

• 复杂性科学 •

融合用户行为同步指标的链路预测研究



王 曦, 许 爽, 许小可*

(大连民族大学信息与通信工程学院 辽宁 大连 116600)

【摘要】 现有链路预测方法大多基于网络结构相似性及连边的权重特征, 没有有效挖掘连边权重形成的时序信息。考虑到两个节点行为的时间同步性往往是由于两个节点存在链接造成的, 因此在网络结构的重构研究中通常利用节点的行为同步性来反推它们之间是否存在链接关系。该文尝试将节点同步性信息这一网络重构的方法引入链路预测领域, 提出一种网络拓扑相似性上融合节点行为同步指标的链路预测算法。经过两类 6 种真实网络数据的比较分析, 发现该算法可有效提高链路预测准确率, 相比现有方法, Precision 指标提高了 15.3%~68.2%。该研究不仅发现节点局部结构相似性和节点行为同步指数对链路预测的共同影响, 也揭示了不同类别真实加权网络的内在结构和动态特征。

关键词 链路预测; 网络重构; 结构相似; 同步指数

中图分类号 TP391 **文献标志码** A **doi**:10.12178/1001-0548.2020241

Link Prediction by Fusing Synchronization Index of User Behaviors

WANG Xi, XU Shuang, and XU Xiao-ke*

(College of Information and Communication Engineering, Dalian Minzu University Dalian Liaoning 116600)

Abstract Most of existing methods of link prediction are based on the similarity of network structures and the weight of edges, but they do not effectively use temporal information of forming the weight of edges. The behavior synchronization of two nodes is often caused by the link relationship between them, so the behavior synchronization of nodes has been widely used in many researches of network structure reconstruction to conjecture whether there is a link relationship between any pair of nodes. In this study, we attempt to introduce node synchronization information into the field of link prediction, and propose a novel link prediction algorithm which integrates the synchronization index of node behaviors with network topological similarity. By analyzing and comparing two types of six real-life network data, the proposed method can effectively improve the accuracy of link prediction. Compared with the existing methods, the performance of precision can increase by 15.3% to 68.2%. This study not only finds the joint influence of local structure similarity and behavior synchronization index on link prediction, but also reveals intrinsic structures and dynamic characteristics of different types of real-life weighted networks.

Key words link prediction; network reconstruction; structural similarity; synchronization index

真实世界存在各种复杂系统, 如果将复杂系统的主要元素抽象为节点和连边, 就可将其抽象为复杂网络, 如科学家合作网络^[1]、万维网络^[2]、社交网络^[3]和交通网络^[4]等。当网络的某些结构信息缺失时, 可应用链路预测^[5]方法来分析节点之间不确定的或潜在的链接关系^[6]。如在线社交服务中, 链路预测可以挖掘社交用户的兴趣并挖掘用户之间的潜在关系, 为用户推荐感兴趣的朋友^[7]。在蛋白质新陈代谢网络中, 链路预测可推断蛋白质之间的交

互作用, 从而有针对性地进行实验来节约成本^[8]。在科学家合作网络中, 可以预测科学家的合作模式和强度, 对理解科研合作的组织方式具有重要的作用^[9]。

链路预测研究中, 最经典的方法是节点局部结构的相似性, 如共同邻居 (common neighbors, CN)、Adamic-Adar 指标^[10]、资源分配指标 (resource allocation, RA)^[11]、局部路径相似性指标^[12]及局部随机游走的相似性指标^[13]等, 但通常不考虑节点之

收稿日期: 2020-04-21; 修回日期: 2020-09-02

基金项目: 国家自然科学基金 (61773091); 辽宁省高等学校创新人才支持计划 (LR2016070); 辽宁省“兴辽英才”计划 (XLYC1807106)

作者简介: 王曦 (1995-), 女, 主要从事社交网络数据挖掘和链路预测方面的研究。

通信作者: 许小可, E-mail: xuxiaoke@foxmail.com

间链接的权重大小。真实复杂系统中元素之间的交互强度往往也蕴含着关键信息, 可将其抽象为加权网络进行分析, 因此链路预测研究开始重视如何充分利用网络连边的权重信息。文献 [14] 提出了基于加权共同邻居指标 (weighted common neighbors, WCN)、加权 Adamic-Adar 指数 (weighted Adamic-Adar, WAA) 及加权资源分配指标 (weighted resource allocation, WRA) 等链路预测方法, 通过对多个实证网络的研究发现了加权链路预测中“弱链接的强作用”。文献 [15] 提出了一种基于可靠线路的链路预测方法 (rWCN, rWAA, rWRA), 通过对可靠路线赋予更高权重来达到更鲁棒的预测性能。文献 [16] 提出了一种局部加权路径指标, 精细利用路径的权重特征来实现准确预测。文献 [17] 改进了结构算法使其更适用于有向加权网络的链路预测, 如改进的共同邻居算法 (modified common neighbors, MCN)、改进的 Jaccard 系数 (MJC)、改进的 Adamic Adar 算法 (modified Adamic Adar, MAA) 和改进的优先链接算法 (modified preferential attach, MPA)。文献 [18] 设计了一种高效共同邻居指标, 该指标通过分析共同邻居节点的拓扑有效性对连边端点两侧资源分配过程的影响来刻画节点间相似性, 在 15 个实际网络数据实验中表明, 该方法具有较高的预测精度。文献 [19] 提出了一种链路权重预测的无监督混合方法, 结合网络的权重一致性和节点的链路权重潜在特征可有效解决链路权重预测问题。文献 [20] 通过对网络中真实存在的连边进行异常度排名, 将链路预测应用在大规模瘫痪状态下的电力系统的恢复, 该方法不仅可以快速连通电源发电机, 也能及时恢复重要线路。文献 [21] 提出了基于链路线形的危险链路预测模型, 对新加入交通网络的链路进行预测, 从而在碰撞发生之前进行相应的整治, 该模型可有效准确地对危险链路进行预测。

