信息传播网络推断综述

王宇辰, 高超, 王震

引用本文

王宇辰, 高超, 王震. 信息传播网络推断综述[J]. 计算机科学, 2024, 51(1): 99-112.

WANG Yuchen, GAO Chao, WANG Zhen. Survey of Inferring Information Diffusion Networks[J]. Computer Science, 2024, 51(1): 99-112.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230500127

信息传播网络推断综述

王宇辰[1] 高 超[1] 王 震[1,2]
1 西北工业大学光电与智能研究院 西安 710072
2 西北工业大学网络空间安全学院 西安 710072
(wany810@mail.nwpu.edu.cn)

摘 要 信息的传播扩散可以建模为在潜在传播网络上发生的随机过程。由于在实际应用场景中,潜在的传播网络拓扑结构和清晰的传播过程往往是不可见的,因此根据观测到的传播结果,如节点感染时间、状态等信息,推断传播网络拓扑结构,对于分析与理解传播过程、跟踪传播路径以及预测未来传播事件起着重要作用。近年来,传播网络推断问题吸引了众多研究者的目光。文中对近年来的信息传播网络推断工作进行系统性的介绍和总结,为传播网络推断提供一个新视角。

关键词:传播网络;网络推断;信息传播;信息级联;关系推断
中图分类号 TP181

Survey of Inferring Information Diffusion Networks

WANG Yuchen[1],GAO Chao[1] and WANG Zhen[1,2]
1 School of Artificial Intelligence,OPtics and ElectroNics(iOPEN),Northwestern Polytechnical University,Xi'an 710072,China
2 School of Cybersecurity,Northwestern Polytechnical University,Xi'an 710072,China

Abstract Information diffusion can be modeled as a stochastic process over a network. However,the topology of an underlying diffusion network and the pathways of spread are often not visible in real-world scenarios. Therefore,the inference of diffusion networks becomes critical in the analysis and understanding of the diffusion process,tracking the pathways of spread,and even predicting future contagion events. There has been a surge of interest in diffusion network inference over the past few years. This paper investigates and summarizes the representative research in the field of diffusion network inference. Finally,this paper analyzes the existing problems of diffusion network inference and provides a new perspective on this field.

Keywords Diffusion network,Network inference,Information diffusion,Information cascades,Relationship inference

1 引言

网络作为现实世界中复杂系统的表示方式,不仅能直观地显示实体之间的关系,还蕴含着高阶结构,为生态学[1-2]、神经系统科学[3]、流行病学[4]和传播学[5]等多个领域的分析提供了服务。随着社交媒体平台和用户数量的急速增多,信息的传播、思想的传递在生活中无处不在。在研究信息传播领域相关问题时,研究者通常将信息的传播扩散建模为在一个潜在传播网络上发生的随机过程[6-8],即传播事件从网络中的某一节点开始,沿着边进行随机扩散。因此,清晰的传播网络拓扑结构有助于分析宏观的传播过程,揭示传播动力学规律,并在社会影响力分析(Social Influence Analysis)[9-11]、病毒式营销(Viral Marketing)[12-13]、虚假信息和谣言传播控制[14-17]等应用中起到了至关重要的作用。具体地:1)在社交影响力研究中,可在已知网络拓扑结构的情况下,通过网络中心性(Centrality)衡量用户在社交网络中的重要程度[18];2)在病毒式营销中,利用已知的社交网络,鼓励客户与其朋友分享产品信息,以达到吸引潜在用户的目的;3)在虚假信息、谣言传播控制上,可通过对网络中的关键节点进行阻塞或删除来达到阻断虚假信息传播和减小传播范围的目的[19]。例如,Zhu等[15]对发生在复杂网络上的谣言传播行为进行分析,提出了一种同时包含定向免疫控制和熟人免疫控制的策略来实现谣言的传播控制。

虽然目前有关信息传播的研究大多都是建立在网络拓扑结构已知的情景下,但现实世界中,由于资源和技术等限制,研究者通常无法观测或直接获取完整的信息传播网络,多数观测到的数据都是传播最终产生的结果。因此,传播网络推断旨在根据观测到的传播结果信息推断潜在的传播网络结构,进而服务于传播动力学分析及其相关应用。

信息传播网络推断任务最早可追溯至2005年,Adar

到稿日期:2023-05-01 返修日期:2023-10-07
基金项目:科技部重点研发(2022YFE0112300);国家自然科学基金(62261136549,61976181)
This work was supported by the National Key R&D Program of China(2022YFE0112300) and National Natural Science Foundation of China(62261136549,61976181).
通信作者:高超(cgao@nwpu.edu.cn)

等[20]将其作为一个监督分类问题，结合信息中丰富的文本特征及支持向量机(Support Vector Machines，SVM)来衡量单一链接出现的概率。2010年，Gomez-Rodriguez等[21]和Myers等[22]先后利用信息的级联数据对传播网络推演任务进行探索，引起了各国学者们的广泛关注并取得了系列成果。近年来，还出现了基于节点感染状态等针对特定场景的传播网络推断方法，使推断网络的条件更加灵活，推断网络的准确性得到进一步提升。在以上背景下，文中根据输入数据对现有的信息传播网络推断方法进行分类与分析，为研究人员描述出一个较为清晰的概貌，以期为网络分析、数据挖掘等相关领域的研究提供有益的参考。

本文第2章根据输入数据对信息传播网络推断问题进行分类与分析；第3章对常用的衡量指标进行汇总；第4章介绍了常用的真实世界数据集；第5章总结信息传播网络推断问题的现状并对未来进行展望。

2 信息传播网络推断方法的分类与分析

2.1 信息传播网络结构推断的问题定义

目前，信息传播与扩散动力学的相关问题主要分为3类：预测传播规模问题、预测个体采纳问题，以及传播关系推断问题[23]。前两类问题的目标为通过历史传播数据预测未来信息传播的趋势以及用户未来是否会参与传播过程；而本文所关注的信息传播网络推断为第三类问题，旨在根据观测的历史传播数据推断社交网络中已存在的传播关系，即信息是如何在活跃用户间传播的。因此，信息传播网络推断问题的形式化定义如下：记一个潜在的信息传播网络为有向网络$G^*=(V,E^*)$，其中V是节点集合，E^*是不可见的传播边集合。传播网络结构推断的目标为根据发生在传播网络G^*上的历史传播数据推断隐藏的传播边集合E^*，也即恢复G^*的拓扑结构。

在考虑发生在社交网络上的信息传播过程时，一般认为相同或相似信息的传播只受到“由节点间关系所采取的行动”的影响[24]。进而，共享和参与传播同一种信息的用户构成了一个信息级联，信息级联可包含信息内容、用户的交互行为以及信息转发时间等数据。如今，将多个信息级联组成的集合作为推断传播网络的证据已成为绝大多数方法的选择，但各方法所需输入数据的种类和丰富程度亦有不同。因此，首先根据输入数据，将传播网络推断方法分为3个类别：基于时间序列的方法、基于感染状态的方法，以及结合特征信息的方法。之后，按照建模方式和思路继续对相同输入数据下的不同方法进行细粒度归类，如图1所示。常用符号及其含义说明如表1所列。

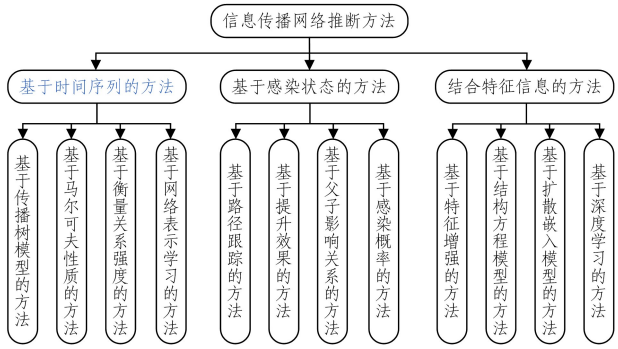

图1 信息传播网络推断方法分类图

Fig. 1 Classification chart of algorithms of inferring information diffusion networks

表1 符号含义表

Table 1 Brief summary of common notations

符号	含义
G^*	信息传播网络
V	节点集合
E^*	参与传播过程的目标边集合
C	级联数据集合
t_i	节点v_i的感染时间
α_{ij}	节点v_j与v_j之间的有向传输率
$\boldsymbol{A}$	传输率矩阵
$\boldsymbol{z}_i$	节点v_i的d维嵌入向量
x_i	节点v_i的感染状态
S	感染状态数据集合

2.2 信息传播模型

信息传播模型作为描述信息如何在社交网络上传播的理论框架，对研究和理解信息传播的过程、机制和效果具有重要意义。信息传播模型主要分为两大类，分别为解释性模型与预测性模型[25]。其中，解释性模型主要以传染病学模型为主，例如SI模型[26]和SIR模型[27]等。这类模型从宏观角度出发描述信息传播的过程，并主要服务于预测信息在群体中的传播规模，分析影响传播率的关键因素等问题。预测性模型则从微观角度出发，对个体间的影响过程展开讨论，在影响力最大化问题[28-29]、预测个体采纳问题[30-31]与传播关系推断等问题中起到至关重要的作用。因此，本文将对预测性模型中的两个经典模型即独立级联(Independent Cascade，IC)模型[32]以及线性阈值(Linear Threshold，LT)模型[33]进行介绍。

