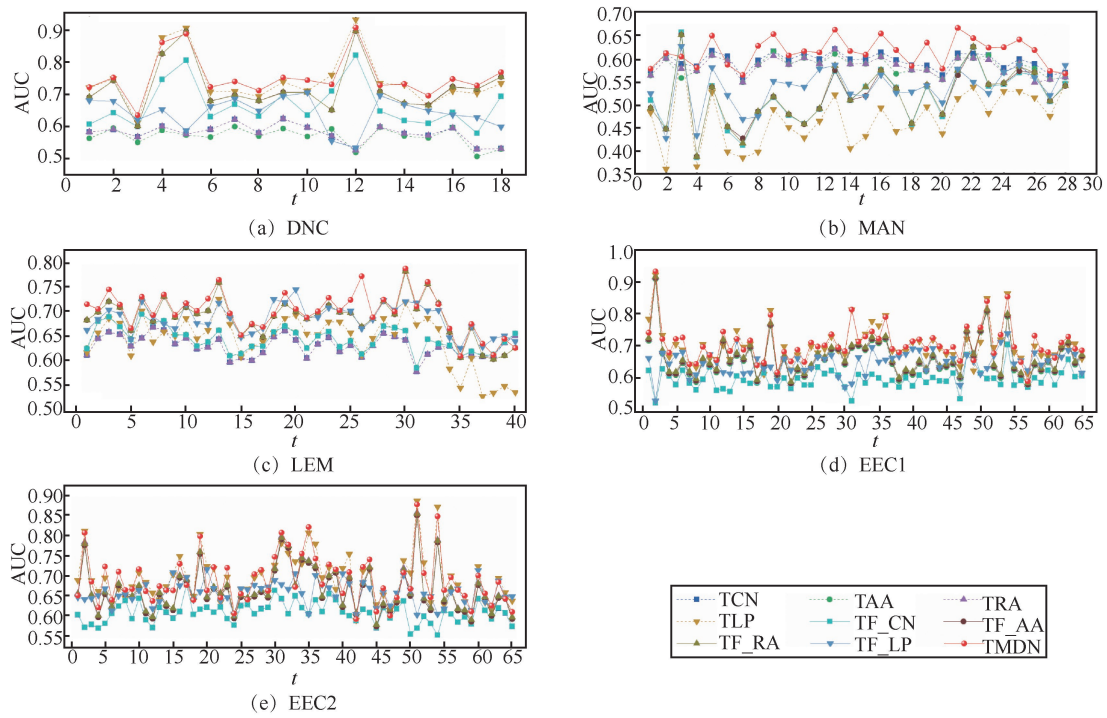
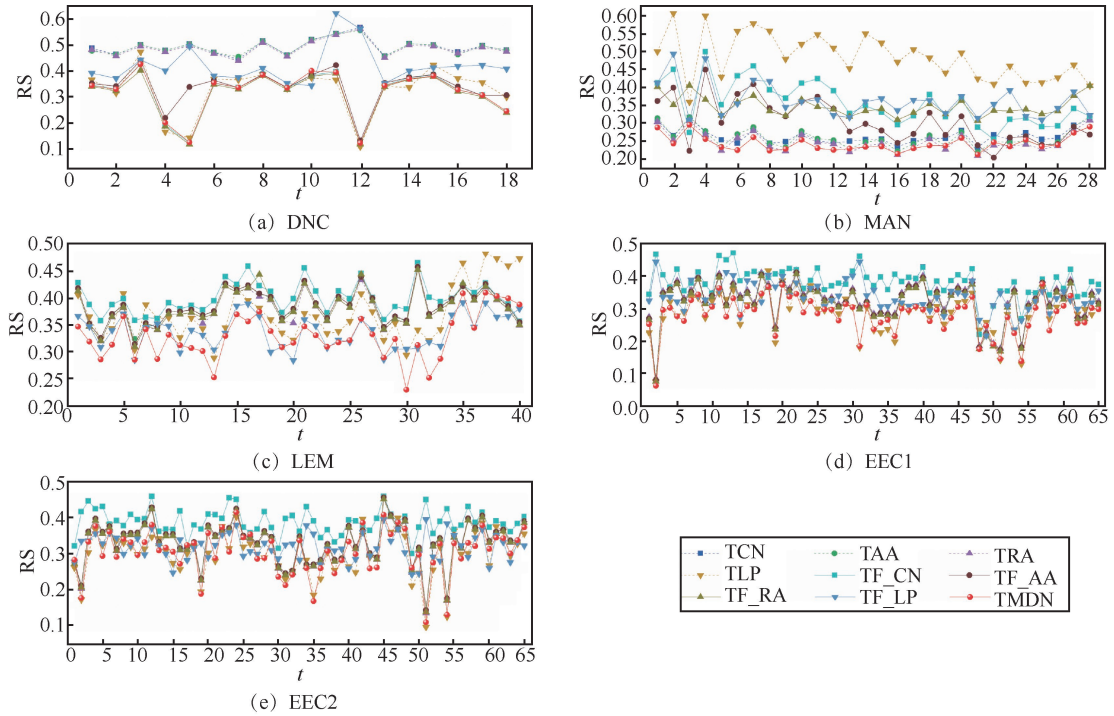