虽然上述研究相对于无权链路预测方法都取得了更好的性能, 但是这些方法仅仅是使用连边的权重信息, 没有考虑挖掘该连边权重形成的时序信息。在很多加权网络尤其是社会网络中, 连边的权重是由该连边连接的两个节点多次互动形成的, 这种互动行为具有一定的时间特性, 也是人类动力学研究的重要内容^[22]。经过研究发现, 两个节点行为的时间同步性往往是由于两个节点存在链接造成的^[23], 因此可利用两个节点的行为同步性来反推它们之间是否存在链接关系。目前该类方法已经广泛应

用于网络结构的重构研究中^[24]。文献 [25] 提出检测嘈杂实验数据中相位同步的方法, 实验发现该方法可有效衡量网络群体中个体的同步程度, 并探究了同步性如何影响个体之间的关系。文献 [26] 利用鸽子飞行轨迹的时序相关性, 在成员之间重构了完整明确的网络层次结构。文献 [27] 分析了时间滞后衰减作用对于网络重构的潜在影响, 取得了较好的网络重构效果。文献 [28] 通过对具有同步和异步动态特性的网络进行研究, 揭示了机器学习方法可以重构交互网络并识别网络内在的动力学特征。文献 [29] 根据由个人状态的时间序列组成的传播数据, 重建了人与人之间牢固关系的主干, 并成功地将该方法整合到了有针对性的免疫策略中。上述研究都是将节点行为同步信息应用在网络重构中, 目前将此类信息应用于链路预测的研究还很少见。

网络重构和链路预测通常采用不同的研究范式和算法, 本文尝试将节点同步性信息这一经典网络重构的方法引入到链路预测领域, 提出一种网络拓扑结构上融合节点行为同步指数的链路预测方法。该方法将节点局域结构相似性 (共同邻居指标) 和节点行为在时间上的相似性 (行为同步指数) 相融合, 充分利用节点之间的拓扑结构域和行为时间域特征。相对于仅使用网络域基于结构和权重的主流方法, 其链路预测效果更显著。同时, 为了探究局域结构和同步指数对链路预测的共同关系, 尝试使用一个可变参数调整两者的融合比率, 实验表明可变参数在较大范围内都不会影响链路预测的结果, 说明了本文所提方法具有较好的鲁棒性。

1 问题描述与评价指标

1.1 问题描述

定义加权网络 $G = (V, E, W)$, 其中 V 是所有节点的集合, E 是网络中连边的集合, W 是网络中连边权重的集合。定义 T 为网络中最大与最小时间戳的差值, 将时间进行相对时间处理后基于 $[0, T]$ 将整个数据划分为两段时间间隔 $[0, T_n]$ 和 $[T_{n+1}, T]$ 。将 $[0, T_n]$ 时间上数据的连边集合作为训练集 E^T , 将 $[T_{n+1}, T]$ 时间上的连边集合作为测试集 E^P 。这两个边集之间的关系为: $E^T \cup E^P = E$, $E^T \cap E^P = \phi$, 且训练集 E^T 与测试集 E^P 的样本为 9 : 1。在科学家合作网络中, 时效信息按年计。在短信网络中时效信息按秒计。在本研究中, 待分析的所有节点既在一

个周期内存在,也在第二个周期内存在,即 $[T_{n+1}, T]$ 中出现的所有节点都是 $[0, T_n]$ 上已经存在的节点。对于仅在 $[0, T_n]$ 和 $[T_{n+1}, T]$ 中出现的节点本文做了删除处理。

1.2 评价指标

Precision 只关注前 L 条高分连边的性能,根据出现连边可能性的分数值从大到小排序,如果有 m 条边是属于存在边集合,那么 **Precision** 定义为:

$$\text{precision} = \frac{m}{L} \quad (1)$$

由该式可知, m 越大则 **Precision** 值越高,预测越准确。

2 链路预测方法

2.1 基于共同邻居的相似性指标

1) CN 指标

共同邻居 (CN) 指标是通过计算两个节点共同邻居的数量来判断节点相似性。定义节点 x 的邻居集合 $\Gamma(x)$, 节点 y 的邻居集合为 $\Gamma(y)$, 共同邻居指标定义为:

$$S_{xy}^{\text{CN}} = |\Gamma(x) \cap \Gamma(y)| \quad (2)$$

一般来说,节点 x 和节点 y 之间的共同邻居越多,连边的分数值越大,节点的相似性越大。

2) AA 指标

在共同邻居的基础上,考虑两端节点度的影响,则演变为 AA 指标。AA 指标的思想是度小的共同邻居节点贡献大于度大的共同邻居节点。该方法通过共同邻居的度,为每个节点附上一个权重值。这个值为该节点度的对数的倒数,定义为:

$$S_{xy}^{\text{AA}} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log k_z} \quad (3)$$