IC模型假设每个在t时刻感染的节点只有一次机会在$t+1$时刻以一定的概率去尝试感染它的邻居节点。以图2为例，其中有向边上的权重为节点间尝试感染成功的概率。在时间步t中，深灰色节点b为刚被感染的节点，因此有一次机会分别以0.2和0.6的概率去感染节点a和c，最终成功感染节点c。在时间步$t+1$中，节点b不能再感染其他邻居节点，而节点c尝试感染节点e，以此类推，直至没有节点被感染或所有节点都被感染结束。

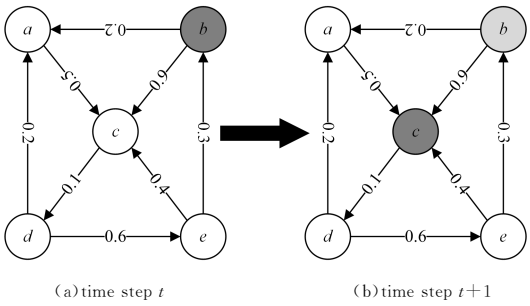

图 2 独立级联模型示意图

Fig. 2 Schematic diagram of independent cascade model

LT 模型假设每个节点v_i都有对应的阈值θ_i（通常范围为[0,1]），阈值θ_i可反映节点受到影响的难易程度。在每一个时间步中，当节点v_i受到被感染邻居节点的影响之和大于阈值θ_i时，v_i将被感染。以图 3 为例，有向边上的权重为邻居节点的影响力，通常邻居影响力总和不大于 1。在时间步 t 中，节点 a 的已感染邻居节点 b 的影响大于阈值θ_a，因此节点 a 会被感染，而节点 c 则不会被感染。

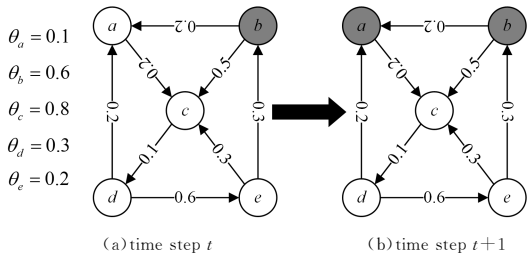

图 3 线性阈值模型示意图

Fig. 3 Schematic diagram of linear threshold model

IC 模型与 LT 模型最大的区别在于随机性：IC 模型的每次感染过程都是随机的，因此会产生完全不同的结果；而 LT 模型的随机性通常表现在节点的边权重以及阈值的确定上，一旦边权重和节点阈值被确定，传播过程即可被预见。这两种模型虽然各自针对信息传播过程提出不同的假设，但都一定程度地反映了信息扩散的特点。后续，以 IC 和 LT 模型为基础，结合更多现实场景中的细节与特征，衍生出一系列拓展模型。例如，引入传输时间延迟的级联传输模型[21]，引入不同用户类别之间影响关系的模型[34]等。

2.3 基于时间序列的传播网络推断方法

感染时间作为反映信息传播动力学的重要因素，在传播网络推断问题中发挥了重要作用。因此，首先对仅利用信息级联中时间信息推断传播网络结构的方法即基于时间序列的方法展开介绍。此类在信息传播结束后收集的节点感染时间集合被研究者称为信息级联的数据。记 M 个级联的时间序列集合为$C=\{c_1,\cdots,c_M\}$，其中级联$c=\{t_1,\cdots,t_{|V|}\}$，t_i为节点v_i对应的感染时间，若未被感染则记为无穷。

基于时间序列的推断方法具有理论和实践意义，在研究初期就受到了广泛关注，并涌现出大量相关方法。本文进一步根据建立模型的思路将相关方法分为 4 类：基于传播树模型的方法、基于衡量关系强度的方法、基于马尔可夫性质的方法，以及基于网络表示学习的方法。

2.3.1 基于传播树模型的方法

IC 模型作为经典的信息传播模型，隐式地模拟了传播发生的过程，并成为推断传播网络相关工作的起点之一。由于 IC 模型是离散时间的模型，并没有考虑感染时间的连续性以及传播的延迟性，因此在 IC 模型基础上，Gomez-Rodriguez 等[21]建立了级联传输模型（Cascade Transmission Model）。在级联传输模型中，若节点v_i成功感染节点v_j，则节点v_j的感染时间为$t_j=t_i+\Delta_{ij}$，其中Δ_{ij}通常称为传输时间（Transmission Time），且一般认为Δ_{ij}服从某种特定的传输时间分布，例如常被用于社交网络扩散过程且具有单调性的指数分布（Exponential Distribution），以及常被用于传染病场景中且具有非单调性的瑞利分布（Rayleigh Distribution）等。级联传输模型产生级联数据的过程如图 4 所示。图 4 中节点 b 为初始感染节点，其分别尝试感染节点 a 和 c，最终成功感染节点 c，节点 c 的感染时间$t_c=t_b+\Delta_{bc}$，其中Δ_{bc}是从传输时间分布中采样的结果。除了引入传输时间这一概念外，级联传输模型的整体传播过程与 IC 模型相近。

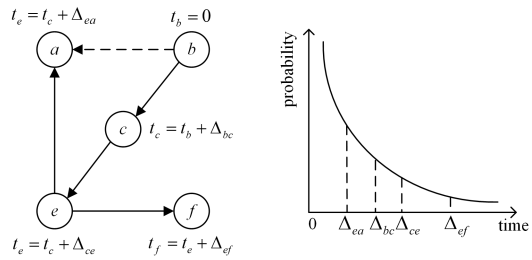

图 4 级联传输模型示意图

Fig. 4 Diagram of a cascade simulated by cascade transmission model

在级联传输模型中，级联数据中被感染的节点均有唯一的传播来源，因此形成级联数据的传播路径可被视为树形结构（被称为传播树 T），如图 5 所示。最早的基于传播树模型的方法是由 Gomez-Rodriguez 等[21]提出的 NetInf 算法。NetInf 算法中提到对于给定网络结构 G 下的一种级联数据，其产生原因是多样且不确定的，所以建立似然函数时需要同时考虑所有传播树的可能性。基于传播树方式建立的似然函数可形式化地表示为：

$$P(c|G)=\sum_{T\in T_c(G)}P(c|T) \tag{1}$$

其中，$T_c(G)$为在级联数据 c 中网络结构 G 的传播树集合。

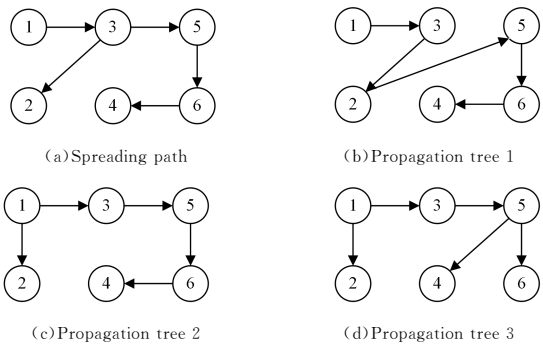

图 5 传播树示意图

Fig. 5 Diagram of propagation trees

对于传播树 T 的似然函数 $P(c|T)$ 可形式化地表示为：

$$P(c|T)=\prod_{(v_i,v_j)\in E_T}\beta P_c(v_i,v_j)\times\prod_{v_i\in V_T,(v_i,v_k)\in E\setminus E_T}(1-\beta) \quad (2)$$

其中，β 为 ICM 中节点尝试感染其邻居节点的概率，$P_c(v_i,v_j)$ 为级联中节点 v_i 与 v_j 之间的传输概率。在仅观测到感染时间的场景下，$P_c(v_i,v_j)$ 为观测到节点之间传输时间差 Δ_{ij} 的概率。虽然 NetInf 提出了对所有可能传播树的建模方式，但为降低求解难度，最终仅利用级联中最有可能的传播树建立似然函数，采用子模优化(Submodular Optimization)提取 k 条边构建推断的传播网络。随后，Gomez-Rodriguez 等继续在 NetInf 的基础上，利用 Kirchhoff's matrix tree 理论，成功解决了在计算级联中全部传播树可能性时产生的指数级时间复杂度的优化问题，进一步提高了网络推断的准确性[35]。

