

# 一种基于博弈论的时序网络链路预测方法

刘 留<sup>1</sup> 王煜尧<sup>2</sup> 倪琦瑄<sup>1</sup> 曹 杰<sup>2</sup> 卜 湛<sup>1</sup>

<sup>1</sup>(南京财经大学信息工程学院 南京 210013)

<sup>2</sup>(南京理工大学计算机科学与工程学院 南京 210094)

(545108883@qq.com)

## A Link Prediction Approach in Temporal Networks Based on Game Theory

Liu Liu<sup>1</sup>, Wang Yuyao<sup>2</sup>, Ni Qixuan<sup>1</sup>, Cao Jie<sup>2</sup>, and Bu Zhan<sup>1</sup>

<sup>1</sup>(College of Information Engineering, Nanjing University of Finance and Economics, Nanjing 210013)

<sup>2</sup>(School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094)

**Abstract** Link prediction is an important task in complex network analysis, which can be applied to many real-world practical scenarios such as recommender systems, information retrieval, and marketing analysis. Different from the traditional link prediction problem, this paper predicts the existence of the link at any time in the future based on the set of temporal links in a given time window, that is, the evolution mechanism of the temporal network. To explore this question, we propose a novel semi-supervised learning framework, which integrates both survival analysis and game theory. First, we carefully define the  $\epsilon$ -adjacent network sequence, and make use of time stamp on each link to generate the ground-truth network evolution sequence. Next, to capture the law of network evolution, we employ the Cox proportional hazard model to study the relative hazard associated with each temporal link, so as to estimate the covariate's coefficient associated with a set of neighborhood-based proximity features. To compress the searching space, we further propose a game theory based two-way selection mechanism to inference the future network topology. We finally propose a network evolution prediction algorithm based on autonomy-oriented computing, and demonstrate both the effectiveness and the efficiency of the proposed algorithm on real-world temporal networks.

**Key words** link prediction; temporal network; survival analysis; game theory; autonomy oriented computing

**摘 要** 链路预测是复杂网络分析领域的一项重要研究课题,可被应用于许多实际应用场景,如推荐系统、信息检索和市场分析等。不同于传统的链路预测问题,针对有时间窗口的时序链路集合,需预测未来任意时刻链路的存 在情况,即探究时序网络的演化机制。为解决这一问题,结合生存分析和博弈论,提出一种有效的半监督学习框架。首先,定义一个  $\epsilon$ -邻接网络序列模型,并利用每条链路的时间戳信息生成真实的网络演化序列。为捕捉网络演化规律,为每条链路定义一组基于邻居相似性的特征向量,并采用 Cox 比例风险模型来估计该特征向量的协变量系数。为缩小搜索空间,提出一种基于博弈的双向选择机制来预测未来的网络拓扑结构。最后,提出一种基于多智能体自治计算的 网络演化预测算法,并在多个真实时序网络数据集上验证了算法的有效性和高效性。

收稿日期:2018-12-24;修回日期:2019-04-02

基金项目:国家自然科学基金项目(71871109,91646204,71801123,71871233)

This work was supported by the National Natural Science Foundation of China (71871109, 91646204, 71801123, 71871233).

通信作者:卜湛(buzhan@nuaa.edu.cn)

**关键词** 链路预测; 时序网络; 生存分析; 博弈论; 自治计算

**中图分类号** TP393

许多现实世界的网络系统是非静态的, 节点间的连接关系随时间动态变化, 此类网络可由一系列包含时间戳的链路集合<sup>[1]</sup>表示. 在动态网络研究领域中, 除了研究最为广泛的社区发现问题<sup>[2-3]</sup>, 链路预测问题<sup>[4-6]</sup>研究如何利用网络历史拓扑信息预测未来的链路存在情况, 是一项具有重要研究意义的课题. 链路预测对于人们理解网络演化<sup>[7]</sup>规律至关重要, 且在诸多实际应用中发挥着重要作用, 如预测蛋白质间的相互作用<sup>[8]</sup>、社交网络中的朋友推荐<sup>[9]</sup>和科研网络的合作推荐<sup>[10]</sup>等.

区别于已有的链路预测研究, 本文旨在研究时序网络<sup>[11]</sup>的演化机制, 即给定一组由固定节点集合构成的静态网络序列, 如  $g^1, g^2, \dots, g^l$ , 我们尝试预测该网络在未来时刻  $t+1$  到  $t+T$  的拓扑结构. 一方面, 传统的链路预测方法<sup>[12]</sup>多关注于静态网络的链路预测问题, 忽略了每条链路上的时间戳信息, 故不能被直接应用于解决时序网络的链路预测问题. 另一方面, 现实世界的网络系统通常具有较大的规模, 包含海量有复杂联系的网络参与者. 例如, 一个包含  $10^5$  个节点和  $10^6$  条链路的稀疏网络, 若预测未来可能生成的新链路, 其搜索空间的数量级为  $10^{10}$ ; 而预测未来可能消失链路的搜索空间的数量级为  $10^6$ , 仅占前者的 0.01%. 受限于上述真实网络的这种稀疏性特性, 有监督的链路预测方法会面临不同程度的分类样本分布失衡问题.

基于上述分析, 本文构建了一个高效的链路预测框架. 首先, 采用  $\epsilon$ -邻接网络序列定义时序网络的表示模型, 从包含时间戳的链路集合中生成真实的网络演化序列. 另外, 由于生存分析中的 Cox 比例风险模型不需对生存时间的分布作出假定, 仅通过一个模型就可分析生存时间的分布规律. 因此, 我们为每条链路关联一组基于拓扑相似性的特征向量, 并借助 Cox 比例风险模型分别学习链路的消失和重连概率. 然而, 现实中的大多数网络为稀疏网络, 即大多数的节点对之间不存在链路. 因此, 为压缩搜索空间, 我们将节点对之间的连边情况模拟为 2 个玩家之间的博弈, 进而提出一种基于动态博弈的双向选择机制来预测未来的网络拓扑结构. 除此之外, 由于计算特征向量所采用的邻近性指标是基于每个节点的邻居来计算得到的, 故我们将可能与节点  $i$  发生重连的节点约束在距  $i$  的最短路径距离不超过 2

的范围内. 此举有效地压缩了链路预测的搜索空间, 提高了算法的计算效率, 缓解了稀疏网络带来的分类样本分布失衡问题.

## 1 相关工作

传统的复杂网络链路预测方法主要分为 3 类: 1) 无监督预测方法; 2) 有监督预测方法; 3) 基于概率模型的预测方法.

无监督预测方法, 也被称为评分方法<sup>[13-14]</sup>. 此类方法的基本假设是: 若 2 个节点拓扑相似度越大, 它们在未来形成链路的概率越高. 基于此假设, 无监督预测方法通过定义一系列节点对的相似性函数, 来刻画网络中任一链路的存在概率. 其中最常用的相似性函数包括 CN (common neighbor index)<sup>[15]</sup>, JC (Jaccard coefficient)<sup>[16]</sup>, PA (preferential attachment index)<sup>[17]</sup>, AA (adamic-adar index)<sup>[18]</sup> 等. 由于这些相似性函数多采用网络的局部拓扑信息, 故此类无监督预测方法通常具有较低的时间复杂度. 此外, 一些无监督预测方法则基于节点对连通路径特征来定义节点对的相似性函数. 此类方法的基本假设是: 若 2 个节点间的距离越短, 则它们在未来形成链路的概率越高. 例如, Katz<sup>[19]</sup> 不仅考虑 2 个节点之间所有的路径, 同时还利用路径的权重来计算它们之间的相似性. Lü 等人<sup>[20]</sup> 提出的局部路径 (local path, LP) 指标同时考虑直接邻居和间接邻居, 获取的路径信息相对更为稳定、可靠. Yang 等人<sup>[21]</sup> 提出了一种基于聚类和决策树的链路预测方法. 基于随机游走的链路预测算法的基本假设是: 若某个节点进行随机游走, 到达另外一个节点所需要的平均游走步数最少, 此时 2 个节点的相似性最高. Lichtenwalter 等人<sup>[22]</sup> 在随机游走的基础上进行了一定的改进, 提出了 PropFlow 方法. Tong 等人<sup>[23]</sup> 提出了一种新的随机游走方法, 充分考虑网络中普遍存在的社区结构特性. Jin 等人<sup>[24]</sup> 提出一种基于路径反转影响和随机游走矩阵的链路预测方法. 与经典的基于邻居相似度的无监督预测方法相比, 基于节点对连通路径特征的无监督预测方法考虑了网络的全局拓扑特征, 其预测准确率相对较高, 但节点对路径特征的计算过程具有较高的时间复杂度.