式中, k_z 表示 x 和 y 的共同邻居节点 z 的度。

3) RA 指标

RA 与 AA 相似,都对节点度大的共同邻居的贡献进行抑制,即节点度越小,对网络连接产生的贡献就越大,RA 指标的抑制程度相对更大。因此,当网络的节点的度都较小时,两者区别不大。而当网络节点中存在较大度时,因为对度大节点有更高的抑制性,所以 RA 展现出相对较好的预测性能^[1]。RA 指标定义如下:

$$S_{xy}^{\text{RA}} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k_z} \quad (4)$$

2.2 基于共同邻居的加权相似性指标

加权网络进行链路预测时,不仅要考虑网络的拓扑结构,也要考虑连边的权重,才能取得较好的预测效果。将 CN、AA、RA 扩展到加权的形式,就可以将连边权重的信息应用到加权网络的链路预测中。常用的 3 种加权指标 WCN、WAA 和 WRA 的定义为如下形式:

1) WCN 指标:

$$S_{xy}^{\text{WCN}} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} W(x, z) + W(z, y) \quad (5)$$

式中, $W(x, z)$ 为两节点 x 和 z 的连边权重; $W(z, y)$ 为两节点 z 和 y 的连边权重。

2) WAA 指标:

$$S_{xy}^{\text{WAA}} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{W(x, z) + W(z, y)}{\log(1 + s(z))} \quad (6)$$

式中, $s(z)$ 表示邻居节点 z 的强度; WAA 根据每个共同邻居的强度值进行了加权。

3) WRA 指标:

$$S_{xy}^{\text{WRA}} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{W(x, z) + W(z, y)}{s(z)} \quad (7)$$

式中, WRA 是 WCN 的另一种权重赋予方式,分母不再采用强度的对数,而直接采用节点强度进行了加权。

2.3 基于可靠路线的加权相似性指标

在加权网络链路预测中,如何利用权重信息是研究重点。在通信网络中,信息的传输往往需要依赖最可靠的线路将数据包从源节点传输到目标节点,以保证数据传输过程中不被损坏。受通信网络的启发,文献 [15] 提出了可靠路线加权指标的链路预测方法,假设所有连边的权重都是独立的,通过连边权重的乘积计算两个节点连边的相似性得分,定义如下:

1) 加权可靠路线 CN 指标 (rWCN):

$$S_{xy}^{\text{rWCN}} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} W(x, z)W(z, y) \quad (8)$$

2) 加权可靠路线 AA 指标 (rWAA):

$$S_{xy}^{\text{rWAA}} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{W(x, z)W(z, y)}{\log(1 + s(z))} \quad (9)$$

3) 加权可靠路线 RA 指标 (rWRA):

$$S_{xy}^{\text{rWRA}} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{W(x, z)W(z, y)}{s(z)} \quad (10)$$

上述 3 种指标利用了可靠路径的计算思路,改

进了 WCN、WAA、WRA 指标, 并且在没有提高计算复杂度的前提下, 对可靠路线赋予了更高的权重, 从而提高了链路预测的准确性。

2.4 基于行为同步指数的相似性指标

在很多真实社会网络中, 通过两个节点多次互动形成的权重是普遍存在的, 并且这种互动行为往往具有一定的时间特性, 该时间特性会通过两个节点的链接而产生^[22], 因此可利用这两个节点时间上的同步性来推断它们之间的潜在链接关系。

从链路预测的角度分析, 网络中节点之间的互动存在时间上的延迟, 这种时间特性可以作为一种同步现象来衡量节点的行为同步性。如果节点之间存在链接, 它们之间的交互时间间隔越短, 则这两个节点的行为将越趋于同步, 以此可推断他们未来存在链接的概率就越大。如节点 x 给节点 y 发消息, 节点 y 每次都立即回复, 说明两个节点产生链接的可能性极大。如果节点 x 给节点 z 发消息, 节点 z 经常不回复或是时间间隔很长才回复, 那么这两个节点产生链接的可能性就较小。因此, 在不知道两个节点之间是否存在链接的情况下, 观察两个节点之间的行为同步性可以推断他们之间是否存在联系。

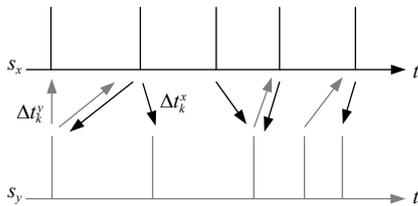


图1 用户同步指数的计算过程图

用户同步指数的计算过程如图1所示, 取网络中的一条连边 (x, y) , 将节点 x 与其他节点的信息交互时间列于一条时间轴上, S_x 为节点 x 的时间序列; 将节点 y 与其他节点信息交互时间列于一条时间轴上, S_y 为节点 y 的时间序列。 Δt_k^x 是在时间序列 S_x 上第 k 个信息交互事件到 S_y 上最近的信息交互事件的时间距离。 Δt_k^y 是在时间序列 S_y 上第 k 个信息交互事件到 S_x 上最近的信息交互事件的时间距离。时间序列 S_x 和 S_y 的相似性表示为:

$$\theta_{xy} = -\frac{D_{xy} + D_{yx}}{2} \quad (11)$$

$$D_{xy} = \frac{1}{N_x} \sum_k \Delta t_k^x \quad k \in O_x \quad (12)$$

$$D_{yx} = \frac{1}{N_y} \sum_k \Delta t_k^y \quad k \in O_y \quad (13)$$

式中, θ_{xy} 是时间序列 S_x 和 S_y 的平均距离; D_{xy} 是节点 x 到节点 y 的时间距离; D_{yx} 是节点 y 到节点 x 的时间距离; O_x 是节点 x 邻居节点的集合; O_y 是节点 y 邻居节点的集合; N_x 是节点 x 与其他节点信息交互的次数; N_y 是节点 y 与其他节点信息交互的次数。本文将平均距离作为衡量两个节点时间序列相似性的标准并应用于链路预测, 平均距离越短则时间序列相似性越强, 产生连边的可能性越大^[30]。式中的 D_{xy} 和 D_{yx} 不仅可以取所有距离的平均值, 也可以取所有距离的最小值, 此时距离定义为:

$$\theta_{xy} = -\frac{D_{xy} + D_{yx}}{2} \quad (14)$$

$$D_{xy} = \min\{\Delta t_k^x\} \quad k \in O_x \quad (15)$$

$$D_{yx} = \min\{\Delta t_k^y\} \quad k \in O_y \quad (16)$$

2.5 融合结构和行为同步指数的相似性指标

本文提出的融合结构和行为同步指数指标是同时计算两个节点之间的共同邻居信息和行为同步指数。事件交互对应消息的发送或接收, 因此定义节点发送消息的融合行为同步指数指标 STCN, 其中 S 表示发送, T 表示时间序列, CN 为共同邻居数量。定义接受消息的融合行为同步指数指标 RTCN, 其中 R 表示接收, 则 STCN 和 RTCN 的具体公式为:

$$\text{STCN} = \text{CN} + \theta_s \quad (17)$$

$$\text{RTCN} = \text{CN} + \theta_r \quad (18)$$

式中, CN 为两节点的共同邻居数; 由于事件交互对应消息的发送或接收, 所以 STCN 中, θ_s 是时间序列 S_x 与时间序列 S_y 发送消息的时间序列相似性; RTCN 中, θ_r 是时间序列 S_x 与时间序列 S_y 接收消息的时间序列相似性。融合行为同步指数的相似性指标包含两部分, 一部分是以 CN 为主的结构信息, 一部分是以行为时效信息为主的同步指数。

为了探究局域结构和同步指数对链路预测的共同关系, 本文使用一个可变参数 α 来调整两者的融合比率, 并分析本文所提方法的鲁棒性和两类指标的相关关系, 具体公式为:

$$\alpha \text{STCN} = (1 - \alpha) \text{CN} + \alpha \theta_s \quad (19)$$

$$\alpha \text{RTCN} = (1 - \alpha) \text{CN} + \alpha \theta_r \quad (20)$$

式中, α 取值范围为 $[0,1]$ 。当 $\alpha=0$ 时, 此时公式中只有结构信息存在; $\alpha=1$ 时, 公式中只有用户同步性存在; 在两者之间时, 公式中既有结构信息存在, 也有行为同步性存在, 并且在 $0\sim 1$ 之间可变参数的取值可以改变两部分信息在预测中的比例。

3 实证分析

3.1 数据说明

本文使用了两类实证数据, 一类是 3 种科学家合作网络数据^[31], 其中节点表示从事研究的科学家, 连边代表两个科学家有合作研究的关系, 连边的权重代表科学家合作的次数, 连边时间代表科学家之间的合作时间。另一类是由运营商提供的某城市某月的 3 种短信数据, 其中节点代表用户, 连边代表用户之间的短信往来, 连边的权重代表用户之间短信网络的次数, 连边的时间代表用户之间短信往来的时间。为了保护隐私, 所有的电话号码被隐藏, 没有任何信息能表明用户身份。

1) JIS 数据网络是以两本期刊 *Journal of Informetrics*、*Scientometrics* 作为研究对象, 从 *Web of Science* 数据库下载了 2013–2017 年的文章记录。包括 *Article*、*Review*、和 *Proceeding paper* 这 3 种类型。该网络包含 3 540 个节点, 6 413 条连边。

2) BMJ 网络是以期刊《英国医学期刊》(BMJ) 作为研究对象, 从 *Web of Science* 数据库下载了 2013–2018 年的所有论文。该网络包含 4 389 个节点, 4 335 条连边。

3) Nature 网络是以期刊《自然》(Nature) 作为研究对象, 从 *Web of Science* 数据库下载了 2010–2018 年的所有论文。该网络包含 21 060 个节点, 21 259 条连边。

4) SD01 网络: 包含某运营商一个月内 44 430 个用户节点, 和 68 834 条短信记录。

5) SD02 网络: 包含某运营商一个月内 72 146 个用户节点, 100 974 条短信记录。

6) SD03 短信网络: 包含某运营商一个月内 14 433 个用户节点, 23 285 条短信记录。

3.2 预测结果分析

本文使用上述两类实证网络, 每一类有 3 种真实数据集。每个网络按时间截点划分数据集, 训练集与测试集比例为 9 : 1。从不存在的边中抽取负样本, 让测试集中正负样本的比例为 1 : 1。