使用贝叶斯推断代替最大似然估计方法，在融入网络结构先验知识以及估计不确定性等方面具有优势。Gray 等[36]将基于传播树建立的似然函数与指数随机图模型相结合，利用马尔可夫链蒙特卡洛(Markov Chain Monte Carlo，MCMC)对网络的后验分布进行采样估计。他们在提出算法的同时，还讨论了在少量数据的情况下使用贝叶斯推断方法来量化不确定性的优势。Ghalebi 等[37]提出一种在线动态网络推断框架 DYFERENCE，并提供了一种非参数的可交换边(Edge-exchangeable)网络模型。DYFERENCE 利用混合的狄利克雷网络分布(Mixture of Dirichlet Network Distributions，MDND)对网络中的动态社区结构进行建模。在每个社区结构中，使用两个狄利克雷分布分别对社区中的输出链接和输入链接进行建模。DYFERENCE 将级联数据按时间间隔划分为不同阶段，并使用一种吉布斯测度[38]计算级联作为传播树 T 的概率，最后使用贝叶斯推断探索不断演变的社区结构并计算潜在传播网络的显式预测分布。Hao 等[39]考虑到传播网络中节点异质性的影响，提出了一种从级联数据中衡量节点重要程度(Importance)的方法。文献[39]假设节点的重要程度会影响传播过程，重要程度高的节点有较大的概率传播和接收信息，因此将节点的重要程度加入到级联数据的似然函数中可以影响贝叶斯推断的采样结果，有效地提高传播网络推断的准确性。

上述对传播树的建模方式默认每个级联只有一个根节点(感染源)。Xu 等[40]提出一种针对传播树建立对数线性模型的算法 DST，打破了以往模型中根节点数量的限制。相较于最大似然的求解思路，DST 提出了一种对比训练(Contrastive Training)方式，利用部分节点感染的时间顺序来代替标记真实边进行训练，无监督地对传播网络结构进行探索。

除了对传播树进行不同方式的建模外，还有部分方法针对已有算法的局限性提出解决方案。例如，Dou 等[41]针对不完整的级联数据，提出 NIIC 算法推断传播网络。NIIC 算法利用蒙特卡洛模拟(Monte Carlo Simulation)对缺失的级联数据进行恢复，再根据恢复后的级联数据最大化基于传播树方式建立的似然函数，最终提取 k 条边构建推断的传播网络。Kong 等[42]针对文献[35]中初始潜在边冗余导致推演速度慢的问题，通过设计一种融合传播时间和状态信息的潜在边评估指标，对初始潜在边进行预剪枝处理，显著提高了网络推演的效率。

2.3.2 基于衡量关系强度的方法

基于传播树建模的方法旨在恢复传播网络的拓扑结构，但其中一个缺陷是无法提供衡量节点间关系强度的依据。为了解决这一问题，以衡量节点间的关系强度为出发点，推断传播网络结构的方法开始出现。

通过估计节点间的感染(传输)概率来反映链接关系强度是最先被考虑到的方式。Myers 等[22]提出一种以节点之间的感染率作为求解目标的算法 ConNIe。ConNIe 分别对感染节点和非感染节点的存在概率建立似然函数，其中节点的感染时间不仅与节点间感染概率关联，同时也考虑了传输时间存在的可能性。最终，ConNIe 提出了一种基于凸优化(Convex Optimization)的最大似然方法，同时加入了类似 l_1 正则化的惩罚项，以使网络结构具有稀疏性。

除了直接求解节点间感染概率的方法外，估计传输时间分布的传输率来反映关系强度也是一种主流方式。记传输率矩阵 $\boldsymbol{A}=R^{|V|\times|V|}$，$\alpha_{ij}\geqslant 0$ 为节点 v_i 与 v_j 间的传输率，传输率通常为传输时间分布的参数，可反映节点之间传输的快慢。Gomez-Rodriguez 等[43]将传播过程建模为在网络上以不同速率发生的连续时间过程，进而提出 NetRate 算法。NetRate 算法利用生存分析理论对级联数据中各个节点的状态建模，利用节点之间的传输似然函数 $f(t_j|t_i;\alpha_{ij})$ 产生生存函数(Survival Function) $S(t_j|t_i;\alpha_{ij})$ 和风险函数(Hazard Function) $H(t_j|t_i;\alpha_{ij})$，这两个函数可分别形式化地表示为：

$$S(t_j|t_i;\alpha_{ij})=1-F(t_j|t_i;\alpha_{ij}) \quad (3)$$

$$H(t_j|t_i;\alpha_{ij})=\frac{f(t_j|t_i;\alpha_{ij})}{S(t_j|t_i;\alpha_{ij})} \quad (4)$$

其中，$F(t_j|t_i;\alpha_{ij})$ 为传输似然函数的累积分布函数。

对于级联中的感染节点 v_j，其似然函数 $P(t_j|c;\boldsymbol{A})$ 是它所有可能发生的感染事件概率之和，每个事件为其可能的邻居节点 v_i 在 t_j 成功感染 v_j，且 v_i 以外的其他邻居未能感染 v_j。该似然函数可形式化地表示为：

$$P(t_j|c;\boldsymbol{A})=\sum_{t_i<t_j}f(t_j|t_i;\alpha_{ij})\times\prod_{t_k<t_j,i\neq k}S(t_j|t_k;\alpha_{kj}) \quad (5)$$

对于未被感染的节点 v_j，则考虑其在观察时间范围 T^c(一般为级联 c 中最大的感染时间)内的存活概率 $P(t_j=\infty|c;\boldsymbol{A})$。其可形式化地表示为：

$$P(t_j=\infty|c;\boldsymbol{A})=\prod_{t_k<T^c}S(t_j|t_k;\alpha_{kj}) \quad (6)$$

最终结合式(5)与式(6)，对级联数据 c 建立似然函数 $P(c|\boldsymbol{A})$。其可形式化地表示为：

$$P(c|\boldsymbol{A})=\prod_{t_i\leqslant T^c}\prod_{t_m>T^c}S(T^c|t_i;\alpha_{im})\times\prod_{t_k<t_i}S(t_i|t_k;\alpha_{ki})\sum_{t_j<t_i}H(t_i|t_j;\alpha_{ji}) \quad (7)$$

NetRate 利用凸优化求解不同节点之间的传输率，将传输率大于一定阈值的节点对视为传播网络中存在的边。Gomez-Rodriguez 等[44]将 NetRate 算法与 l_1 正则化相结合来应对稀疏网络，并进一步讨论了在该方法下推断传播网络所需的级联数量。凸优化虽然能精确地计算不同节点之间的传输率，但需要大量的计算和时间开销，难以快速地应对动态演化的网络。因此，Gomez-Rodriguez 等[45]通过对级联数据抽样和利用投影随机梯度(Projected Stochastic Gradient)方法，来大幅提高网络推演的效率，有效地解决了大规模动态网络的

推断问题。Du 等[46]考虑到节点间传输的异质性，即节点间的传输时间服从多种未知分布，提出了 KernelCascade 算法。KernelCascade 通过对生存分析中的风险函数进行核化（Kernelizing），解除了以往方法中对节点间传输似然函数的限制，利用块坐标下降（Block-coordinate Descent）方法求解节点之间的传输率。Wang 等[47]考虑多模式（主题）级联下的网络推断问题，并提出了 MMRate 算法。他们对真实世界社交媒体中的大量级联数据进行分析发现，节点间的传输模式在不同主题下不尽相同。因此，MMRate 将不同主题 k 下节点间的传输率α_{ij}^{k}和级联的扩散模式参数δ^{c}结合为 $g(\alpha_{ij}^{k},\delta^{c})$，并将其作为传输似然函数 $f(t_j \mid t_i; g(\alpha_{ij}^{k},\delta^{c}))$中的参数，最终利用期望最大化算法（Expectation Maximization，EM）推断网络结构，以及节点间在不同主题下的传输率。Tan 等[48]将基于生存分析的方法与网络基元（Network Motif）知识相结合，并提出了 MANDI 算法。网络基元是复杂网络中频繁出现的子图模式，是复杂网络的“构建块”。MANDI 算法根据级联数据挖掘在传播网络中可能频繁出现的基元结构，并将它们作为正则先验影响网络推断的结果，然后根据网络的推断结果对候选的基元结构进行更新。基元结构挖掘和网络推断两者相互影响且持续更新，直至算法收敛。

在基于传播树模型和基于衡量关系强度的方法中，还存在一些方法通过挖掘时间序列中节点感染时间所暗含的强度关系来提升已有模型的推演表现。例如，Zhao 等[49]指出整个传播过程可划分为不同阶段，忽略不同阶段影响强度的异质性，会导致推演结果不准确。因此，他们提出了 LSH 方法将整个时间序列缩放至[0，1]范围并等分为不同阶段，根据不同阶段内新增感染节点的数量体现影响强度的异质性，进而提高现有方法包括 NetInf，ConNIe 和 NetRate 的推演表现。Gao 等[50]通过实验发现节点对在级联中互动的归一化概率能一定程度地反映边出现的概率，进而提出 PBI 方法。PBI 将互动的归一化概率转为权重融入以生存分析理论建模的似然函数中，以提高网络推演的准确性，并利用贝叶斯推断对潜在边采样，在有限的迭代次数下完成网络的推演。