有监督预测方法将链路预测问题转化为二进制

分类问题,根据历史的网络演化序列构建训练样本,并借助现有的分类器,如逻辑回归<sup>[25]</sup>、支持向量机<sup>[26]</sup>、决策树<sup>[27]</sup>、多层感知器神经网络<sup>[28]</sup>和朴素贝叶斯<sup>[29]</sup>等,训练分类器参数,进而再利用训练好的分类器进行链路预测.例如,Pujari 等人<sup>[30]</sup>提出一种监督式的排序融合方法对复杂网络进行链路预测.Chen 等人<sup>[31]</sup>利用正则化参数来完成贝叶斯推理,进而提高学习潜在特性的辨别能力.Nguyen-Thi 等人<sup>[32]</sup>提出一种基于径向基函数(radial basis function, RBF)的支持向量机(support vector machine, SVM)分类器,可有效提高预测精度且降低分类器的训练时间.基于监督学习的链路预测方法在构建训练样本时都不同程度地面临样本类别分布失衡问题.真实的网络系统呈现极强的稀疏特征,假定时刻  $t$  存在一个包含  $n$  个节点和  $m$  条链路的无向网络  $g^t$ ,其稀疏性表现为  $m \ll n(n-1)/2$ ,其中  $n(n-1)/2$  表示  $g^t$  网络最大可能的无向链路数量.当  $g^t$  演化至时刻  $t'$ ,我们可以根据网络  $g^{t'}$  与网络  $g^t$  的拓扑差异,分别构建相应的训练集来预测未来的消失链路和重连链路.以重连链路预测为例,我们可以将  $\{e_{ij} | e_{ij} \in (g^{t'} - g^t)\}$  中的链路标记为正例样本,将  $\{e_{ij} | e_{ij} \notin (g^{t'} - g^t)\}$  中的链路标记为负例样本.考虑到真实网络的动态演化呈现显著的渐进性,网络拓扑结构的变化是一个相对缓慢的过程.由于  $|\{e_{ij} | e_{ij} \in (g^{t'} - g^t)\}| \ll |\{e_{ij} | e_{ij} \notin (g^{t'} - g^t)\}|$ ,采用上述方法构建的训练集中,正例样本和负例样本存在严重的分布失衡现象,传统的机器学习分类器难以训练出理想的重连链路预测模型.

基于概率模型的链路预测的基本思想是建立一个含有多个可调参数的模型,利用某种优化策略寻找参数的最优值,从而使得网络的结构和关系特征能够在模型中得到更好的体现.网络中任意节点对的重连概率可表示为基于最优参数配置的条件概率.Wang 等人<sup>[33]</sup>提出了一个基于局部概率模型(local probabilistic model, LPM)的链路预测方法,可极大地降低大规模复杂网络链路预测问题的时间复杂度和空间复杂度.Coutant 等人<sup>[34]</sup>提出了一种基于聚类不确定性的概率关系模型(probabilistic relational models with clustering uncertainty, PRM-CU)用于分析复杂网络的级联失效,该概率模型能够充分利用网络的拓扑信息,衡量节点对的交互能力,能够有效地提高链路预测的准确性.由于网络结构本身较为复杂,为表示更高层次的非线性网络结构,Wang 等人<sup>[35]</sup>提出了结构化深层网络嵌入模型

(structural deep network embedding, SDNE).与传统链路预测方法相比,SDNE 模型能够有效地提取网络的局部和全局结构信息,重构网络结构,进而得到更高的预测精度.

此外,还有一些链路预测方法考虑了链路的时间戳信息和网络的社团结构特征.例如,Sharan 等人<sup>[36]</sup>提出了基于不同时间关系的分类模型,采用 2 个实体间动态关系信息的关系模型.Rossi 等人<sup>[37]</sup>提出了基于时序关系从属分类的集成预测方法.Clauset 等人<sup>[38]</sup>提出了一种利用网络层次结构进行链路预测的方法,且在具有明显层次结构的网络中,该方法能取得更好的预测性能.这些链路预测方法大多对网络的拓扑结构有较高的要求,且计算时间复杂度普遍较高.

## 2 问题定义

现实世界的时序网络可表示为一个时序链路集  $\gamma = \{\langle i_p, j_p, \tau_p \rangle | p = 1, 2, \dots, M\}$ ,其中  $\langle i_p, j_p, \tau_p \rangle$  表示一个带有时间戳的时序链路,表明节点  $i_p$  和节点  $j_p$  在时刻  $\tau_p$  是连通的.当忽略时序链路的时间戳和重复链路时,时序网络就退化为一个静态网络  $g(\gamma) = \{e_{ij} | \exists \langle i_p, j_p, \tau_p \rangle \in \gamma\}$ .

**定义 1.**  $\delta$ -时序范式.时刻  $t$  的  $\delta$ -时序范式是一个有序链路序列  $M^t(\delta) = \langle i_{t_s}, j_{t_s}, \tau_{t_s} \rangle, \dots, \langle i_{t_e}, j_{t_e}, \tau_{t_e} \rangle$ ,满足  $(\tau_{t_s} \leq \dots \leq \tau_{t_e}) \wedge (\delta = \tau_{t_e} - \tau_{t_s})$ ,且静态网络  $g^t(\delta) = (V^t(\delta), E^t(\delta))$ ,  $E^t(\delta) \subseteq V^t(\delta) \times V^t(\delta)$  是连通的.

**定义 2.**  $\epsilon$ -邻接.给定 2 个长度为  $\delta$  的重叠时间窗口  $[t_s, t_e]$  和  $[t'_s, t'_e]$ ,满足  $\tau_{t_e} - \tau_{t_s} = \tau_{t'_e} - \tau_{t'_s} = \delta \wedge (t_s \leq t'_s \leq t_e \leq t'_e)$ .窗口滑动率定义为  $\epsilon = (t'_s - t_s)/\delta$ ,此时称静态网络  $g^t$  与  $g^{t'}$  是  $\epsilon$ -邻接的.

静态网络  $g^t$  演化至与其  $\epsilon$ -邻接的网络  $g^{t'}$  包含 3 种可能情况:1)从  $g^t$  中消失的链路;2) $g^t$  中任意 2 个不连通的节点形成的链路;3)至少有 1 个外部节点参与形成的链路.由于外部信息具有较大的随机性,本文将重点关注前 2 种情况的链路预测问题,并假设出现在所有周期的节点集合是固定的.

**定义 3.**  $\epsilon$ -邻接网络序列.从初始网络  $g^0(\delta) = (V, E^0)$  开始, $\epsilon$ -邻接网络序列由一系列耦合的静态网络组成,即  $G^\epsilon(g^0(\delta)) = \{g^1, g^2, \dots\}$ ,满足  $\forall t \geq 1$ :1) $g^{t-1}$  和  $g^t$  是  $\epsilon$ -邻接的;2) $V^t \subseteq V, E^t \subseteq V \times V$ .

为强调从  $g^{t-1}$  到  $g^t$  的网络演化过程,令  $M^t$  和  $R^t$  分别表示从  $g^{t-1}$  消失的链路集合和在  $g^t$  中新

生成的链路集合. 显然,  $M^t = E^{t-1} - E^t$  和  $R^t = E^t - E^{t-1}$ .

**定义 4.** 多层网络<sup>[39]</sup> 序列. 令  $\epsilon$ -邻接网络序列  $G^\epsilon(g^0(\delta)) = \{E^1, E^2, \dots\}$  中任意第  $t$  个周期的静态网络都对应一个多层网络  $M^t(L) = \{E^t, P^t(L)\}$ , 其中,  $E^t \subseteq V \times V$  是周期  $t$  中静态网络的链路集,  $P^t(L) \subseteq V \times V$  是  $g^t$  最近  $L$  个静态网络投影所得的链路集:

$$P^t(L) = \begin{cases} \bigcup_{l=0}^t E^l, & 0 \leq t \leq L-1; \\ \bigcup_{l=1}^L E^{t-l+1}, & t > L-1. \end{cases} \quad (1)$$

这样, 由任意周期窗口  $L$  导出的多层网络序列表示为  $MNS(L) = \{M^0(L), M^1(L), \dots\}$ .