表 1 为 6 种网络中, 融合行为同步指数指标与基于局部信息相似性指标、加权局部信息相似性指标、可靠路线相似性指标的 Precision 结果比较。

表 1 6 种实证网络链路预测 Precision 值比较

指标	Precision					
	JIS	BMJ	Nature	SD01	SD02	SD03
CN	0.519	0.534	0.496	0.482	0.488	0.493
AA	0.553	0.534	0.500	0.483	0.491	0.482
RA	0.551	0.527	0.496	0.481	0.493	0.480
WCN	0.520	0.527	0.497	0.478	0.491	0.492
WAA	0.551	0.526	0.497	0.486	0.487	0.482
WRA	0.553	0.531	0.502	0.484	0.488	0.496
rWCN	0.551	0.532	0.497	0.481	0.491	0.493
rWAA	0.553	0.536	0.498	0.479	0.490	0.498
rWRA	0.552	0.532	0.498	0.477	0.490	0.482
STCN (mean)	0.692	0.584	0.649	0.618	0.686	0.646
RTCN (mean)	0.654	0.560	0.522	0.715	0.691	0.811
STCN (min)	0.576	0.618	0.647	0.687	0.660	0.625
RTCN (min)	0.655	0.554	0.538	0.683	0.688	0.808

可以看出, 本文提出的融合行为同步指数的链路预测方法的预测精度最好, 其预测准确率比共同邻居、加权共同邻居、可靠路线策 Precision 指标提高了 15.3%~68.2%。

计算融合行为同步指数指标时分别给出了行为同步指数取最小值及平均值的实验结果。每种数据中最高的预测性能加粗表示。表 1 中可发现行为同步指数总是取得最佳的性能, 虽然取最小值及平均值在不同网络中预测结果有些不同, 在 BMJ 网络中, 行为同步指数取最小值预测更准确。而在 SD01、SD02、SD03、Nature 和 JIS 网络中, 取平均值预测性能更好。

除了行为同步指数的取值影响算法精度外, 时间序列的方向(如发送和接收)同样也影响着预测精度。有些网络利用发送消息计算同步指数能取得更好的预测结果, 有些网络中用接收消息计算同步指数能取得更好的预测结果。表 1 中, SD01、SD02、SD03 网络在行为同步指数取平均值的情况下, 接受消息的 RTCN (mean) 的预测结果更高。而在 Nature、JIS 和 BMJ 科学家合作网络的预测中, 发送消息的 STCN (mean) 预测效果更好。在短信网络中接收消息更能有效挖掘网络节点之间的预测关系。而在科学家合作网络中, 发送消息更能有效挖掘网络节点之间的预测关系。

基于局部结构共同邻居数量演化出来的一系

列相似性指标, 预测结构没有明显变化, 因此它们相互之间的性能波动较小。而当局部结构信息加入了行为同步指数之后, 预测结果得到了大幅提升, 这说明这两类信息具有很强的互补性。如在 SD03 网络中, 基于局部结构特征的 Precision 值最高为 0.498, 加入行为同步指数之后, Precision 值为 0.811, 相对提高了 62.8%。因此, 节点的行为同步指数不仅能用于网络重构中, 还可以应用在链路预测上。无论是在科学家合作网还是在短信交互网中, 生成网络连边权重的时效信息都是不能忽略的重要特征, 由时效信息演变的同步效果都是普遍存在的, 融合同步指数的链路预测方法性能更具有优势。

3.3 融合比例变化对预测结果的影响

实验中利用可变参数 α 改变网络结构信息和时效信息的融合 α 值, 同时观察该值对预测性能的影响。两类网络中各选一种, 以科学家合作网 Nature 和短信网络 SD03 为例进行分析, 每一类中其他网络的结果与本文呈现的该类网络结果是相似的。

图 2a 是 Nature 网络特征融合比例变化的性能曲线图。纵轴是预测结果的 Precision 值, 横轴是以 0.1 为跨度的 α 值。预测结果显示, 不管是 STCN 方法还是 RTCN 方法, α 取 0 是单一结构信息的预测结果, 此时并不能达到一个较好的预测性能。 $\alpha = 1$ 时是单一同步指数的预测结果, 而预测效果却是最不理想的, 甚至要低于单一结构特征预测。尤其是接收消息的 RTCN 特征, 比单一结构信息的 Precision 值下降了 48.9%。当 $0 < \alpha < 1$ 时, 无论 α 取何值都能取得较高的预测性能, 说明本文的方法具有很好的鲁棒性。综上所述, 对于 Nature 等科学家合作网络来说, 科学家的共同邻居数量与行为同步性共同影响着合作网络中链路的出现和动态演化。

图 2b 是 SD03 短信交互网络特征融合比例变化的性能曲线图。纵轴是预测结果的 Precision 值, 横轴是以 0.1 为跨度的 α 值。在 4 种特征的预测下, 预测精度最低时对应的 α 等于 0, 此时是结构信息的单一预测结果, 说明短信交互网络中只用结构信息不足以挖掘网络的潜在链接信息。预测精度最高时对应 $0 < \alpha < 1$, 此时为结构信息和行为同步指数共同的预测结果。 α 在一定范围内取值, 预测性能都没有显著变化, 说明了本文所

提方法具有很好的鲁棒性。当 $\alpha = 1$ 时, 图 2b 中没有出现预测值明显下跌的情况, 结果与 $0 < \alpha < 1$ 时的 Precision 值相差无几, 此时是行为同步指数单一预测结果。综上所述, SD03 网络用单一同步指数预测也会出现很好的精度, 用户之间的行为同步性可以在不依赖结构信息的情况下, 实现单一特征预测并且达到较好的预测精度。