2.3.3 基于马尔可夫性质的方法

马尔可夫性质指在一个随机过程中，当前状态只与前一状态有关，与更早的状态无关。基于马尔可夫性质的方法通常根据“节点当前状态的改变受到之前邻居节点状态的影响”这一观点，使用状态转移概率矩阵描述节点状态改变的随机过程。

Eslami 等[51]将网络上的信息扩散过程建模为一个马尔可夫随机游走（Markov Random Walk）过程，并提出了 Dne 算法。Dne 根据节点被感染的顺序，将所有可能存在的感染关系构建为一个初始图。在随机游走的每一个离散时间步中，已感染的节点都有一个与初始图相关联的转移概率矩阵，用于描述节点之间的感染关系。通过借鉴马尔可夫随机游走中的到达时间（Hitting Time）这一概念，Dne 定义了一个参数 RT 来反映边存在的概率，RT 越小表明边存在的概率越大，最终提取 k 条最小 RT 的边构建传播网络。Ramezani 等[52]考虑到社区结构对推断过程的影响，提出了 Dani 方法。Dani 利用转移概率矩阵描述传播的随机过程，并加入 Jaccard 系数衡量节点在网络结构中的社区关联性。Dani 使用节点感染的次序取代具体的感染时间评估传输概率，使算法不受特定传输时间分布假设的限制，同时该算法还具有运行速度快的优点。Foroutain 等[53]利用马尔可夫决策过程（Markov Decision Process，MDP）模拟信息传播过程。对于每一次传播，从初始状态开始，代理（Agents）找到可能的行动，选择其中一个执行并更新状态，之后继续重复执行这 3 个步骤，直至到达最终状态。在可能的状态和行动已知下，通过强化学习求解节点间的转移概率，同样取得了有效的推断结果。Crawford[54]在假设节点v_i在网络中的度d_i已知的场景下，提出了一种对传播网络的重构方法。文献[54]将传播过程建模为一个连续时间的马尔可夫模型，将级联的似然函数解释为指数随机图模型（Exponential Random Graph Models，ERGMs），最终利用贝叶斯推断估计网络中潜在边的后验概率。

Tahani 等[55]指出推断动态传播网络更具有现实意义和挑战性，并针对该问题提出了 DDNE 算法。DDNE 利用隐马尔可夫模型（Hidden Markov Model，HMM）将网络中边的存在建模为一个随机过程。DDNE 根据在同一时间段内的节点感染次序定义各个边存在可能性的变量，并通过当前时间段内边的状态以及可能性大小预测下一时间段内边的存在概率。最终，选取存在概率超过人工设置阈值的边构建传播网络。

2.3.4 基于网络表示学习的方法

嵌入指将对象转化为低维稠密的实数向量来存储其蕴含的特征信息。常见的概率模型将网络的邻接矩阵 $\boldsymbol{A} \in R^{|V|\times|V|}$ 作为求解目标，虽然求解目标简单直接，但邻接矩阵与节点数量$|V|$呈指数关系，难以应对大规模的网络。因此，研究者利用表征学习领域提供的压缩能力对网络中的实体进行嵌入，来应对更加复杂的场景。

利用级联数据对网络中的节点进行嵌入是一种直观的方式。记节点v_i的 d 维嵌入向量为$\boldsymbol{z}_i \in R^{d}$，基于节点嵌入方法的核心思想是利用嵌入向量在低维空间中的距离来反映节点之间存在边的概率，可形式化地由图 6 和式(8)表示。

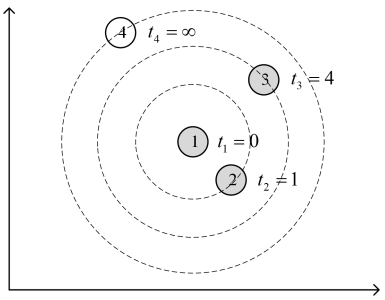

图 6 节点嵌入方法示意图

Fig. 6 Diagram of a node-embedding method

$$P(v_i,v_j)\propto|\boldsymbol{z}_i-\boldsymbol{z}_j|^{2} \tag{8}$$

Bourigault 等[56]将级联数据中的感染事件映射到一个连续的潜在空间上，并提出了 CDK 算法。CDK 算法使用热扩散核（Heat Diffusion Kernel）反映节点在传播过程中被感染概率随时间的变化，然后根据感染时间的先后顺序对所有节点的嵌入向量提出了限制关系，使用蒙特卡洛模拟更新节点

的嵌入向量。Hu 等[57]基于 CDK 的思想，并简化其目标函数，提出了 P-CENI 的预剪枝方法，通过提前排除未参与传播过程的边来提高传播网络推断方法的效率。Bourigault 等[58]在 ICM 的基础上提出 Embedded IC 算法。Embedded IC 算法假设每个节点均有接收者和发送者两种身份，采用 sigmoid 函数建立信息发送者和接收者之间的传播概率，最后利用蒙特卡洛模拟分别更新节点作为发送者和接收者身份的嵌入向量。Kurashima 等[59]基于生存分析的建模方式提出了 PLNV 算法。PLNV 算法将节点之间传输似然函数中的传输率替换为两个节点嵌入向量在空间中的距离，使用梯度下降对每个节点的坐标进行更新，获得一种可以在低维空间衡量和可视化网络结构的方式。Zhang 等[60]使用 Wald 分布作为全局传输时间分布，利用节点嵌入向量间的距离作为分布中的参数，提出了 COSINE 算法。COSINE 算法使用高斯混合模型(Gaussian Mixture Model，GMM)作为用户嵌入向量的先验，以保持嵌入向量与社区结构之间的相关性，使用 EM 算法分别对节点的嵌入向量和社区结构进行更新和探索。

Zhuo 等[61]指出上述对节点表征学习的方法在训练中只考虑在级联数据中的节点，尽可能使级联中的节点在潜在空间中相互靠近，而忽略了不在级联中的节点。为了解决这一问题，他们提出了 DiffusionGAN 算法，利用生成式对抗网络(Generative Adversarial Networks，GAN)对级联中的用户进行表征学习，区分每个用户的表示，以确保在同一级联中的用户在嵌入空间中彼此接近，而在不同级联中的用户彼此远离。在 DiffusionGAN 算法中，生成器(Generator)试图生成用户的嵌入向量，以匹配传播级联数据中的真实用户分布；鉴别器(Discriminator)试图区分所述采样用户是否来自真实值。生成器和鉴别器进行博弈论的极大极小博弈来相互优化。

对网络中的边进行嵌入也是一种有效的方式。边嵌入的方法将节点对(v_i, v_j)视为个体，利用它在不同级联中的时间差集合 $D=\{|t_i^k-t_j^k|\}_{k=1}^{M}$ 对其进行嵌入。在获取所有节点对的嵌入向量后，利用聚类方法对节点对集合进行二分类，并使用平均时间间隔小的一类节点对构建的传播网络。

Rong 等[62]通过实验发现传播节点对和非传播节点对之间平均时间间隔分布不同的统计特征，提出了零模型(Null Model)的 NPDC 算法。如图 7 所示，虽然参与传播的节点对的平均时间间隔更小，但另一方面，时间间隔过小的节点对也可能由于传播过程中的时间延迟效应而不存在传播关系。基于此发现，NPDC 算法假设传播节点对和非传播节点对分别有各自的平均时间间隔分布，且传播节点对的平均时间间隔分布的平均值低于非传输节点对的分布，利用两种分布之间在统计学上的差异以及通过衡量节点对的时间间隔集合与假设时间间隔分布之间的距离，对节点对的嵌入向量进行更新。NPDC 无须提前假设节点之间的传输似然函数，在无模型的假设下推断传播网络，易于在真实世界场景下使用。Hu 等[63]同样根据传播节点对与非传播节点对在时间间隔分布上具有显著差异这一特性，利用再生希尔伯特空间(Reproducing Kernel Hilbert Space，RKHS)的嵌入方法进一步提升推断表现。

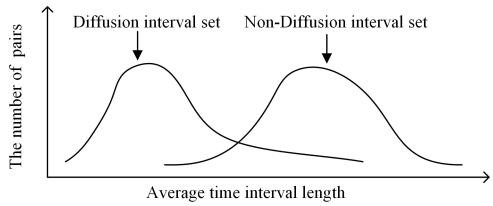

图 7 传播节点对与非传播节点对的平均时间间隔比较示意图

Fig. 7 Diagram of average time interval comparison between diffusion node pairs and non-diffusion node pairs