图 7 不同方法关于参数 α 的性能曲线

Figure 7 Performance curves of different methods with respect to parameters α

图 8 不同方法随时间 t 的 AUC 性能曲线Figure 8 AUC performance curve of different methods with time t 图 9 不同方法随时间 t 的 RS 性能曲线Figure 9 RS performance curve of different methods with time t

据集网络按照每 7 天 1 个快照, 划分了 38 个快照, 选择的时间窗口长度是 10, 预测过程被分为 28 个时刻, 首先会看到一个特别明显的特征变化,

即在 T_3 时刻全局方法 TLP、TF_LP 有一个明显的跳跃, 而考虑时间衰减因子的局部方法 TF_CN、TF_AA、TF_RA 也发生了跳跃, 并且超越了

TMDN。在其他时刻,各项指标变化平缓,TMDN方法性能遥遥领先。在图8(c)中,LEM网络按照每6个月划分了50个时间快照,并按照时间窗口长度10选取了40个时刻,在这些时刻中,TMDN方法性能保持领先,但从 T_{32} 开始,各项方法性能开始递减,并在 T_{35} 达到最小,这说明在此时间段,LEM受时间的影响较小,考虑了时间因素的拓展指标没有发挥出优势。在图8(d)和图8(e)中,两个网络各划分出了75个时间快照,并选取了65个时刻,各指标在两个网络中的变化特征大致相同,在开始后和结束前,各方法的AUC值发生了幅度较大的振荡。在整个过程中,TMDN指标的AUC在绝大多数时刻保持领先。

从不同方法随时间 t 的AUC性能曲线中可以看出,整体而言,TMDN方法的性能优于其他拓展指标,尤其是优于基于时间序列的拓展方法。通过不同时刻的AUC对比实验中,同样考虑时间影响因素的不同方法能表现出不一致的变化趋势,这说明时间衰减因子能准确描绘时间影响特征,更验证了所提算法的普适性和全面性,不仅考虑网络拓扑的局部信息,也考虑了拓扑信息的三阶路径;而相对加入时间衰减因子的拓展指标,本文方法展现出了较好的预测性能,主要是因为本文方法更加详细和全面地考虑网络拓扑的信息。

在RS评价标准下,其他方法在5个网络上表现各异,而TMDN表现十分稳定。在图9(a)中,全局方法没有展现出方法优势,相反,考虑邻居节点度的局部做拓展方法TF_RA、TF_AA与TMDN的RS得分几乎无差异,而共同邻居拓展方法由于考虑的拓扑信息较少,预测性能也较差。在图9(b)中,局部相似性方法展现出不俗的预测性能,在一些时刻,RS得分甚至比TMDN的RS得分更低,这是由于此网络节点数较少,只考虑局部信息的局部方法更能发挥出优势。在图9(c)中,全局方法TF_LP开始发挥优势,在一些时刻的RS得分比TMDN低,且两者在 T_{16} 至 T_{22} 之间的RS得分比TF_CN低40%。而在图9(d)和图9(e)中,基于时间序列的全局方法TLP的预测性能良好,反而TF_LP预测性能一般。