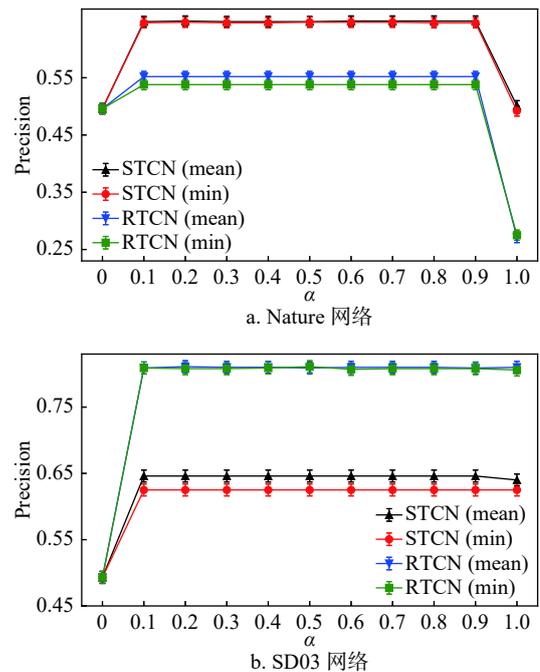


图2 特征融合比例对预测性能的影响

考虑到科学家合作网络和短信网络的形成机制不同, 科学家合作网络因为 n 个合作者连成完全图, 所以共同邻居指数很重要。这里假设同步性在合作网络中起的作用实际上只能在 CN 值相同的时候提高区分度, 因此将科学家合作网和短信网络的预测参数 α 值调节到一个非常小的值 ($\alpha=0.001$), 再调节到一个非常接近 1 的值 ($\alpha=0.999$), 实验结果如图 3 所示。

图 3 是 Nature 期刊科学家合作网络在 $\alpha=0.001$ 和 $\alpha=0.999$ 的预测结果, 并将该结果与 α 取 0 和 1 的预测结果做对比。在科学家合作网络之中, $\alpha=0.001$ 就可以达到一个提升预测性能的作用, 并且与 $\alpha=0.999$ 并无差别。说明此时同步性指标的作用类似于局域性指标里面的高阶项, 因此预测的过程中以 CN 为主, 同步性可以辅助 CN 达到提高预测分辨率的效果。

图 4 是短信交互网络 SD03 在 $\alpha=0.001$ 和 $\alpha=0.999$

的预测结果，并将该结果与 α 取 0 和 1 的预测结果做对比。短信网络的实验结果中可以看出 CN 指标作用很小，而同步性无论在 $\alpha=0.001$ 还是 $\alpha=0.999$ 都可以达到良好的预测精度。本文统计了 6 种网络的平均聚类系数，结果如表 2 所示。