2.3.5 总结

基于时间序列的传播网络推断方法作为最早被研究的主流方法，已取得了不俗的进展。本文所归纳的 4 类方法也各有优势和劣势。基于传播树模型的方法聚焦恢复网络拓扑结构，但忽略了节点间联系的强度，同时以最大似然估计或贝叶斯推断作为求解手段，因此运行效率并不高。基于衡量关系强度的方法以凸优化作为求解方式需要大量的计算资源，并且难以规避过拟合问题[64]。基于马尔可夫性质的方法通常假设传播过程是简单且独立的，难以更好地体现传播过程中的复杂性和不确定性，并且在计算概率转移矩阵时，计算空间和时间复杂度都与节点数量的平方成正比。基于网络表示学习的方法将网络的实体(多数方法为节点)嵌入到一个固定维度的潜在空间中，在有限空间内可应对更大规模的网络结构，并且表征结果可服务于其他任务中，例如预测未来可能的感染事件等。此外，前 3 种方法在建模过程中都遵循着严格的假设，例如节点间的传输时间服从特定分布，节点只能由一个父节点影响等。而基于网络表示学习的方法通常对模型的假设较为宽松，但这也意味着基于网络表示学习的方法需要从更多的数据中挖掘潜在的规律，因此只利用时间序列数据或数据的数量和质量不足都可能使算法无法取得较好的表现[23]。

2.4 基于感染状态的传播网络推断方法

虽然使用级联数据推断传播网络取得了阶段性的进展，但在真实世界场景中精确及时地获取时间序列数据依旧需要消耗大量的人力与物力资源。因此，使用更易获取的节点感染状态数据推断传播网络成为研究者的新的研究方向，并涌现出了一系列基于感染状态的传播网络推断方法。

感染状态数据指在网络上发生传播后，记录各个节点的状态集合。记 M 次传播后的感染状态数据集合为 $S=\{s_1,\cdots,s_M\}$，第 i 次传播产生的感染数据为 $s_i=\{x_1,\cdots,x_{|V|}\}$，其中 $x_i=1$ 表示节点 v_i 被感染，$x_i=0$ 表示其未被感染。相较于时间序列数据带来的节点感染次序和节点传输时间间隔等利于缩小推断关系范围的信息，感染状态数据的无序性使推断传播网络在理论上更为困难。目前，根据建立模型的角度和方式，可将基于感染状态的方法进一步分为基于路径跟踪的方法、基于提升效果的方法、基于父子影响关系的方法以及基于节点间感染概率的方法。

2.4.1 基于路径跟踪的方法

研究早期，Gripon 等[65]提出一种在传播路径中只有 3 个节点的场景下推断传播网络的算法。文献[65]提出，如果在一个传播路径中有且仅有 3 个任意节点 v_i，v_j 和 v_k 处于感染

状态，则无向边(v_i,v_j)和(v_i,v_k)中至少有一个是真实参与传播的边。基于此引理，文献[65]利用传播路径中节点对的出现频率推断传播网络。该方法虽然提供了严谨的数学理论基础并具有良好的运行效率，但对输入数据的格式要求较为严格，难以适用于实际场景中。

2.4.2 基于提升效果的方法

后续，Amin 等[66]基于感染状态数据定义了一种提升效果 $L(v_i|v_j)$来反映当一个节点v_j为感染源时对节点v_i感染状态的解释程度，并根据所有潜在边的提升效果排序结果构建传播网络。但该算法需要人工设置边的数量作为算法的停止条件，同时需要获取各个传播过程的起始节点，因此具有一定的局限性。

2.4.3 基于父子影响关系的方法

Huang 等[67]进一步地考虑传播网络中父子之间的影响关系，并提出了 TWIND 算法。TWIND 算法根据“节点只能被其父节点感染”这一性质，将所有节点状态的联合概率转换为多个相互独立的基于父节点状态集合的条件概率乘积，可形式化地表示为：

$$P(x_1,\cdots,x_{|V|})=\prod_{n=1}^{|V|}P(x_i|x_{F_i}) \tag{9}$$

其中，F_i是节点v_i的父节点集合，x_{F_i}是F_i的状态。TWIND 算法对每一个节点启发式地搜索最大化 $P(x_i|x_{F_i})$的父节点集合，根据感染状态数据的数量估计父节点数量的理论上限作为算法的停止条件。Huang 等[68]基于式(9)构建了基于相对熵的概率生成模型并提出 SIDN 算法。SIDN 算法利用 KL 散度(Kullback-Leibler Divergence)衡量不同父节点集合概率分布之间的差距，并证明条件熵 $H(x_i|x_{F_i})$与条件概率 $P(x_i|x_{F_i})$成反比，对每一个节点启发式地搜索最小化条件熵的父节点集合构建传播网络。Han 等[69]对节点之间的父子关系定义一种新的评分机制并提出了 TENDS 算法。该评分机制由两部分组成：一部分为节点v_i在父节点F_i下的似然函数；另一部分是关于父节点数量的惩罚函数，以防算法过度地加入父节点。为了快速推断网络结构，TENDS 算法将衡量节点之间相关性的互信息(Mutual Information，MI)指标修改为更符合传播数据性质的 IMI 指标，结合 K-means 聚类算法对潜在边进行预剪枝。Gan 等[70]提出在不完整的感染状态数据下推断传播网络的框架 POIND。该框架分为两步：1)根据观测的部分数据估计缺失的节点感染状态概率以及节点和其可能父节点集合之间的感染概率，并利用已知的算法如 TWIND 推断传播网络；2)根据推断的传播网络结构，利用类 EM 算法更新节点与可能父节点集合之间的感染概率。POIND 框架使基于感染状态数据推断传播网络的方法在现实世界中更具有可行性。

2.4.4 基于节点间感染概率的方法

上述方法从父子节点间影响关系的角度出发，探索各节点在观测数据下最有可能的父节点集合，进而构建传播网络，但仍然不能衡量节点间的关系强度。因此，Sun 等[71]从节点间感染概率的角度出发，提出了 FINITI 算法。FINITI 算法与 ConNIe 算法的建模思路相似，都构建了以节点间感染传播概率为变量的似然函数。但感染状态数据体现的感染事件会相互冲突(由于无法确定感染的次序，因此两个有关联的节点都有可能是对方的父节点)，导致似然函数具有非凸性，因此 FINITI 采用 EM 算法作为优化算法。实验表明，与前 3 类方法相比，该方法在提高传播网络推断准确性的同时，不仅可以评估节点间的关系强度，还具有良好的运行效率。

2.4.5 总结

尽管基于感染状态数据推断传播网络结构更具实际应用性，但其同样存在困难和挑战。早期的研究方法更适用于特定场景，例如可追踪路径或已知感染源的信息。当前，基于父子影响关系的推断方法已成为主流方法之一，同时实验证明在多感染源的场景下，这类方法的推断表现可以达到和基于时间序列的方法相近甚至更好的表现。然而，与时间序列数据不同，感染状态的无序性会使推断范围更广泛，导致算法时间复杂度较高。即使使用聚类方法对候选边进行预剪枝处理以缩小推断范围，该算法也难以应对大规模网络结构的推断。此外，基于节点间感染概率的方法容易受到优化过程中过拟合问题的影响，并且依赖于预剪枝的结果。

2.5 结合特征信息的传播网络推断算法

除了信息传播过程产生的感染事件外，信息内容本身以及用户特征等额外可挖掘的数据特征同样可提供有利于推断网络结构的证据。本文根据建立模型的方式以及思路，将这一部分结合特征信息的传播网络推断方法分为基于特征增强的方法、基于扩散模型嵌入的方法、基于结构方程模型的方法以及基于深度学习的方法。

2.5.1 基于特征增强的方法

基于特征增强的方法通常将除感染时间和感染状态外的信息特征融入现有成熟的概率模型中，以提高传播网络推断的准确性。例如，在 DST 模型中融入节点特征和文本内容特征，进一步提高了算法在真实世界数据集中的表现[40]。

在利用节点的特征信息时，研究者一般假设用户具有不同的传播倾向。例如在社交网络中，共同爱好交集多的两个用户间更有可能存在传播关系。Wang 等[72]基于此假设，利用特征信息增强概率模型，提出 MONET 算法，如图 8 所示，特征增强后的概率模型不再将感染时间作为计算节点间传输概率的唯一依据。MONET 算法在根据生存分析理论构建级联数据的似然函数时，融入节点的特征信息，进一步提升了推断网络的准确性。

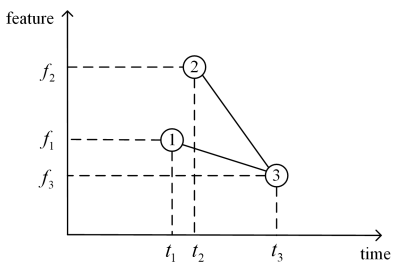

图 8 特征增强的概率模型示意图

Fig. 8 Diagram of a feature-enhanced probabilistic model

除了节点特征外，研究者认为传播内容的特征也会影响传播过程。Du 等[73]针对主题敏感的信息传播网络，提出了 TOPICCASCADE 算法。对主题敏感的信息传播网络，指在

该网络上不同主题的信息的扩散速度不同。TOPICCASCADE算法考虑到普通的传输时间分布仅利用了节点间的感染时间因素而未考虑到主题对传输速率的潜在影响，将节点间的传输率表示为多个主题下的传输率总和，可形式化表示为：

$$\alpha_{ij}=\sum_{l} m_l \alpha_{ij}^{l} \tag{10}$$

其中，α_{ij} 为节点 v_i 和 v_j 之间的传输率，m_l 为主题 l 的权重，α_{ij}^{l} 为在主题 l 下节点 v_i 和 v_j 之间的传输率。TOPICCASCADE算法基于生存分析对级联数据建模，利用块坐标下降法求解节点间的传输率。