**定义 5.** 拓扑相似性函数. 令  $N_i = \{j | e_{ij} \in E^*\}$  表示节点  $i$  在某个静态网络中的邻居集合, 链路  $\langle i_p, j_p, \tau_p \rangle$  的拓扑相似性可由一些相似性函数度量计算得到. 例如, CN<sup>[15]</sup>, JC<sup>[16]</sup>, Sa (salton index), So (sorensen index), HP (hub promoted index), HD (hub depressed index), LN (Leicht-Newman index), PA<sup>[17]</sup>, AA<sup>[18]</sup> 等, 具体为

$$\begin{aligned} f_p^{CN} &= |N_i \cap N_j|; \\ f_p^{JC} &= \frac{|N_i \cap N_j|}{|N_i \cup N_j|}; \\ f_p^{Sa} &= \frac{|N_i \cap N_j|}{\sqrt{|N_i| |N_j|}}; \\ f_p^{So} &= \frac{2|N_i \cap N_j|}{|N_i| + |N_j|}; \\ f_p^{HP} &= \frac{|N_i \cap N_j|}{\min(|N_i|, |N_j|)}; \\ f_p^{HD} &= \frac{|N_i \cap N_j|}{\max(|N_i|, |N_j|)}; \\ f_p^{LN} &= \frac{|N_i \cap N_j|}{|N_i| |N_j|}; \\ f_p^{PA} &= |N_i| |N_j|; \\ f_p^{AA} &= \sum_{k \in N_i \cap N_j} \frac{1}{|N_k|}. \end{aligned}$$

我们可以采用不同的相似性函数<sup>[11]</sup> 来计算得到时刻  $t$  静态网络中每条链路的特征向量  $\mathbf{F}_p^t = (f_p^1, f_p^2, \dots, f_p^n)^T \in \mathbb{R}^n$ , 其中  $n$  为相似性指标的个数. 然而对于真实的大规模网络, 节点对的数目较多, 相似性指标的选取将直接影响到算法的执行效率. 因此, 本文选取定义 5 中基于节点局部信息的 9 种经典的相似性函数来计算得到  $\mathbf{F}_p^t$ , 即  $\mathbf{F}_p^t = (f_p^{CN}, f_p^{JC}, f_p^{Sa}, f_p^{So}, f_p^{HP}, f_p^{HD}, f_p^{LN}, f_p^{PA}, f_p^{AA})^T \in \mathbb{R}^9$ .

**定义 6.** 消失链路序列 (missing link sequence, MLS). 给定一个  $\epsilon$ -邻接网络序列  $G^\epsilon(g^0(\delta))$ , 由此导出的消失链路序列表示为  $G^M(g^0(\delta)) = \{M^1, M^2, \dots\}$ , 满足  $\forall t \geq 1, M^t = E^{t-1} - E^t$ .

**定义 7.** 重连链路序列 (reconnected link sequence, RLS). 给定一个  $\epsilon$ -邻接网络序列  $G^\epsilon(g^0(\delta))$ , 由此导出的重连链路序列表示为  $G^R(g^0(\delta)) = \{R^1, R^2, \dots\}$ , 满足  $\forall t \geq 1, R^t = E^t - E^{t-1}$ .

**问题 1.** 时序网络链路预测. 给定一个由  $\delta$ -时序范式导出的初始静态网络, 时序网络链路预测旨在生成一个网络演化序列  $\tilde{G}(g^0(\delta)) = \{\tilde{E}^1, \tilde{E}^2, \dots, \tilde{E}^T\}$ , 满足  $\forall t \geq 1, \tilde{E}^t$  尽可能匹配  $\epsilon$ -邻接网络序列  $G^\epsilon(g^0(\delta)) = \{E^1, E^2, \dots\}$  中的某个静态网络:

$$\max_{\tilde{G}(g^0(\delta))} \left( \frac{1}{2P} \sum_{t \geq 1} \max_{\tau} F1(\tilde{E}^t, E^\tau) + \frac{1}{2T} \sum_{\tau \geq 1} \max_t F1(\tilde{E}^t, E^\tau) \right), \quad (2)$$

其中,  $F1(\tilde{E}^t, E^\tau)$  表示链路集  $\tilde{E}^t$  和  $E^\tau$  的  $F1$ -score, 满足

$$F1(\tilde{E}^t, E^\tau) = \frac{2P(\tilde{E}^t, E^\tau)R(\tilde{E}^t, E^\tau)}{P(\tilde{E}^t, E^\tau) + R(\tilde{E}^t, E^\tau)}, \quad (3)$$

$$P(\tilde{E}^t, E^\tau) = \frac{|\tilde{E}^t \cap E^\tau|}{|\tilde{E}^t|}, \quad (4)$$

$$R(\tilde{E}^t, E^\tau) = \frac{|\tilde{E}^t \cap E^\tau|}{|E^\tau|}. \quad (5)$$

### 3 时序网络链路预测模型

给定一个  $\epsilon$ -邻接网络序列  $G^\epsilon(g^0(\delta))$ , 本文拟利用  $G^M(g^0(\delta))$  和  $G^R(g^0(\delta))$  中的链路标签推理链路消失或重连的概率. 具体来说, 我们将借助 Cox 比例风险模型来学习特征向量  $\mathbf{F}_p^t$  的权重系数, 在此基础上计算给定链路的消失或重连概率, 再借助于动态博弈模型, 预测网络未来的拓扑结构.

**定义 8.** 消失链路生存数据模型 (SDGM-MLP). 给定一个多层网络序列  $MNS(L) = \{M^0(L), M^1(L), \dots\}$  和对应的消失链路序列  $G^M(g^0(\delta)) = \{M^1, M^2, \dots\}$ , 预测消失链路的生存数据定义为  $H^M = \{\langle T_p, \mathbf{F}_p^M \rangle | p \in [1, \sum_{t \geq 1} |M^t|]\}$ .  $H^M$  中的每个实例  $\langle T_p, \mathbf{F}_p^M \rangle$  表示  $G^M(g^0(\delta))$  中的一条消失链路, 其中  $T_p$  表示节点对  $(i, j)_p$  保持连通的持续时间,  $\mathbf{F}_p^M$  表示节点对  $(i, j)_p$  相对于静态网络  $E^{t-T_p}$  的特征向量.

**定义 9.** 重连链路生存数据模型 (SDGM-RLP).

给定一个多层网络序列  $MNS(L) = \{M^0(L), M^1(L), \dots\}$  和对应的重连链路序列  $G^R(g^0(\delta)) = \{R^1, R^2, \dots\}$ , 预测重连链路的生存数据定义为  $H^R = \{\langle T_q, \mathbf{F}_q^R \rangle \mid q \in [1, \sum_{t \geq 1} |R^t|]\}$ .  $H^R$  中的每个实例  $\langle T_q, \mathbf{F}_q^R \rangle$  表示  $G^R(g^0(\delta))$  中的一条重连链路, 其中  $T_q$  表示节点对  $(i, j)_q$  保持不连通的持续时间,  $\mathbf{F}_q^R$  表示节点对  $(i, j)_q$  相对于投影网络  $P^{t-T_q}(L)$  的特征向量.

基于定义 8 和定义 9, 我们可借助于 Cox 比例风险模型分别学习链路消失和重连所对应特征向量的权重系数. Cox 比例风险模型采用最大似然估计方法对模型参数进行估计, 上述 2 个事件对应的协变量系数  $\mathbf{W}^M$  和  $\mathbf{W}^R$  可分别通过式 (6) (7) 学习得到:

$$\mathbf{W}^M = \arg \max_{\mathbf{W} \in \mathbb{R}^p} \sum_p (\mathbf{W}^T \mathbf{F}_p^M - \lg \sum_{o \in R(t_p)} \mathbf{W}^T \mathbf{F}_o^M), \quad (6)$$

其中,  $p \in [1, \sum_{t \geq 1} |M^t|]$ .

$$\mathbf{W}^R = \arg \max_{\mathbf{W} \in \mathbb{R}^q} \sum_q (\mathbf{W}^T \mathbf{F}_q^R - \lg \sum_{o \in R(t_q)} \mathbf{W}^T \mathbf{F}_o^R), \quad (7)$$

其中,  $q \in [1, \sum_{t \geq 1} |R^t|]$ .

令  $\mathbf{A}^t = [A_{ij}^t] \in \mathbb{R}^{n \times n}$  表示静态网络  $g^t$  的邻接矩阵,  $\mathbf{F}_{ij}^{t,M}$  和  $\mathbf{F}_{ij}^{t,R}$  分别表示  $E^t$  和  $P^t(L)$  中节点对  $(i, j)$  的特征向量. 节点对  $(i, j)$  在周期  $t+1$  时的消失概率和重连概率分别被定义为

$$\theta_{ij}^t = A_{ij}^t (\mathbf{W}^M)^T \mathbf{F}_{ij}^{t,M}, \quad (8)$$

$$\phi_{ij}^t = (1 - A_{ij}^t - \vartheta_{ij}) (\mathbf{W}^R)^T \mathbf{F}_{ij}^{t,R}. \quad (9)$$

我们进一步将时序网络链路预测问题模拟为一种动态博弈过程. 假设在一个离散时间动态系统 ( $t=0, 1, \dots$ ) 中, 包含 2 个超级玩家  $\alpha$  和  $\beta$ . 在每个周期  $t$  中, 每个玩家有相同的策略空间, 即  $S_\alpha = S_\beta = V = \{1, 2, \dots, n\}$ . 令  $i \in S_\alpha$  和  $j \in S_\beta$  分别表示玩家  $\alpha$  和  $\beta$  所采取的策略. 这样一个策略点  $s^t = (i, j) \in V \times V$ , 可以看作是静态网络  $g^t$  中的一个节点对  $(i, j)^t$ . 我们进一步假设给定一个策略点  $s^t = (i, j)$ , 玩家  $\alpha$  和  $\beta$  具有相同的收益值, 即  $u_\alpha^t(i, j) = u_\beta^t(i, j) = u_{ij}^t$ . 这样在当前周期下, 超级玩家  $\alpha$  和  $\beta$  的收益矩阵  $\Psi^t$  可表示为

$$\Psi^t = \begin{pmatrix} u_{11}^t & \cdots & u_{1n}^t \\ \vdots & & \vdots \\ u_{n1}^t & \cdots & u_{nn}^t \end{pmatrix}, \quad (10)$$