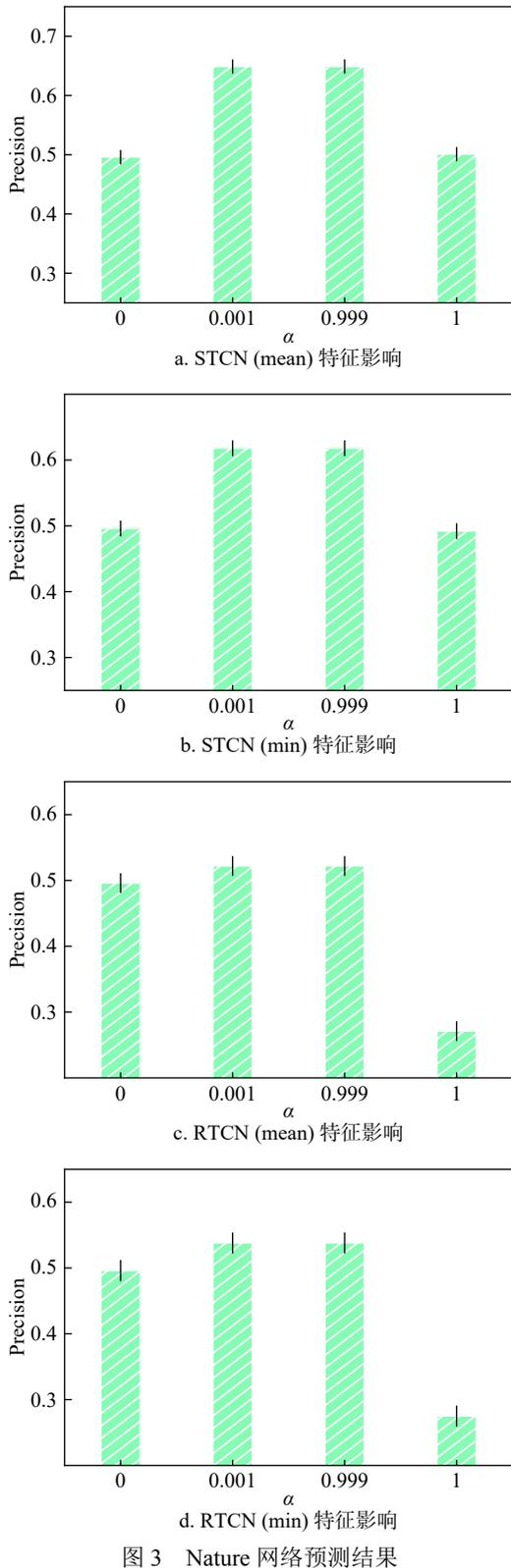


图 3 Nature 网络预测结果

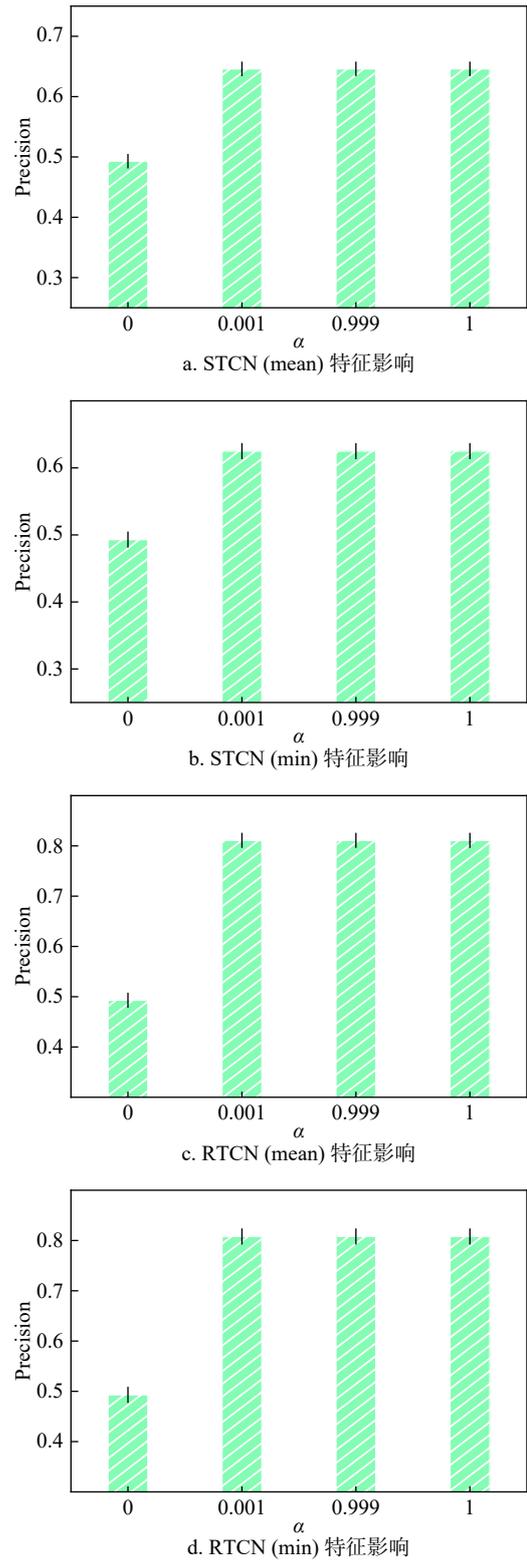


图 4 SD03 网络预测结果

从表 2 中可以看出，3 个短信网络的平均聚类系数都很小，说明短信网络的结构都十分松散，因此预测过程中邻域结构信息的价值很小，这也导致了 $\alpha=1$ 时图 2b 中 SD03 网络没有出现预测值明显

下跌的情况。而合作网络的平均聚类系数相对较大, 也说明了科学家合作网络的共同朋友结构信息明显, 对 CN 的依赖性较大, 同步信息仅起到提高更好分辨率的作用。

表 2 6 种实证网络平均聚类系数

网络类型	网络名称	平均聚类系数
短信网络	SD01	0.045
	SD02	0.042
	SD03	0.059
	JIS	0.398
科学家合作网	BMJ	0.252
	Nature	0.211

4 结束语

本文将用户行为同步性信息这一网络重构的方法引入到链路预测领域, 提出一种网络拓扑相似性上融合节点行为同步指数的链路预测方法。通过对该方法在真实社会时效网络上与局部结构相似性指标、加权局部结构相似性指标、可靠路线相似性指标进行对比分析, 发现该方法可有效提高网络的链路预测性能。相对于经典方法, 本文方法在 Precision 指标上提高了 15.3%~68.2%。

本文方法不仅可提高链路预测的性能, 也可揭示加权网络的内在结构特征和行为同步指数对链路预测的相互作用与共同影响。研究发现, 在科学家合作网络中, 需要结构信息和同步指数共同作用, 才会得到良好的预测效果。而在短消息交互网络中, 结构相似性指标的预测性能较差, 而仅仅使用行为同步指数就可以接近最优的预测性能。在以后的工作中, 将采取更优的方式将节点行为同步指标融合与网络结构相似特性进行融合, 并且对结构信息的时效性展开更深入研究, 以达到更优的链路预测性能。同时, 也将对短信网络和科学家合作网络的细节机制做更深入分析, 包括发送短信的模式、群发和短时间频繁发送等, 根据其结构特点来设计链路预测方法。本文所提方法可以拓展到社团检测、信息传播等其他网络科学领域的实际应用中。