Zhou 等[74]同时结合了拓扑特征、节点属性和信息内容来推断节点间的传输率，提出了 NIMFC 算法。NIMFC 算法提取信息文本中的主题特征，考虑节点和级联数据各自的主题分布，将节点间的相似度和节点与级联信息间的相似度融入以生存分析建模的似然函数中，最后通过最大化似然函数求解传输率矩阵。Manco 等[75]提出了一种结合传播动力学和传播内容主题的传播网络推断方法。文献[75]将各主题下节点间的传输率分解为两部分，分别为节点在主题上的权威程度和敏感程度。节点的权威程度反映其影响其他节点的可能性；敏感程度反映它接受传播内容的可能性。基于此分解方式，该方法建立了一种概率图模型来阐述主题影响传播过程的方式，最终通过 EM 算法分别求解权威程度矩阵、敏感程度矩阵以及主题的分布参数。

2.5.2 基于结构方程模型的方法

结构方程模型(Structural Equation Modeling，SEM)是一种基于统计学的多变量分析方法，可用于探究多个变量之间的因果关系[76]。

Baingana 等[77]采用动态 SEM 来捕获观察到的感染时间与边的未知权值之间的关系。他们假设网络拓扑结构会随时间产生变化，但在一段时间内保持固定。在第 t 时间段内，节点的感染时间可建模为一个动态线性 SEM，如式(11)所示：

$$y_{ic}^{t}=\sum_{j\neq i} a_{ij}^{t} y_{jc}^{t}+b_{ii}^{t} x_{ic}^{t}+e_{ic}^{t} \tag{11}$$

其中，y_{ic}^{t} 是节点 v_i 在第 t 时间段内的感染时间，a_{ij}^{t} 是边 (v_i, v_j) 上的权重，x_{ic}^{t} 是节点 v_i 对感染的易感性(假设是已知的)，b_{ii}^{t} 获取外部感染源的影响力，e_{ic}^{t} 是对测量误差和未建模的动力学进行解释。因此，节点的感染事件可被划分为 $\sum_{j\neq i} a_{ij}^{t} y_{jc}^{t}$ 的外因(被其他节点感染)以及 $b_{ii}^{t} x_{ic}^{t}$ 所代表的内因(自身易感性)两方面。此外，Baingana 等还进一步提出了一种在计算复杂度较高时有效的随机梯度算法以求解参数。后续，Baingana 等[78]指出网络拓扑有时可能在有限数量的离散状态之间“跳转”，进而表现为级联行为的突然变化。例如，电子邮件网络可能会在周末从主要的基于工作的拓扑结构切换到基于朋友的拓扑结构。针对这种现象，他们提出了一种可切换的动态 SEM 来捕获与网络拓扑结构相关的级联演化，以及拓扑结构的离散状态。

虽然 SEM 能够合并外部影响来解决内在的“directional ambiguities”问题，然而传统的 SEM 假设有外部影响因素的全部知识，在实际环境中缺乏可行性。对此，Shen 等[79]利用并行因子分解(Parallel Factor Decomposition)求解各个节点之间的传输率。

2.5.3 基于扩散嵌入模型的方法

以数据作为驱动的基于扩散模型嵌入的方法同样取得了不俗的表现。不同于上文提及的基于特征增强的方法依旧对模型进行较为严格的假设，基于扩散嵌入模型的假设会更为宽松，倾向于从数据中学习到潜在规律。而与基于时间序列数据中的网络表示学习方法相比，结合特征信息的扩散模型嵌入方法通常会将感染时间、传播内容、用户特征等信息共同作为扩散模型的组成部分，因此通常具有更好的推断表现。

Bourigault 等[56]在利用时间序列数据提出一种将节点嵌入到潜在空间的 CDK 算法后，进一步地考虑到文本内容对传播的影响，即每个可能的信息内容将对应于潜在空间中的一个特定度量，从而产生不同的传播方案。他们通过在扩散核中加入文本内容 q 的嵌入向量 $\boldsymbol{f}(q)$，改变感染源在潜在空间中的位置，提出了 CSDK 方法。如图 9 所示，作为感染源的节点 1 在不同级联中受文本特征 $f(q_1)$ 与 $f(q_2)$ 的影响，感染源的位置发生了改变，虽然其他节点在潜在空间中的位置是固定的，但最终导致两个级联观测到的时间序列结果完全不同。

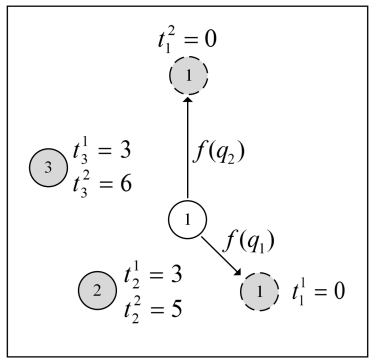

图 9 CSDK 方法示意图

Fig. 9 Diagram of CSDK

Gao 等[80]考虑到信息级联中的内容特征对传播过程的影响，提出了 IEDP 算法。IEDP 算法根据用户向量和传播内容向量在嵌入空间中的几何关系构建有关传播的评分函数 d_{ij}，可形式化地表示为：

$$d_{ij}=\| \boldsymbol{z}_i+\boldsymbol{p}_c-\boldsymbol{z}_j \|_{l_2} \tag{12}$$

其中，$\boldsymbol{p}_c$ 为传播内容的特征向量，$\boldsymbol{z}_i+\boldsymbol{p}_c$ 为用户 v_i 传播 $\boldsymbol{p}_c$ 在嵌入空间的位置。$\boldsymbol{z}_i+\boldsymbol{p}_c$ 与用户 v_j 的嵌入向量 $\boldsymbol{z}_j$ 距离越近，代表用户 v_i 与 v_j 之间有更高的概率存在传播内容 $\boldsymbol{p}_c$ 的关系。利用此评分函数，IEDP 算法根据感染时间构建目标函数，并利用随机梯度下降更新节点的嵌入向量。相较于基于显式推理的模型，该算法大大降低了计算复杂度，为处理海量数据提供了一种可靠的方法。

2.5.4 基于深度学习的方法

基于深度学习的传播网络推断方法通常利用部分已知的传播网络结构对分类器进行监督与修正，训练结束后对其余所有可能的边进行判别，最终利用其中判别为参与传播的边构建传播网络结构。近年来，由于深度学习在分类任务上取得了不俗的成绩，研究者开始考虑将深度学习应用到传播网络推断中。Wang 等[81]提出传播网络推断存在两个难点：一是生成传播数据的传播路径不唯一；二是节点之间的传播

过程可能存在延迟，导致推断结果不准确。因此，文献[81]提出了 FNI 算法，利用节点特征和深度学习方法推断传播网络。通过任意节点间特征值的共性来映射它们之间存在传播关系的可能性，映射方式为寻找两个节点相同的特征并在特征向量的对应位置设置为 1，通过深度神经网络（Deep Neural Networks，DNN）对节点对进行分类，进而构建传播网络。

Yang 等[82]基于推荐系统提出了 DIM-SPTF 算法，用于推断参与传播过程的社交网络结构。DIM-SPTF 算法将多种类信息，包括用户接收传播内容的时间信息、用户的历史偏好特征信息以及用户传播内容的文本特征信息，作为节点对构建输入向量。同时，DIM-SPTF 算法利用循环神经网络（Recurrent Neural Network，RNN）对节点对进行分类并构建传播网络。

2.5.5 总结

在当今大数据时代下，收集丰富的数据特征为更精确地推断传播网络结构提供了可能性。上述 4 种方法分别从不同角度通过不同方式对信息特征进行融合与利用。基于特征增强的方法通常遵循着基于时间序列方法的假设，并在概率模型中融入节点和传播内容的特征以提高推演准确性。基于结构方程模型的方法通常从内因和外因两个方面分析感染结果产生的过程，但其同样对假设要求非常严格，包括模型的线性关系、正态性、独立性等，如果假设不成立，模型的拟合效果和推断结论可能会失真。同时，算法的计算复杂度较高，难以应对大规模的网络结构。与仅利用时间序列数据的网络表示方法不同，基于扩散模型嵌入的方法会将信息内容对传播过程的影响纳入信息传播模型中，不仅可以通过节点在嵌入空间中的距离量化用户的关系，还可以嵌入不同类型的扩散信息内容，以捕获级联中用户的不同偏好，有利于迁移到不同的传播环境中。但这也意味着基于扩散模型嵌入的方法对数据本身和提取的质量有着更高的要求，否则可能会适得其反。例如在文献[56]的部分对比实验中，利用文本内容的 CSDK 算法表现不如仅利用时间序列数据的 CDK 方法。基于深度学习的方法通常需要部分已知的网络结构对分类器进行训练，使得适用场景更为有限，并且可迁移性不高，对不同类型的传播网络需要重新训练，成本较高。