$\Psi^t$  对应的纳什均衡点集为

$$D^t = \{(i, j) : \max_i \max_j \{u_{ij}^t\} = \max_j \max_i \{u_{ij}^t\}\}. \quad (11)$$

回到问题 1 定义的时序网络链路预测问题,  $\Psi^t$

可被分别定义为  $\Theta^t = [\theta_{ij}^t] \in \mathbb{R}^{n \times n}$  和  $\Phi^t = [\phi_{ij}^t] \in \mathbb{R}^{n \times n}$ , 这样与收益矩阵  $\Theta^t$  和  $\Phi^t$  所对应的纳什均衡集分别表示为  $D^{t,M}$  和  $D^{t,R}$ . 其中,  $(i, j)^M \in D^{t,M}$  表示互连节点  $i$  和  $j$  都有意愿终止当前的连接关系. 类似地,  $(i, j)^R \in D^{t,R}$  表示节点  $i$  和  $j$  彼此都想建立一条新的链路. 这样, 时刻  $t+1$  的链路集合  $E^{t+1}$  可表示为

$$E^{t+1} = Q(E^t, \Theta^t, \Phi^t) = E^t - D^{t,M} + D^{t,R}. \quad (12)$$

上述动态博弈模型中, 对  $D^{t,M}$  和  $D^{t,R}$  的预测可理解为任意节点对  $(i, j) \in V \times V$  间的一种“双向选择”过程. 给定  $(i, j) \in E^t \Rightarrow i \in N_i^t \wedge j \in N_j^t$ , 当且仅当  $j = \arg \max_{k \in N_j^t} \theta_{jk}^t$  和  $j = \arg \max_{k \in N_i^t} \theta_{ik}^t$  同时成立时,  $(i, j) \in D^{t,M}$ , 也就是说, 节点  $i$  和节点  $j$  同时终止彼此间的相互连接可最大化其自身的效益值. 与此类似, 当且仅当  $j = \arg \max_{k \in N_j^t} \theta_{jk}^t$  和  $j = \arg \max_{k \in N_i^t} \theta_{ik}^t$  同时成立时,  $(i, j) \in D^{t,R}$ , 即节点  $i$  和节点  $j$  同时选择彼此建立一条新的链路. 这种双向选择机制可以极大地压缩链路预测的搜索空间. 以节点  $i$  为例, 我们可以分别计算该节点在时刻  $t$  的最“乐意”终止连接的邻居和建立新链路的节点, 这种计算对所有节点来说是彼此独立的, 可并行化执行. 同时, 由邻居相似性函数定义 (定义 5) 可知, 与节点  $i$  重连的节点只可能存在于距离节点  $i$  最短路径距离不超过 2 的范围内. 这样,  $\forall i \in \{1, 2, \dots, n\}$ , 求解其对应的重连节点的搜索空间将由  $O(n^2)$  压缩为  $O(n \langle k \rangle^2)$ , 其中  $\langle k \rangle$  表示网络的平均度, 极大地降低了链路预测的时间复杂度, 同时可规避传统有监督预测方法的分类样本分布失衡问题.

#### 4 基于自治计算的网络演化算法

我们首先定义一种适用于链路预测的自治计算系统  $AOC^{LP} = \{V, t, L, \mathbf{W}^M, \mathbf{W}^R\}$ , 其中  $V = \{v_1, v_2, \dots, v_n\}$  表示  $n$  个节点智能体,  $t$  表示系统时间,  $L$  表示时间窗口长度,  $\mathbf{W}^M$  和  $\mathbf{W}^R$  分别表示 Cox 比例风险模型学习的协变量系数.  $AOC^{LP}$  的每个节点智能体  $v_i$  可由一个三元组  $\langle \tilde{N}_i^t, \tilde{\omega}_i^t, \tilde{\omega}'_i^t \rangle$  描述, 其中  $\tilde{N}_i^t = \{j \mid (i, j) \in \tilde{E}^t\}$  表示时刻  $t$  静态网络中节点  $i$  的邻居集. 节点  $i$  在对应投影网络  $P^t(L)$  中的邻居集为  $\hat{N}_i^t = \{j \mid \exists x \geq 0, t-L+1 \leq x \leq t, (i, j) \in \tilde{E}^x\}$ , 节点  $i$  的可达节点集定义为  $N_i^{t,\lambda} = \{j \mid dis_{ij}^t \leq \lambda\}$ , 其中  $dis_{ij}^t$  表示投影  $P^t(L)$  中节点  $i$  和节点  $j$  之间的最短距离.  $\tilde{\omega}_i^t = \arg \max_{j \in \tilde{N}_i^t} \theta_{ij}^t$  表示在下一时间周期  $t+1$

中最有可能同节点  $i$  终止连接的某个邻居,  $\tilde{\omega}_i^t = \arg \max_{j \in (\tilde{N}_i^{t,\lambda} - \tilde{N}_i^t)}$  表示在下一时间周期  $t+1$  中最有可能同节点  $i$  建立新连接的某个节点.

### 算法 1. 网络演化算法.

输入: 初始静态网络  $g^0(\delta) = (V, E^0)$ 、时间窗口长度  $L$ 、最大迭代次数  $I$ 、消失链路的协变量系数  $W^{M^*}$ 、重连链路的协变量系数  $W^{R^*}$ ;

输出: 网络序列  $\tilde{G}(g^0)$ .

- ①  $t \leftarrow 0$ ;
- ②  $\forall i \in V, v_i \leftarrow \langle N_i^0, \emptyset, \emptyset \rangle$ ;
- ③ repeat
- ④  $\forall i \in V$ , 并行计算:
  - ⑤ 更新  $\tilde{N}_i^t$  和  $N_i^{t,2}$ ;
  - ⑥  $\forall j \in N_i^{t,2}$ , 更新  $F_{ij}^t$ ;
  - ⑦  $\forall j \in \tilde{N}_i^t, \theta_{jk}^t \leftarrow (W^{M^*})^T F_{ij}^t$ ;
  - ⑧  $\forall j \in N_i^{t,2} - \tilde{N}_i^t, \phi_{ij}^t \leftarrow (W^{R^*})^T F_{ij}^t$ ;
  - ⑨  $\tilde{\omega}_i^t = \arg \max_{j \in \tilde{N}_i^t} \theta_{ij}^t$ ;
  - ⑩  $\tilde{\omega}_i^t = \arg \max_{j \in (\tilde{N}_i^{t,\lambda} - \tilde{N}_i^t)} \phi_{ij}^t$ ;
  - ⑪  $\tilde{D}^{t,M}.insert(\langle i, \tilde{\omega}_i^t \rangle), \tilde{D}^{t,R}.insert(\langle i, \tilde{\omega}_i^t \rangle)$ ;
- ⑫ for each  $\langle i, \tilde{\omega}_i^t \rangle \in \tilde{D}^{t,M}$  do
- ⑬ if  $\nexists \langle j, \tilde{\omega}_j^t \rangle \in \tilde{D}^{t,M}, \langle i, \tilde{\omega}_i^t \rangle = \langle \tilde{\omega}_j^t, j \rangle$  then
- ⑭  $\tilde{D}^{t,M}.delete(\langle i, \tilde{\omega}_i^t \rangle)$ ;
- ⑮ end if
- ⑯ end for
- ⑰ for each  $\langle i, \tilde{\omega}_i^t \rangle \in \tilde{D}^{t,R}$  do
- ⑱ if  $\langle j, \tilde{\omega}_j^t \rangle \in \tilde{D}^{t,R}, \langle i, \tilde{\omega}_i^t \rangle = \langle \tilde{\omega}_j^t, j \rangle$  then
- ⑲  $\tilde{D}^{t,R}.delete(\langle i, \tilde{\omega}_i^t \rangle)$ ;
- ⑳ end if
- ㉑ end for
- ㉒  $\tilde{E}^{t+1} \leftarrow \tilde{E}^t - \tilde{D}^{t,M} + \tilde{D}^{t,R}$ ;
- ㉓  $\tilde{G}.insert(\tilde{E}^{t+1})$ ;
- ㉔  $t \leftarrow t+1$ ;
- ㉕ until  $t \leq I$ ;
- ㉖ return  $\tilde{G}(g^0)$ .