各指标在5个网络中的性能表现不太稳定,说

明这些方法虽然针对不同的网络可以表现出比较好的预测效果,但普适性和稳定性远远不如TMDN方法。在5个网络中,TMDN的RS得分最高降低了58%,在这几个网络中普遍降低3%~58%。

4.4 时间复杂度分析

假设网络中节点数目为 N , $k = \max\{k_i\}$, k 表示网络中节点最大的度, l 为节点之间的路径长度。在表征节点的局部影响力时,利用的是二阶邻居节点信息,时间复杂度为 $O(Nk^3)$;在考虑三阶路径时,增加了时间复杂度,具体为 $O(IN^2k)$ 。文中所提TMDN算法在考虑局部影响力的同时,考虑到了三阶路径,由于动态网络结构特殊性和不确定性,特将算法运行时间进行对比,并进行了归一化处理(如图10所示),表示为

$$T_{\text{avg}}^i = \frac{T^i}{\sum_i T} \quad (26)$$

其中, T^i 为方法 i 的运行时间。

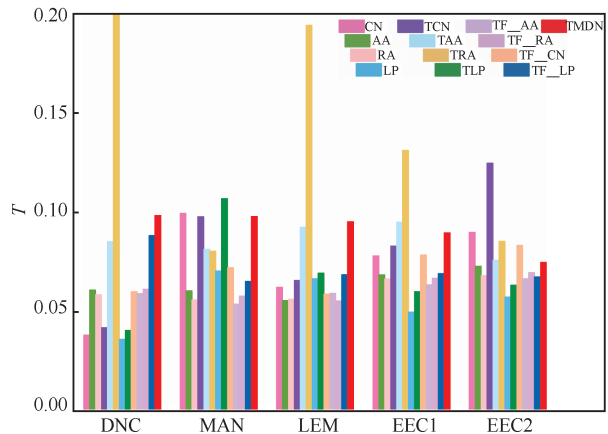


图10 不同方法时间消耗对比
Figure 10 Comparison of time consumption of different methods

从运行时间归一化对比图可看出,本文所提TMDN指标在提升性能的同时并未增加太多运行时间。

5 结束语

动态网络上的链路预测问题近年来受到网络科学领域学者的广泛关注。针对真实网络中连边的形成存在偏好性这一现象,结合网络节点自身属性特征,本文提出一种基于节点匹配度的动态

网络链路预测方法。该方法首先定义了衡量节点重要性的原生影响力和次生影响力方法；随后，将节点重要性与时间衰减因子相结合，提出用于动态网络链路预测的节点匹配度方法。在多个真实网络中的实验结果表明，所提 TMDN 方法相比现有方法在 AUC 和 RS 评价标准下均表现出更优的预测效果，验证了节点间匹配程度对动态网络连边形成的影响。所提方法实现简单，计算复杂度较低，可推广至超大规模动态网络上的链路预测场景。今后的研究中，可不局限于对网络拓扑的局部结构化研究，通过加入多样化的特征信息，对动态真实网络作进一步深层次分析可对更准确预测出其演化规律，具有现实意义。

参考文献:

- [1] MCGLOIN J M, KIRK D S. Social network analysis[M]. Springer TMDN York, 2010.
- [2] CUI Y, CAI M, DAI Y, et al. A hybrid network-based method for the detection of disease-related genes[J]. *Physica A: Statistical Mechanics and Its Applications*, 2018, 492: 389-394.
- [3] SLAVIN S. Medical student mental health: challenges and opportunities[J]. *Medical Science Educator*, 2018, 28(1): 13-15.
- [4] SHANMUKHAPPA T, HO I W H, TSE C K. Spatial analysis of bus transport networks using network theory[J]. *Physica A: Statistical Mechanics and Its Applications*, 2018, 502: 295-314.
- [5] 刘树新, 季新生, 刘彩霞, 等. 一种信息传播促进网络增长的网络演化模型[J]. *物理学报*, 2014, 63(15): 429-439.
- [6] LIU S X, JI X S, LIU C X, et al. A complex network evolution model for network growth promoted by information transmission[J]. *Acta Physica Sinica*, 2014, 63(15): 429-439.
- [7] ROLI A, VILLANI M, FILISETTI A, et al. Dynamical criticality: overview and open questions[J]. *Journal of Systems Science and Complexity*, 2018, 31(3): 647-663.
- [8] SCELLATO S, NOULAS A, MASCOLO C. Exploiting place features in link prediction on location-based social networks[C]//Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. 2011: 1046-1054.
- [9] KERRACHE S, ALHARBI R, BENHIDOUR H. A scalable similarity-popularity link prediction method[J]. *Scientific Reports*, 2020, 10: 6394.
- [10] WANG Y C, WANG Y, LIN X Y, et al. The influence of network structural preference on link prediction[J]. *Discrete Dynamics in Nature and Society*, 2020, 2020: 1-9.
- [11] HISANO R. Semi-supervised graph embedding approach to dynamic link prediction[J]. *CoRR*, 2016, abs/1610.04351.
- [12] 吕琳媛. 复杂网络链路预测[J]. *电子科技大学学报*, 2010, 39(5): 651-661.
- [13] LYU L Y. Link prediction on complex networks[J]. *Journal of University of Electronic Science and Technology of China*, 2010, 39(5): 651-661.
- [14] VON MERING C, JENSEN L J, SNEL B, et al. STRING: known and predicted protein-protein associations, integrated and transferred across organisms[J]. *Nucleic Acids Research*, 2005, 33(suppl_1): D433-D437.
- [15] SARKAR P, CHAKRABARTI D, JORDAN M. Nonparametric link prediction in large scale dynamic networks[J]. *Electronic Journal of Statistics*, 2014, 8(2): 2022-2065.
- [16] 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, 12(4): 12-19.
- [17] LYU L Y, REN X L, ZHOU T. Network link prediction: concept and frontier[J]. *Communications of CCF*, 2016, 12(4): 12-19.
- [18] AHMED N M, CHEN L. An efficient algorithm for link prediction in temporal uncertain social networks[J]. *Information Sciences*, 2016, 331: 120-136.
- [19] GÜNEŞ İ, GÜNDÜZ-ÖĞÜDÜCÜ Ş, ÇATALTEPE Z. Link prediction using time series of neighborhood-based node similarity scores[J]. *Data Mining and Knowledge Discovery*, 2016, 30(1): 147-180.
- [20] AHMED N M, CHEN L, WANG Y L, et al. Sampling-based algorithm for link prediction in temporal networks[J]. *Information Sciences*, 2016, 374: 1-14.
- [21] SETT N, BASU S, NANDI S, et al. Temporal link prediction in multi-relational network[J]. *World Wide Web*, 2018, 21(2): 395-419.
- [22] 陈敏, 崔大方, 黄平, 等. 3 种乡土水生植物对富营养化水体净化能力比较[J]. *安徽农业科学*, 2019, 47(7): 63-65, 69.
- [23] CHEN M, CUI D F, HUANG P, et al. Comparison of purification ability of three native aquatic plants to eutrophic water bodies[J]. *Journal of Anhui Agricultural Sciences*, 2019, 47(7): 63-65, 69.
- [24] CROMIE G, COLLINS J, LEACH D. Sequence interruptions in enterobacterial repeated elements retain their ability to encode well-folded RNA secondary structures[J]. *Molecular Microbiology*, 1997, 24(6): 1311-1314.
- [25] 严丽军. 自然灾害的灾情信息集成: 理论与实证研究[D]. 上海: 上海师范大学, 2016.
- [26] YAN L J. Disaster information integration of natural disasters: theoretical and empirical research. Shanghai: Shanghai Normal University, 2016.
- [27] 苗会强, 刘会平, 范九生, 等. 汶川地震次生灾害的成因、成灾与治理[J]. *地质灾害与环境保护*, 2008, 19(4): 1-5.
- [28] MIAO H Q, LIU H P, FAN J S, et al. Secondary disasters, hazard chains and the curing in Wenchuan earthquake-hit area, West China[J]. *Journal of Geological Hazards and Environment Preservation*, 2008, 19(4): 1-5.
- [29] 陈嘉颖, 于炯, 杨兴耀, 等. 基于复杂网络节点重要性的链路预测算法[J]. *计算机应用*, 2016, 36(12): 3251-3255, 3268.
- [30] CHEN J Y, YU J, YANG X Y, et al. Link prediction algorithm based on node importance in complex networks[J]. *Journal of Computer Applications*, 2016, 36(12): 3251-3255, 3268.
- [31] LIU F X, YANG H, WANG L M, et al. Biosynthesis of the high-value plant secondary product benzyl isothiocyanate via functional expression of multiple heterologous enzymes in *Escherichia coli*[J]. *ACS Synthetic Biology*, 2016, 5(12): 1557-1565.