3 性能评价指标

本章总结了传播网络推断算法常见的性能评价指标以供研究者参考。

记真实参与传播的网络为$G^*=(V,E^*)$，算法的推断结果为$\hat{G}=(V,\hat{E})$，真实的传输率矩阵为$\boldsymbol{A}^*$，矩阵$\boldsymbol{A}^*$中节点v_i与v_j之间的传输率为α_{ij}^*，算法推断的传输率矩阵为$\hat{\boldsymbol{A}}$，矩阵$\hat{\boldsymbol{A}}$中节点v_i与v_j之间的传输率为$\hat{\alpha}_{ij}$。

评估指标可分为 3 类，分别为评估传输率的指标、评估拓扑结构的指标，以及评估概率的指标。评估传输率的指标有均方根误差（Mean Squared Error，MSE）和均绝对误差（Mean Absolute Error，MAE）。评估拓扑结构准确性的指标有准确率（Accuracy）、精确率（Precision）、召回率（Recall）和 F-分数（F-score）。评估基于概率的分类器指标有 AUC 值，即 ROC 曲线下与坐标轴围成的面积。评估性能指标的总结如表 2 所列。

表 2 常用性能评估指标总结

Table 2 Summary of evaluation metrics

评估指标	公式	描述	文献
Mean Squared Error(MSE)	$\frac{1}{\lvert V\rvert^2-\lvert V\rvert}\sum_{i,j}(\alpha_{ij}^*-\hat{\alpha}_{ij})^2$	测量值误差的平方和取平均值	[22,45,53,58,71,79]
Mean Absolute Error(MAE)	$\frac{1}{\lvert V\rvert^2-\lvert V\rvert}\sum_{i,j}\lvert\alpha_{ij}^*-\hat{\alpha}_{ij}\rvert$	测量值误差的绝对值取平均值	[43,47,73-74]
Accuracy	$1-\frac{\sum_{i,j}\lvert I(\alpha_{i,j}^*)-I(\hat{\alpha}_{i,j})\rvert}{\sum_{i,j}I(\alpha_{i,j}^*)+\sum_{i,j}I(\hat{\alpha}_{i,j})}$	通过传输率评估推断的传播网络的准确性	[35,43,45,81-82]
Precision	$\frac{\lvert E^*\cap\hat{E}\rvert}{\lvert\hat{E}\rvert}$	真实边在推断边中的比例	[21-22,35-37,40-41,45-46,48-49,56-59,63,72,81-82]
Recall	$\frac{\lvert E^*\cap\hat{E}\rvert}{\lvert E^*\rvert}$	推断正确的边在所有真实边中的比例	[21-22,35-37,40,45-46,48-49,56-59,63,72,81-82]
F-score	$\frac{2\times Precision\times Recall}{Precision+Recall}$	同时兼顾 Precision 和 Recall 的指标	[37,41-42,44,46-53,55,57-59,67-75,82]
Area Under Curve(AUC)	—	衡量整个 ROC 曲线下的整个区域面积	[21-22,36,39,49,54,60-61]

4 人工网络和真实网络数据集

本章收集了常用于评估传播网络推断方法的人工网络模型和真实世界数据集以供研究者参考。

4.1 人工网络模型

为了验证和探索算法在不同规模、不同特征网络上的推断表现，研究者通常利用人工网络模型生成可控的网络结构，在人工网络上利用传播模型仿真传播数据作为传播网络推断算法的输入。常见的人工网络模型有 ER 随机图（Erdős-rényi Random Graphs）模型[83]、核心-边缘网络（Core-periphery Network）模型[84]、分层社区网络（Hierarchical Community Network）模型[85]、森林火灾（Forest Fire）模型[86]以及 LFR 基准图（LFR Benchmark Graphs）模型[87]。

ER 随机图的特点是每一对节点都以一定的概率相连；核心-边缘网络的特点是存在两个不同的组成部分，一部分是紧密连接的“核心”节点，另一部分是稀疏的“外围”节点，“核心”节点与“外围”节点之间存在着稀疏的连接；分层社区网络模型的特点是具有层次化社区结构。上述 3 种网络都可由

克罗内克图(Kroncker Graphs)模型[88]通过不同的参数生成。Leskovec 等[89]观察到真实世界中的网络结构存在两种现象:会随时间推移密集化;节点之间的平均距离会随着时间推移缩小。为了仿真这两种现象,文献[89]基于森林火灾模型提供了一种网络生成器。LFR 基准图模型考虑节点度分布和社区规模分布异质性这两种特征,仿真真实世界中具有社区结构的网络,可用于评估算法对具有社区性质的网络的推断效果。本文进一步地总结了文中提及的算法所用的人工网络模型,如表 3 所列。

表 3 人工网络模型总结

Table 3 Summary of synthetic network models

模型	文献
Erdős Rényi random graph	[21-22,35-36,41-44,46-48,50-51,53,62,65-66,73-74]
Core-periphery network	[21,35-37,41-43,45-48,50-51,53,55,62-63,73-74]
Hierarchical community network	[21,35-37,41-44,45,47-48,51,53,55,62-63,73-74]
Forest Fire model	[21,36-37,43-44,48,51,55,62]
LFR benchmark graphs	[50,52,60,67-71]

4.2 真实世界数据集

真实世界数据集可分为两种类型。

第一类为通过在真实世界中存在的网络上利用传播模型仿真得到感染数据,进而验证算法的表现。本文收集和总结了常见的真实世界网络结构,如表 4 所列。

表 4 常用真实世界网络结构总结

Table 4 Summary of real-world networks

网络	$\|V\|$	$\|E\|$	文献
Scientists collaboration network	1461	2742	[22,50-52,66-70]
Email communication in European	986	16064	[22,36,50]
Adolescent health	2539	12969	[42]
Polblogs	1490	19025	[42]
Highschool	327	5818	[39]
DUNF	750	2974	[67-70]
Football network	115	615	[51]

第二类是根据真实世界中发生的信息传播数据,例如博客的转发和社交平台的订阅等,推断传播网络。例如,Memetracker 数据集收集了来自 Memetracker 的 9600 万条博客信息。此类数据集有两种基于传播数据构建基准传播网络的方式。1)以站点的超链接为基础的构建方式。假如发布信息的站点 P 引用了站点 L 的超链接,则在站点 P 与站点 L 之间构建传播链路,并继续回溯站点 L 引用的超链接,构建完整的传播链路。如果站点 L 是第一个发布这个信息的站点,则它为该信息的传播源。2)以传播信息内容为基础的构建方式,将网站中发布的内容解析成多个短语(模因),寻找发布过相同或近似词语的站点,并根据发布的时间构建传播链路。本文总结了常见的真实世界信息传播数据集,如表 5 所列。

表 5 常用真实世界信息传播数据集总结

Table 5 Summary of real-world information diffusion datasets

数据集	描述	数据链接	文献
Memetracker	一个包含数以百万计博客和新闻文章的数据集,可用于追踪信息传播链路	https://snap.stanford.edu/data/memetracker9.html	[21,35,37,41,43,45-46,48-49,52-53,56-62,73,75,80]
Twitter	一个来自 Twitter 平台的数据集,收集了该平台上推文和其相应转发的记录	https://snap.stanford.edu/data/higgs-twitter.html	[36-37,58,61-63,72,75]
Sina Weibo	一个来自微博平台的数据集,收集了平台上微博内容和其相应转发的记录	https://www.aminer.org/influencelocality	[39,57,74,82]
Digg	一个来自新闻门户网站的数据集,收集了用户在平台上转发新闻的记录	https://www.isi.edu/~lerman/downloads/digg2009.html	[56,58,61,80]
Flixster	一个来自一款移动社交网络应用的数据集,包括用户的社交网络和信息的转发记录	https://networkrepository.com/soc-flixster.php	[63]
Lastfm	一个来自 Last.fm 音乐社交平台的数据集,收集了近 1000 个用户在音乐平台上订阅记录	http://ocelma.net/MusicRecommendationDataset/lastfm-1K.html	[58]
Irvine	一个来自加州大学学生在线社区的数据集,记录用户之间发送消息的时间信息	https://toreopsahl.com/datasets/#online_forum_network	[58]
ICWSM	一个收集了 1 年内 4400 万篇博客引用信息的数据集	https://www.icwsm.org/data/	[40,56,58]

4.3 算法性能评估

为进一步了解不同算法推断传播网络的准确性,本文对常见的基于时间序列数据的传播网络推断方法进行对比实验,分别使用人工网络模型生成的拓扑结构和真实世界存在的拓扑结构作为推断目标。对于人工网络模型,本文采用 Erdős-Rényi random graph,Core-peripherynetwork 以及 LFR benchmark graphs 3 种模型生成 4 个节点数量在 1000 左右的人工网络。本文使用 ICM 模型产生级联数据,设置节点间的传输概率 $\beta=0.3$ 控制级联数据的长度,并保证传播过程覆盖网络边数的 95% 以上,节点间传输时间均采样于指数分布 $Exp(1)$ 中。使用 F-score 作为统一的评估指标,数据集的细节与结果如表 5 所列。