基于上述定义的自治计算系统, 本文提出了基于博弈论的时序网络预测算法, 具体过程见算法 1. 从单个节点智能体的角度来看, 步骤⑫初始化节点智能体  $v_i$  的三元组为  $\langle N_i^0, \emptyset, \emptyset \rangle$ , 其中  $N_i^0$  表示

初始静态网络  $g^0$  中节点  $i$  的邻居集. 在每轮迭代中, 步骤⑤~⑪并行化地计算  $\tilde{\omega}_i^t$  和  $\tilde{\omega}_i^t$ ; 步骤⑫~⑳计算相应收益矩阵  $\Theta^t$  和  $\Phi^t$  的纳什均衡点集; 步骤㉑生成下一周期静态网络的链路集合. 当系统时间达到预设的最大迭代次数  $I$  时, 算法将会输出一个预测的网络演化序列  $\tilde{G}(g^0) = \{\tilde{E}^1, \tilde{E}^2, \dots, \tilde{E}^I\}$ .

假定初始的静态网络  $g^0(\delta)$  包含  $n$  个节点, 每个节点对的特征向量维度为  $d$ , 投影网络  $P^t(L)$  的链路数为  $\hat{m}^t$ , 该投影网络的平均度  $\langle \hat{k}^t \rangle = 2 \hat{m}^t / n$ . 在算法 1 的每轮迭代中, 计算可达节点集节点  $i$  的特征向量占用最大的时间开销, 其时间复杂度为  $O(nd(\langle \hat{k}^t \rangle)^3)$ . 因此算法 1 的时间复杂度为  $O(\sum_{t=0}^I nd(\langle \hat{k}^t \rangle)^3)$ . 在后续的仿真实验中, 每轮迭代过程中,  $\hat{m}^t \leq m$ . 这样算法 1 的时间复杂度可简化为  $O(Ind(\langle \hat{k}^t \rangle)^3)$ , 其中  $\langle k \rangle = 2m/n$  表示初始网络  $g^0(\delta)$  的平均度.

## 5 实验和结果

### 5.1 实验设置

我们选择斯坦福网络分析平台 SNAP 中的 4 个时序网络 Math, Ask, Super, Stack, 生成 4 个  $\epsilon$ -邻接网络序列. 具体来说, 对于每个时序网络, 我们使用第 1 年的数据元组构建一个无向网络, 并抽取该网络的最大连通子图生成初始的静态网络  $g^0$  (初始静态网络的统计特征见表 1). 接下来, 我们进一步将  $\epsilon$  设置为 2%, 生成相应的  $\epsilon$ -邻接网络序列. 图 1 分别记录了 4 个  $\epsilon$ -邻接网络序列中相应的  $|E^t|, |M^t|, |R^t|$  包含的链路数量. 可见, 随着时间周期  $t$  的增加,  $|E^t|, |M^t|, |R^t|$  都呈显著的下降趋势. 后续实验中, 我们将这 4 个  $\epsilon$ -邻接网络序列作为真实的网络演化序列, 并将算法 1 分别应用于表 1 的 4 个初始静态网络. 通过对比预测得到的网络演化序列和真实网络演化序列的相似度, 来评价链路预测算法的优劣.

Table 1 Statistical Characteristics Associated with the Induced  $\delta$ -Networks ( $\delta = 365$  d)

表 1 初始的 $\delta$ -时序范式静态网络统计特征 ( $\delta = 365$ d)						
$g^0(\delta)$	$n$	$m$	$\langle k \rangle$	$c$	$d$	$\langle dis_{ij} \rangle$
Math	4 199	21 273	9.981	0.173	9	3.454
Ask	7 084	23 520	6.415	0.097	10	3.825
Super	13 733	60 454	8.570	0.054	11	3.789
Stack	46 232	563 699	24.386	0.044	7	2.916

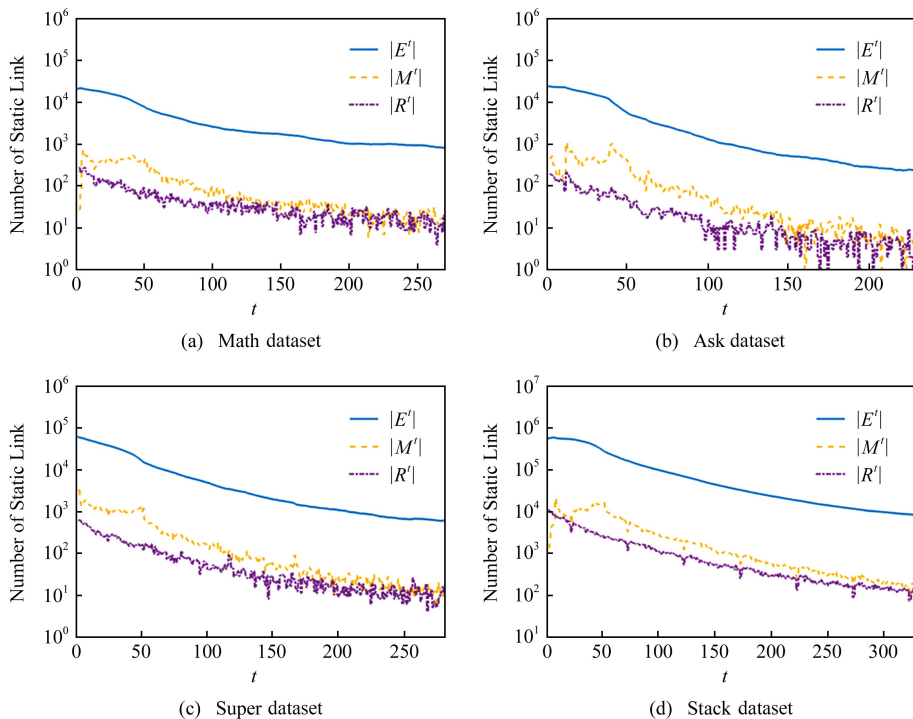


Fig. 1 The movements of the numbers of static links in the three link sets along the corresponding  $\epsilon$ -adjacent network sequences ( $\epsilon=2\%$ )

图 1  $\epsilon$ -邻接网络序列中包含的链路数量( $\epsilon=2\%$ )

我们选择 5 种基于监督学习的链路预测方法,包括 Logistic 回归(logistic regression, LR)、SVM、决策树(C4.5)、多层感知器神经网络(multi-layer perception neural network, MPNN)和朴素贝叶斯(naïve Bayes, NB)作为对比算法.针对每个耦合的静态网络对,如 $(E^{t-1}, E^t)$ ,然后将所有  $M^t$  中的链路标记为正则,并从  $E^{t-1} - M^t$  中随机选择相同数量的链路标记为反例.这些链路实例的特征向量基于相应静态网络  $E^{t-1}$  计算生成.这样,我们可以选择一种分类器在上述训练集上进行训练,并将训练好的分类器用于消失链路预测.采取相同的思路,我们可以构建另一个训练样本进行重连链路预测,其中节点对的特征向量基于相应的投影网络  $P^{t-1}(L)$  计算生成.此外,我们还选择了 3 种基于概率模型的链路预测方法,包括 Wang 等人<sup>[33]</sup>提出的 LPM 模型、Coutant 等人<sup>[34]</sup>提出的 PRM-CU 模型、Wang 等人<sup>[35]</sup>提出的 SDNE 模型.

给定一个  $\epsilon$ -邻接网络序列  $G^\epsilon(g^0(\delta)) = \{E^1, E^2, \dots, E^T\}$ ,令  $\tilde{G}^\epsilon(g^0(\delta)) = \{\tilde{E}^1, \tilde{E}^2, \dots, \tilde{E}^P\}$  表示预测的网络演化序列.预测算法的效果可以通过比较这 2 个网络演化序列的相似性来评价.本文的评价指标为  $AvgF1$ :

$$AvgF1 = \frac{1}{2P} \sum_t \max_{\tau} F1(\tilde{E}^t, E^\tau) + \frac{1}{2T} \sum_{\tau} \max_t F1(\tilde{E}^t, E^\tau), \quad (13)$$

其中  $F1(\tilde{E}^t, E^\tau)$  表示链路集  $\tilde{E}^t$  和  $E^\tau$  的  $F1$ -score,具体定义参见式(3)~(5).

## 5.2 基于 Cox 比例风险模型的参数学习

我们首先分别采用 SDGM-MLP 和 SDGM-RLP 模型基于 4 个  $\epsilon$ -邻接网络序列生成相应的生存数据集  $H^M$  和  $H^R$ ,相应的参数设置为  $\delta=365$  d、 $\epsilon=2\%$  和  $T=50$ .然后,我们利用 Cox 比例风险模型来分别学习特征向量  $F_b^t$  的系数  $W^{M^*}$  和  $W^{R^*}$ .以 Math 数据集为例,与之对应的  $H^M$  和  $H^R$  生存数据集分别包含 32742 和 12287 个数据样本.由于我们仅关心“事件”(如链路消失和链路重连)发生的相对风险,因此  $H^M$  和  $H^R$  剔除了相应的删失数据.实验中,预设的显著性水平设为 0.05,我们采用最大对数似然估计方法来估计协变量的系数,参见式(6)(7).