参 考 文 献

- [1] NEWMAN M E. The structure of scientific collaboration networks[J]. *Proceedings of the National Academy of Sciences*, 2001, 98(2): 404-409.
- [2] ALBERT R, JEONG H, BARABÁSI A L. Diameter of the world-wide web[J]. *Nature*, 1999, 401(6): 130-131.
- [3] BRANDO M A, MORO M M. Social professional networks: A survey and taxonomy[J]. *Computer Communications*, 2017, 100: 20-31.
- [4] 刘宏鲲, 周涛. 航空网络研究综述[J]. *自然科学进展*, 2008, 18(6): 601-608.
LIU Hong-kun, ZHOU Tao. Overview of aviation network research[J]. *Progress in Natural Science*, 2008, 18(6): 601-608.
- [5] 吕琳媛, 周涛. 链路预测[M]. 北京: 高等教育出版社, 2013.
LÜ Lin-yuan, ZHOU Tao. Link prediction[M]. Beijing: Higher Education Press, 2013.
- [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] SCELLATO S, NOULAS A, MASCOLO C. Exploiting place features in link prediction on location-based social networks[C]//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [S.l.]: ACM, 2011: 1046-1054.
- [8] YÜ H, PASCAL B, MUHAMMED A Y, et al. High-quality binary protein interaction map of the yeast interactome network[J]. *Science*, 2008, 322(5898): 104-110.
- [9] LIBEN-NOWELL D, KLEINBERG J. The link-prediction problem for social networks[J]. *Journal of the American Society for Information ence and Technology*, 2007, 58(7): 1019-1031.
- [10] ADAMIC L A, ADAR E. Friends and neighbors on the Web[J]. *Social Networks*, 2003, 25(3): 211-230.
- [11] ZHOU T, LÜ L, ZHANG Y C. Predicting missing links via local information[J]. *The European Physical Journal B*, 2009, 71(4): 623-630.
- [12] 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(4): 046122.
- [13] LIU W, LÜ L. Link prediction based on local random walk[J]. *EPL*, 2010, 89(5): 58007.
- [14] LÜ L, ZHOU T. Link prediction in weighted networks: The role of weak ties[J]. *EPL*, 2010, 89(1): 18001.
- [15] ZHAO J, MIAO L, YANG J. Prediction of links and weights in networks by reliable routes[J]. *Scientific Reports*, 2015, 5(1): 12261.
- [16] YAO Y, ZHANG R, YANG F, et al. Link prediction based on local weighted paths for complex networks[J]. *International Journal of Modern Physics C*, 2017, 28(4): 1750053.
- [17] DEVI S J, SINGH B. Analysis of link prediction in directed and weighted social network structure[C]//The International Symposium on Intelligent Systems Technologies and Applications. [S.l.]: Springer, 2017: 1-13.
- [18] 王凯, 刘树新, 于洪涛, 等. 基于共同邻居有效性的复杂网络链路预测算法[J]. *电子科技大学学报*, 2019, 48(3): 432-439.
WANG Kai, LIU Shu-xin, YÜ Hong-tao, et al. Predicting missing links of complex network via effective common neighbors[J]. *Journal of University of Electronic Science and Technology of China*, 2019, 48(3): 432-439.

- [19] CAO Z W, ZHANG Y C, GUAN J H, et al. Link weight prediction using weight perturbation and latent factor[J]. *IEEE Transactions on Cybernetics*, 2020(99): 1-13.
- [20] 郭婷婷, 赵承业. 异常链路分析在电力网络恢复中的应用[J]. *电子科技大学学报*, 2016, 45(5): 854-859.
GUO Ting-ting, ZHAO Cheng-ye. Application of abnormal links analysis in restoring power networks[J]. *Journal of University of Electronic Science and Technology of China*, 2016, 45(5): 854-859.
- [21] 唐雪飞, 杨陈皓, 牛新征, 等. 复杂网络链路危险度预测模型研究[J]. *电子科技大学学报*, 2013, 42(3): 123-128.
TANG Xue-fei, YANG Chen-hao, NIU Xin-zheng, et al. Research on prediction of complex networks link danger level[J]. *Journal of University of Electronic Science and Technology of China*, 2013, 42(3): 123-128.
- [22] 周涛, 韩筱璞, 闫小勇, 等. 人类行为时空特性的统计力学[J]. *电子科技大学学报*, 2013, 42(4): 481-540.
ZHOU Tao, HAN Xiao-pu, YAN Xiao-yong, et al. Statistical mechanics on temporal and spatial activities of human[J]. *Journal of University of Electronic Science and Technology of China*, 2013, 42(4): 481-540.
- [23] CODRONS E, BERNARDI F N, VANDON M, et al. Spontaneous group synchronization of movements and respiratory rhythms[J]. *PLoS ONE*, 2014, 9(9): e107538.
- [24] FU C, WANG X. Network growth under the constraint of synchronization stability[J]. *Physical Review E*, 2011, 83(6): 066101.
- [25] FRANK T D, RICHARDSON M J. On a test statistic for the Kuramoto order parameter of synchronization: An illustration for group synchronization during rocking chairs[J]. *Phys D Nonlin*, 2010, 239(23): 2048-2092.
- [26] NAGY M, ÁKOS Z, BIRO D. Hierarchical group dynamics in pigeon flocks[J]. *Nature*, 2010, 464(7290): 890-893.
- [27] LIAO H, LIU M, MARIANI M S. Temporal similarity metrics for latent network reconstruction: The role of time-lag decay[J]. *Information Sciences*, 2019: 182-192.
- [28] PANAGGIO M J, CIOCANEL M V, LAZARUS L, et al. Model reconstruction from temporal data for coupled oscillator networks[J]. *Chaos*, 2019, 29(10): 103116.
- [29] SURANO F V, BONGIORNO C, ZINO L, et al. Backbone reconstruction in temporal networks from epidemic data[J]. *Physical Review E*, 2019, 100(4): 042306.
- [30] FELDT S, WADDELL J, HETRICK V L, et al. Functional clustering algorithm for the analysis of dynamic network data[J]. *Physical Review E, Statistical, Nonlinear, and Soft Matter Physics*, 2009, 79(5): 56104.
- [31] NWEAMAN M E. The structure and function of complex networks[J]. *SIAM Review*, 2003, 45(2): 167-256.

编辑 蒋晓