- [25] 杨育捷. 复杂网络下基于拓扑相似性的链路预测研究[D]. 北京: 北京邮电大学, 2019.
YANG Y J. Research on link prediction based on topological similarity in complex networks[D]. Beijing: Beijing University of Posts and Telecommunications, 2019.
- [26] VÁZQUEZ A. Growing network with local rules: preferential attachment, clustering hierarchy, and degree correlations[J]. Physical Review E, 2003, 67(5): 056104.
- [27] MUBAYI D. Structure and stability of triangle-free set systems[J]. Transactions of the American Mathematical Society, 2007, 359(1): 275-291.
- [28] WANG X Y, CAO J Y, QIN X M. Study of robustness in functionally identical coupled networks against cascading failures[J]. PLoS One, 2016, 11(8): e0160545.
- [29] 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.
- [30] ÖZCAN A, ÖĞÜDÜCÜ Ş G. Supervised temporal link prediction using time series of similarity measures[C]//Proceedings of 2017 Ninth International Conference on Ubiquitous and Future Networks (ICUFN). 2017: 519-521.
- [31] LORRAIN F, WHITE H C. Structural equivalence of individuals in social networks[J]. The Journal of Mathematical Sociology, 1971, 1(1): 49-80.
- [32] ADAMIC L A, ADAR E. Friends and neighbors on the Web[J]. Social Networks, 2003, 25(3): 211-230.
- [33] OU Q, JIN Y D, ZHOU T, et al. Power-law strength-degree correlation from resource-allocation dynamics on weighted networks[J]. Physical Review E, Statistical, Nonlinear, and Soft Matter Physics, 2007, 75(2): 021102.
- [34] LYU L, JIN C H, ZHOU T. Similarity index based on local paths for link prediction of complex networks[J]. Physical Review E, Statistical, Nonlinear, and Soft Matter Physics, 2009, 80(4): 046122.
- [35] DA SILVA SOARES P R, PRUDÊNCIO R B C. Time series based link prediction[C]//Proceedings of 2012 International Joint Conference on Neural Networks (IJCNN). 2012: 1-7.
- [36] KUNEGIS J. KONECT: the Koblenz network collection[C]//Proceedings of the 22nd International Conference on World Wide Web - WWW '13 Companion. 2013: 1343-1350.
- [37] Batagelj V, Mrvar A. Pajek Datasets[EB/OL].
- [38] PARANJAPE A, BENSON A R, LESKOVEC J. Motifs in temporal networks[C]//Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. 2017: 601-610.

[作者简介]



李聪(1996-), 男, 山东菏泽人, 信息工程大学硕士生, 主要研究方向为复杂网络、链路预测、网络安全。



季新生(1969-), 男, 河南驻马店人, 信息工程大学教授、博士生导师, 主要研究方向为通信网架构、内生安全、复杂网络。



刘树新(1987-), 男, 山东临沂人, 信息工程大学助理研究员, 主要研究方向为复杂网络、链路预测、社交网络分析等。



李劲松(1992-), 男, 山东泰安人, 信息工程大学博士生, 主要研究方向为复杂网络、链路预测、大数据挖掘、网络安全。



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安全、数据处理和嵌入式设计。