表 2 展示了 Math 时序网络的  $H^M$  数据集的生存分析结果.其中,SE 表示  $w^M$  的标准误差;Wald 表示卡方检验统计值,该统计值可反映给定协变量  $w^M$  是否与零假设有统计学显著差异; $\exp(w^M)$

表示与协变量  $\omega^M$  单位递增的相对风险. 实验中, 我们选择“向后似然估计过程”来估计协变量系数. 首先, 所有协变量的显著性被同时分析, 当某些协变量的显著性水平超过 0.05 时, 我们在后续迭代过程中忽略这些协变量. 重复上述过程, 最大似然估计一共进行了 5 次迭代. 此时, 所有协变量的显著性水平都小于 0.05, 表示这些协变量对消失链路预测是有意义的. 我们进一步分析这些协变量的系数, 从表 2 的第 6 列, 我们发现  $f_p^{CN}$ ,  $f_p^{So}$ ,  $f_p^{HP}$  是与“链路消失事件”呈现正相关的 3 个协变量; 协变量  $f_p^{Sa}$  和  $f_p^{AA}$  同该事件的发生呈显著的负相关; 其他协变量与此事件的发生是独立的. 我们同时分析了 Math 时序网络的  $H^R$  数据集, 其结果见表 3. 采用“向后似然估计过程”, 对  $H^R$  数据集的最大似然估计共经历了 6 轮迭代. 由表 3 可知, 与“链路重连事件”呈正相关的协变量只有  $f_p^{So}$ , 而  $f_p^{JC}$  和  $f_p^{Sa}$  与此事件发生呈负相关.

采用相同的分析过程, 我们可以分别对其他 3 个时序网络进行类似的分析, 其结果见表 4. 其中, 每个数据集中, 加粗字体的协变量系数表示该特征与对应事件是正相关, 正常字体的协变量系数表示该特征与对应事件是负相关, 而缺失协变量系数表示该特征与对应事件不相关. 在所有数据集中, PA 特征<sup>[17]</sup>与链路的消失和重连都是独立的. 经典的 BA 网络演化模型<sup>[40]</sup>假设一个新的节点同已知网络中的节点  $i$  连接的概率同  $k_i$  呈正比. 而本文考虑的网络

演化序列是基于固定节点集  $V$  生成的, 排除了新节点加入对网络演化的影响. 在这种情境下, 传统的基于 BA 模型的无标度网络生成模型可能并不适用.

**Table 2 Results of Survival Analysis on  $H^M$  Dataset of the Temporal Network Math**

**表 2 基于 Math 时序网络  $H^M$  数据集的生存分析结果**

Covariates	$\omega^M$	SE	Wald	Sig.	$\exp(\omega^M)$
CN	<b>0.038</b>	<b>0.003</b>	<b>203.995</b>	<b>0.000</b>	<b>1.039</b>
Sa	-7.559	0.740	104.217	0.000	0.001
So	<b>5.580</b>	<b>0.715</b>	<b>60.846</b>	<b>0.000</b>	<b>265.038</b>
HP	<b>1.073</b>	<b>0.087</b>	<b>153.450</b>	<b>0.000</b>	<b>2.925</b>
AA	-0.195	0.035	30.154	0.000	0.823

Notes: Bold values represent a positive correlation between the feature and link missing event, while the regular ones represent a negative correlation. Sig. means significance.

**Table 3 Results of Survival Analysis on  $H^R$  Dataset of the Temporal Network Math**

**表 3 基于 Math 时序网络  $H^R$  数据集的生存分析结果**

Covariates	$\omega^R$	SE	Wald	Sig.	$\exp(\omega^R)$
JC	-11.908	4.122	8.346	0.004	0.000
Sa	-2.640	0.582	20.610	0.000	0.071
So	<b>9.561</b>	<b>2.556</b>	<b>13.996</b>	<b>0.000</b>	<b>14 194.692</b>

Notes: Bold values represent a positive correlation between the feature and link reconnection event, while the regular ones represent a negative correlation. Sig. means significance.

**Table 4 Learning Results of the Covariate Coefficients Based on the Cox Proportional Hazards Model**

**表 4 Cox 比例风险模型协变量系数学习结果**

Covariates	Math		Ask		Super		Stack	
	$\omega^M$	$\omega^R$	$\omega^M$	$\omega^R$	$\omega^M$	$\omega^R$	$\omega^M$	$\omega^R$
CN	<b>0.038</b>		<b>0.028</b>	<b>0.007</b>		<b>0.005</b>	-0.009	-0.012
JC		-11.908	-14.291	-73.303	-5.547	<b>20.624</b>	-24.262	<b>13.878</b>
Sa	-7.559	-2.640	-4.999	-2.010	-4.633		-5.383	<b>1.619</b>
So	<b>5.580</b>	<b>9.561</b>	<b>9.879</b>	<b>42.203</b>	<b>6.378</b>		<b>16.596</b>	-5.160
HP	<b>1.073</b>		<b>0.367</b>		<b>0.193</b>	<b>0.303</b>	-0.065	
HD						-11.431	-0.828	-3.846
LN			<b>6.234</b>		<b>2.145</b>	-11.530	<b>8.906</b>	-2.939
PA								
AA	-0.195		<b>0.506</b>	-0.192	<b>0.574</b>	-0.162	<b>1.504</b>	-0.145

Notes: Bold values represent a positive correlation between the feature and link missing/reconnection event, the regular ones represent a negative correlation, while the missing values represent that the features are independent from the related events.

### 5.3 时序网络链路预测结果

基于每个数据集中学习得到的特征向量权重  $W^M$  和  $W^R$ , 利用算法 1 生成一组开始于静态网络

$g^0(\delta)$  的网络演化序列  $\tilde{G}(g^0) = \{\tilde{E}^1, \tilde{E}^2, \dots, \tilde{E}^P\}$ , 最大迭代次数  $P$  设置为  $P = T = 100$ . 图 2 给出了 4 个网络数据集中真实网络演化序列同预测网络



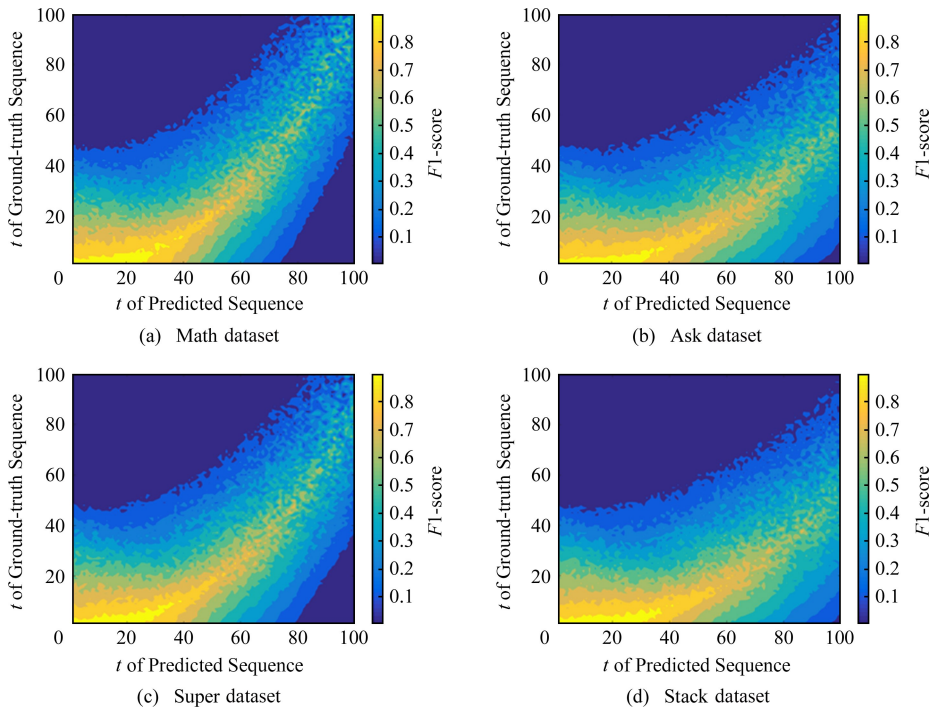


Fig. 2 The  $F1$ -score matrices between the predicted network evolution sequence and the ground-truth one

图 2 真实网络演化序列同预测网络演化序列的  $F1$ -score 矩阵

演化序列  $100 \times 100$  的  $F1$ -score 矩阵.其中  $x$  轴表示预测序列中静态网络的时间索引, $y$  轴表示真实序列中静态网络的时间索引, $(x, y)$  位置的像素代表对应链路集  $\tilde{E}^x$  和  $E^y$  的  $F1$ -score.在算法 1 迭代过程的前半段,本文方法预测得到的静态网络同真实静态网络匹配度较好.随着迭代轮次的增加,算法的预测精度将有所下降.由于本文考虑的是预测给定初始静态网络的演化序列,随着迭代轮次的增加,时刻  $x$  预测的静态网络会受到之前预测结果的影响,因此在迭代过程中,后续的预测误差将被逐级放大.另外,我们还发现时刻  $x$  预测的静态网络同真实网络演化序列中最佳匹配网络的时间索引相比存在一定的滞后性.这是因为在进行网络演化预测时,本文采用的双向选择机制相对较为保守,预测的消失链路集和重连链路集的规模要小于真实情况.

接下来,我们将算法 1 同其他链路预测方法进行性能对比.从图 3 中可观察到,本文所提方法明显优于对比算法.其中,MPNN 只在 Stack 网络上有较好的表现,但无法准确预测其他 3 个网络的演化序列.多数基于监督式学习的链路预测方法对训练样本很敏感,受分类样本失衡问题的影响,这些方法尽管在消失链路预测上可以取得不错效果,但难以准确地预测重连链路.本文方法采用动态博弈的双向选择机制,可以筛选出最可能消失的链路和最可能

重连的节点对.这种相对保守的预测方法尽管在召回率方面有所牺牲,但可以获得更高的准确率.因此,在  $AvgF1$  指标上,本文方法要明显优于基于有监督学习的链路预测方法.与基于概率模型的链路预测方法相比,本文方法在 Super 网络上的预测性能稍逊于 SDNE;而在其他网络上,算法 1 得到的网络演化序列能更好地模拟真实网络的演化过程.

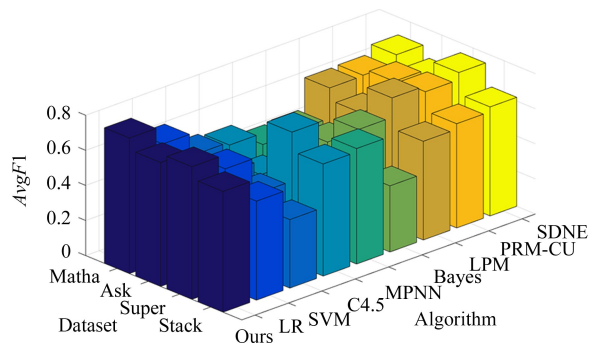


Fig. 3 Accuracy comparison of the network evolution sequences obtained by different link prediction algorithms

图 3 不同算法对网络演化序列的预测准确性比较

最后,我们对比了不同链路预测算法在 100 次连续预测过程中的执行效率,从图 4 可以看出,不论在小规模的 Math 数据集上,还是在大规模的 Stack 数据集上,算法 1 的执行效率都明显优于其他对比

算法.由于本文采用了Cox比例风险模型对每个数据集中不相关的链路特征进行了排序,因此在迭代预测过程中,每个节点对的特征向量更新耗时更少.在动态博弈框架下,每个节点的特征向量更新和策略选择是独立的,故算法1可以很好地并行化.另外,我们发现本文所提方法的执行效率比基于有监督学习的链路预测算法提高了将近2倍,比基于概率模型的链路预测算法的执行效率提高了将近10倍.

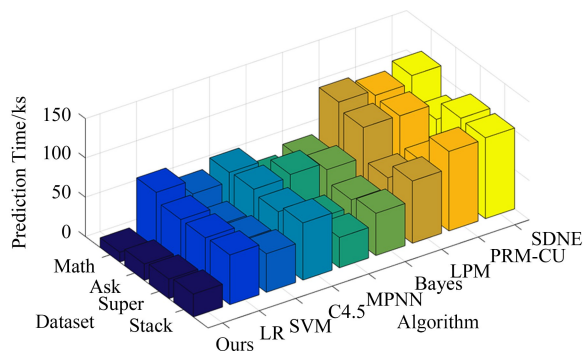


Fig. 4 Comparison of the total running time of different link prediction algorithms

图4 不同链路预测算法的执行时间对比

## 6 总 结

本文提出了一种新颖的用于预测时序网络的演化序列半监督学习框架.为简化研究问题,假设时序网络的演化是在一组固定节点集中进行的,此时网络的演化预测问题转换为:如何根据历史的网络演化序列预测下一时刻静态网络的消失链路集和重连链路集.为解决此问题,首先为每条链路定义一组基于邻居相似度的特征向量,利用Cox比例风险模型学习对应链路消失和重连事件的特征向量权重.为压缩网络演化预测的搜索空间,提出了一种基于动态博弈的双向选择机制来预测未来的网络拓扑结构.在理论研究基础上,提出了一种基于多智能体自治计算的链路预测算法,并在真实时序网络数据集上验证了方法的有效性和高效性.未来将关注于如何将网络的社团结构信息融入预测模型,进而提高预测算法的准确性.

## 参 考 文 献

- [1] Paranjape A, Benson A R, Leskovec J. Motifs in temporal networks [C] // Proc of the 10th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2017: 601-610
- [2] Li Shuangdong, Ni Jinlong. Analysis of college students' association structure based on network graph [J]. Journal of Liaocheng University: Natural Sciences Edition, 2017, 30(2): 106-110 (in Chinese)  
(李双东, 倪晋龙. 基于网络图的大学生社团结构分析[J]. 聊城大学学报: 自然科学版, 2017, 30(2): 106-110)
- [3] Wang Ruiguo, Ye Yaling, Bu Zhan. A community detection approach based on network embedding [J]. Journal of Liaocheng University: Natural Sciences Edition, 2019, 32(4): 72-80 (in Chinese)  
(王瑞国, 叶雅玲, 卜湛. 一种基于网络嵌入的社区发现方法[J]. 聊城大学学报: 自然科学版, 2019, 32(4): 72-80)
- [4] Lü Linyuan. Link prediction in complex networks [J]. Journal of University Electronic Science and Technology of China, 2010, 39(5): 651-661 (in Chinese)  
(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)
- [5] Lee C, Pham M, Jeong M K, et al. A network structural approach to the link prediction problem [J]. Inform Journal on Computing, 2015, 27(2): 249-267
- [6] Wang Zhiqiang, Liang Jiye, Li Ru, et al. An approach to cold-start link prediction: Establishing connections between non-topological and topological information [J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(11): 2857-2870
- [7] Holme P, Saramäki J. Temporal networks [J]. Physics Reports, 2012, 519(3): 97-125
- [8] Yu Haiyuan, Braun P, Yildirim M A, et al. High-quality binary protein interaction map of the yeast interactome network [J]. Science, 2008, 322(5898): 104-110
- [9] Wang Yu, Gao Lin. Social circle-based algorithm for friend recommendation in online social networks [J]. Chinese Journal of Computers, 2014, 37(4): 801-808 (in Chinese)  
(王珂, 高琳. 基于社交圈的在线社交网络朋友推荐算法[J]. 计算机学报, 2014, 37(4): 801-808)
- [10] Zhang Jinzhu. Uncovering mechanisms of co-authorship evolution by multirelations-based link prediction [J]. Information Processing & Management, 2016, 53(1): 42-51
- [11] Cholewo T J, Zurada J M. Sequential network construction for time series prediction [C] // Proc of the 15th IEEE Int Joint Conf on Neural Networks. Piscataway, NJ: IEEE, 1997: 2034-2038
- [12] Hasan M A, Zaki M J. Social Network Data Analytics [M]. Berlin: Springer, 2011
- [13] Lü Linyuan, Zhou Tao. Link prediction in complex networks: A survey [J]. Physica A: Statistical Mechanics and Its Applications, 2011, 390(6): 1150-1170
- [14] Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks [J]. Journal of the American Society for Information Science and Technology, 2007, 58(7): 1019-1031

- [15] Cui Aixiang, Fu Yan, Shang Mingsheng, et al. Emergence of local structures in complex network; Common neighborhood drives the network evolution [J]. *Acta Physica Sinica*, 2011, 60(3): No.038901 (in Chinese)  
(崔爱香, 傅彦, 尚明生, 等. 复杂网络局部结构的涌现; 共同邻居驱动网络演化[J]. *物理学报*, 2011, 60(3): No.038901)
- [16] Jaccard P. Etude comparative de la distribution florale dans une portion des Alpes et des Jura [J]. *Bulletin Del La Societe Vaudoise Des Sciences Naturelles*, 1901, 37(142): 547-579
- [17] Newman M E J. Clustering and preferential attachment in growing networks [J]. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, 2001, 64(2): No.025102
- [18] Adamic L A, Adar E. Friends and neighbors on the Web [J]. *Social Networks*, 2003, 25(3): 211-230
- [19] Katz L. A new status index derived from sociometric analysis [J]. *Psychometrika*, 1953, 18(1): 39-43
- [20] Lü Linyuan, Jin Cihang, Zhou Tao. Similarity index based on local paths for link prediction of complex networks [J]. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, 2009, 80(2): No.046122
- [21] Yang Niya, Peng Tao, Liu Lu. Link prediction method based on clustering and decision tree [J]. *Journal of Computer Research and Development*, 2017, 54(8): 1795-1803 (in Chinese)  
(杨妮亚, 彭涛, 刘露. 基于聚类 and 决策树的链路预测方法 [J]. *计算机研究与发展*, 2017, 54(8): 1795-1803)
- [22] Lichtenwalter R N, Lussier J T, Chawla N V. New perspectives and methods in link prediction [C] //Proc of the 16th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2010: 243-252
- [23] Tong Hanghang, Faloutsos C, Pan Jiayu. Fast random walk with restart and its applications [C] //Proc of the 6th Int Conf on Data Mining. Piscataway, NJ: IEEE, 2006: 613-622
- [24] Jin Zhaoyan, Wu Quanyuan, Shi Dianxi, et al. Random walk based inverse influence research in online social networks [C] //Proc of the 11th IEEE Int Conf on Embedded and Ubiquitous Computing. Piscataway, NJ: IEEE, 2013: 2206-2213
- [25] Popescul A, Ungar L H. Statistical relational learning for link prediction [C] //Proc of the 18th IJCAI Workshop on Learning Statistical Models from Relational Data. Piscataway, NJ: IEEE, 2003: 109-115
- [26] Hasan M A, Chaoji V, Salem S, et al. Link prediction using supervised learning [C/OL] //Proc of the 4th Workshop on Link Analysis, Counterterrorism and Security. 2006 [2019-03-03]. <http://www.cs.rpi.edu/~zaki/PaperDir/LINK06.pdf>
- [27] Lichtenwalter R N, Lussier J T, Chawla N V. New perspectives and methods in link prediction [C] //Proc of the 16th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2010: 243-252
- [28] Schmidhuber J. Deep learning in neural networks: An overview [J]. *Neural Networks*, 2015, 61: 85-117
- [29] Liu Zhen, Zhang Qianming, Lü Linyuan, et al. Link prediction in complex networks: A local naive Bayes model [J]. *Europhysics Letters*, 2011, 96(4): No.48007
- [30] Pujari M, Kanawati R. Supervised rank aggregation approach for link prediction in complex networks [C] //Proc of the 21st Int Conf Companion on World Wide Web. New York: ACM, 2012: 1189-1196
- [31] Chen Ning, Zhu Jun, Xia Fei, et al. Discriminative relational topic models [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, 37(5): 973-986
- [32] Nguyen-Thi A T, Nguyen P Q, Ngo T D, et al. Transfer adaboost SVM for link prediction in newly signed social networks using explicit and PNR features [J]. *Procedia Computer Science*, 2015, 60(1): 332-341
- [33] Wang Chao, Satuluri V, Parthasarathy S. Local probabilistic models for link prediction [C] //Proc of the 7th IEEE Int Conf on Data Mining. Piscataway, NJ: IEEE, 2007: 322-331
- [34] Coutant A, Leray P, Capitaine H L. Probabilistic relational models with clustering uncertainty [C] //Proc of the 33rd IEEE Int Joint Conf on Neural Networks. Piscataway, NJ: IEEE, 2015: 1-8
- [35] Wang Daixin, Cui Peng, Zhu Wenwu. Structural deep network embedding [C] //Proc of the 22nd ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2016: 1225-1234
- [36] Sharan U, Neville J. Temporal-relational classifiers for prediction in evolving domains [C] //Proc of the 8th IEEE Int Conf on Data Mining. Piscataway, NJ: IEEE, 2008: 540-549
- [37] Rossi R, Neville J. Time-evolving relational classification and ensemble methods [C] //Proc of the 16th Pacific-Asia Conf on Knowledge Discovery and Data Mining. Berlin: Springer, 2012: 1-13
- [38] Clauset A, Moore C, Newman M E. Hierarchical structure and the prediction of missing links in networks [J]. *Nature*, 2008, 453: 98-101
- [39] Domenico M D, Granell C, Porter M A, et al. The physics of spreading processes in multilayer networks [J]. *Nature Physics*, 2016, 12(10): 901-906
- [40] Barabási A, Albert R. Emergence of scaling in random networks [J]. *Science*, 1999, 286(5439): 509-512



**Liu Liu**, born in 1992. Master. His main research interests include link prediction and data mining.



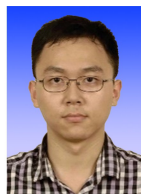
**Wang Yuyao**, born in 1994. PhD candidate. His main research interests include complex network and machine learning.



**Cao Jie**, born in 1969. PhD, professor, PhD supervisor. Member of CCF. His main research interests include business intelligence, data mining, and recommendation system.



**Ni Qixuan**, born in 1997. Master candidate. Her main research interests include data mining and time series analysis.



**Bu Zhan**, born in 1987. PhD, associate professor. Member of CCF. His main research interests include social computing, data mining, complex network, and game theory.

## 2017 年《计算机研究与发展》高被引论文 TOP10

排名	论文信息
1	施巍松, 孙辉, 曹杰, 张权, 刘伟. 边缘计算: 万物互联时代新型计算模型[J]. 计算机研究与发展, 2017, 54(5): 907-924 Shi Weisong, Sun Hui, Cao Jie, Zhang Quan, Liu Wei. Edge Computing—An Emerging Computing Model for the Internet of Everything Era [J]. Journal of Computer Research and Development, 2017, 54(5): 907-924
2	黎建辉, 沈志宏, 孟小峰. 科学大数据管理: 概念、技术与系统[J]. 计算机研究与发展, 2017, 54(2): 235-247 Li Jianhui, Shen Zhihong, Meng Xiaofeng. Scientific Big Data Management: Concepts, Technologies and System [J]. Journal of Computer Research and Development, 2017, 54(2): 235-247
3	张玉清, 周威, 彭安妮. 物联网安全综述[J]. 计算机研究与发展, 2017, 54(10): 2130-2143 Zhang Yuqing, Zhou Wei, Peng Anni. Survey of Internet of Things Security [J]. Journal of Computer Research and Development, 2017, 54(10): 2130-2143
4	祝烈煌, 高峰, 沈蒙, 李艳东, 郑宝昆, 毛洪亮, 吴震. 区块链隐私保护研究综述[J]. 计算机研究与发展, 2017, 54(10): 2170-2186 Zhu Liehuang, Gao Feng, Shen Meng, Li Yandong, Zheng Baokun, Mao Hongliang, Wu Zhen. Survey on Privacy Preserving Techniques for Blockchain Technology [J]. Journal of Computer Research and Development, 2017, 54(10): 2170-2186
5	李敏, 孟祥茂. 动态蛋白质网络的构建、分析及应用研究进展[J]. 计算机研究与发展, 2017, 54(6): 1281-1299 Li Min, Meng Xiangmao. The Construction, Analysis, and Applications of Dynamic Protein-Protein Interaction Networks [J]. Journal of Computer Research and Development, 2017, 54(6): 1281-1299
6	王继业, 高灵超, 董爱强, 郭少勇, 陈晖, 魏欣. 基于区块链的数据安全共享网络体系研究[J]. 计算机研究与发展, 2017, 54(4): 742-749 Wang Jiye, Gao Lingchao, Dong Aiqiang, Guo Shaoyong, Chen Hui, Wei Xin. Block Chain Based Data Security Sharing Network Architecture Research [J]. Journal of Computer Research and Development, 2017, 54(4): 742-749
7	陈龙, 管子玉, 何金红, 彭进业. 情感分类研究进展[J]. 计算机研究与发展, 2017, 54(6): 1150-1170 Chen Long, Guan Ziyu, He Jinhong, Peng Jinye. A Survey on Sentiment Classification [J]. Journal of Computer Research and Development, 2017, 54(6): 1150-1170
8	高玉凯, 王新华, 郭磊, 陈竹敏. 一种基于协同矩阵分解的用户冷启动推荐算法[J]. 计算机研究与发展, 2017, 54(8): 1813-1823 Gao Yukai, Wang Xinhua, Guo Lei, Chen Zhumin. Learning to Recommend with Collaborative Matrix Factorization for New Users [J]. Journal of Computer Research and Development, 2017, 54(8): 1813-1823
9	傅艺琦, 董威, 尹良泽, 杜雨晴. 基于组合机器学习算法的软件缺陷预测模型[J]. 计算机研究与发展, 2017, 54(3): 633-641 Fu Yiqi, Dong Wei, Yin Liangze, Du Yuqing. Software Defect Prediction Model Based on the Combination of Machine Learning Algorithms [J]. Journal of Computer Research and Development, 2017, 54(3): 633-641
10	刘洋. 神经机器翻译前沿进展[J]. 计算机研究与发展, 2017, 54(6): 1144-1149 Liu Yang. Recent Advances in Neural Machine Translation [J]. Journal of Computer Research and Development, 2017, 54(6): 1144-1149