表 6 中的结果显示,Dani[52]在多个数据集上的推断表现均位于前列,但 Dani 需要人工设置推断边数作为停止条件,导致其不能保证在真实使用场景(即未知传播网络的边数时)中依旧保持可信赖的推断表现。基于生存分析理论建立模型的方法 InfoPath[45]虽然可以同时推断传播网络以及节点间联系强弱的关系,但其推断准确率低于其他方法。相较之下,文献[54]所提出的贝叶斯推断方法以及基于边嵌入的 KEBC[63]方法,可在无需设置推断边数的条件下,依旧达到较高的推断准确性,在适用性方面领先于其他方法。

表 6 基于时间序列的传播网络方法推断表现

Table 6 Inference performance of diffusion network methods based on time series

网络	$\|V\|$	$\|E\|$	$\|C\|$	NetInf[21]	[53]	InfoPath[45]	Dani[52]	[36]	KEBC[63]
Erdős Rényi random graph	1024	2044	2000	0.31	0.66	0.25	0.99	0.89	0.63
Core-periphery network	1024	2017	2000	0.35	0.67	0.28	0.79	0.43	0.69
LFR benchmark graphs (average degrees=6)	1000	3046	2000	0.41	0.54	0.52	0.87	0.59	0.91
LFR benchmark graphs (average degrees=10)	1000	4914	2000	0.62	0.51	0.30	0.90	0.37	0.78
Scientists collaboration network	379	914	1000	0.31	0.66	0.25	0.99	0.89	0.63
Email communication in European	1005	16706	1000	0.35	0.67	0.28	0.79	0.43	0.69

结束语 传播网络在分析宏观的传播过程以及揭示传播动力学规律时必不可少,但在真实世界中,传播网络的结构通常难以被完整地观测和获取。因此,研究传播网络推断问题在传播学、社交网络分析等领域有着重要意义。本文先根据输入数据的种类对现有传播网络推断方法进行粗粒度分类,再根据方法的建模思路与方式进行二次分类,同时收集汇总了常用的性能评估指标与基准数据集,为研究人员提供有益的参考。对于未来,传播网络推断问题仍需要从以下几个方面进行加强和改善。

1)目前依旧没有算法能够在无需任何网络先验知识的条件下同时实现高推断效率和高推断结果准确率。在基于极大似然估计的方法中,虽然使用凸优化计算节点之间的传输率可以在无需网络先验知识(例如边的数量等)的条件下得到较为精确的结果,但其需要大量的计算与时间开销。基于贝叶斯推断的方法同样需要进行大量的迭代和采样,保证算法收敛,得到准确结果。基于感染状态数据的方法由于父节点集合的节点复杂多变,因此推断效率较低,难以应对大规模的网络。如何设计出高效、精准且无需任何先验知识的传播网络推断算法,仍是需要解决的问题之一。

2)目前多数传播网络推断方法假设在一次传播过程中各个节点的感染事件相互独立,且节点的感染源是唯一的,但在真实世界环境中这一假设未必成立,例如用户主动传播信息的意念可能是由其多位好友的共同影响而产生的[90]。因此,探索更符合真实世界中信息传播现象的模型,揭示传播动力学规律,是需要重点研究的内容之一。

3)针对质量不确定的数据,需要获得更可靠的传播网络推断结果。现有的方法大多假设观测收集的数据是完整且准确无误的,例如级联数据中节点被感染的时间信息等。但在真实世界中,数据的质量无法得到保证,从而导致推断的结果不可信,例如 Peel 等[91]提出局部的数据误差会影响到全局的网络推断表现。因此,传播网络推断方法需要将数据的不确定性纳入考量范围,使其更具有适用性。虽然目前已有一些工作[41,70,92]分别针对数据的缺失问题和误差问题提出了解决方案,但仍缺少能全方面考虑数据质量不确定性的方法。因此,利用质量不确定的数据进行可靠的传播网络推断是未来的发展方向之一。

4)相较于其他类型的方法,扩散嵌入模型不仅能够推断信息级联中已发生的传播关系,还能根据传播文本语义、用户特征等信息预测未来可能发生的传播事件[61,80],因此具有更广泛的适用场景,是未来重点研究的内容之一。但这类方法需要更多的数据作为驱动,以便模型可以自动地挖掘用户与用户、用户与传播信息之间的潜在关系,因此数据的质量和数量会很大程度地影响算法的表现。此外,通过将网络中的实体嵌入到固定维度的空间中,能够避免传统信息传播模型(例如 IC 模型)产生的空间复杂度问题,从而有效地应对更大规模网络。然而,对于维度的选择,不能简单地通过增加维数来试图提高算法的表现。已有研究工作[56]证实,盲目地增加维数可能会导致过拟合而降低算法性能。对于扩散嵌入模型的未来研究,一方面,研究者需要加强模型的泛化能力,以尽可能减少重新训练模型的成本;另一方面,研究人员需要关注模型训练时可能产生的过拟合以及维度灾难问题,防止算法在应用场景中出现性能退化,甚至失效的现象。

5)需要在一个统一的框架中建模信息扩散过程。信息传播领域涉及的任务除了本文关注的传播网络推断之外,还包括流行度预测[93-94]、影响力最大化[95]、传播源定位[96]等。这些任务虽然都涉及信息传播和扩散过程,但在不同的问题和数据下却有着各自的假设和处理方式。例如,文献[93]根据任务的粒度大小,将信息预测工作分为微观、中观和宏观级,不同层级下的工作对信息扩散过程的描述和假设并不完全相同。此外,同一任务中的不同方法对信息扩散过程也会有着不同的解释。例如,上文涉及的基于时间序列数据方法中的传播树模型在建模信息传播过程时具有 IC 模型所带来的更严格的假设,而多数基于网络表示学习的方法则直接从观察结果中学习传播模型。文献[94]根据信息传播模型的随机性和确定性,将基于时间序列数据的流行度预测方法系统地分为点过程模型(Point Process Models)以及仓室模型(Compartment Models)。因此,未来工作的另一个挑战是如何在同一个框架下建立信息传播的过程,使得各个信息传播领域相关的任务之间更具有解释性和整体性。

参 考 文 献

[1] FANG D, CHEN B. Ecological network analysis for a virtual water network[J]. Environmental Science & Technology, 2015, 49(11): 6722-6730.

[2] HUANG I B, KEISLER J, LINKOV I. Multi-criteria decision analysis in environmental sciences: Ten years of applications and trends[J]. Science of the Total Environment, 2011, 409(19): 3578-3594.

[3] SPORNS O. Contributions and challenges for network models in cognitive neuroscience[J]. Nature Neuroscience, 2014, 17(5): 652-660.

[4] PARK J Y, CHOI J, CHOI J Y. Network analysis in systems epidemiology[J]. Journal of Preventive Medicine and Public

Health, 2021, 54(4): 259.
[5] KUMAR P, SINHA A. Information diffusion modeling and analysis for socially interacting networks[J]. Social Network Analysis and Mining, 2021, 11(1): 1-18.
[6] HETHOTE H W. The mathematics of infectious diseases[J]. SIAM Review, 2000, 42(4): 599-653.
[7] ADAR E, ADAMIC L A. Tracking information epidemics in blogspace[C]// The 2005 IEEE/WIC/ACM International Conference on Web Intelligence(WI'05). IEEE, 2005: 207-214.
[8] ZHOU J, LIU Z, LI B. Influence of network structure on rumor propagation[J]. Physics Letters A, 2007, 368(6): 458-463.
[9] TANG J, SUN J, WANG C, et al. Social influence analysis in large-scale networks[C]// Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2009: 807-816.
[10] SUN J, TANG J. A survey of models and algorithms for social influence analysis[C]// Proceedings of the Sixth ACM International Conference on Web Search and Data Mining(WSDM'13). 2011: 177-214.
[11] SADRI A M, UKKUSURI S V, AHMED M A. Review of social influence in crisis communications and evacuation decision-making[J]. Transportation Research Interdisciplinary Perspectives, 2021, 9: 100325.
[12] LESKOVEC J, ADAMIC L A, HUBERMAN B A. The dynamics of viral marketing[J]. ACM Transactions on the Web (TWEB), 2007, 1(1): 5-es.
[13] BHATTACHARYA S, GAURAV K, GHOSH S. Viral marketing on social networks: An epidemiological perspective[J]. Physica A: Statistical Mechanics and its Applications, 2019, 525: 478-490.
[14] WANG X, ZHU X, TAO X, et al. Anomalous role of information diffusion in epidemic spreading[J]. Physical Review Research, 2021, 3(1): 013157.
[15] ZHU L, YANG F, GUAN G, et al. Modeling the dynamics of rumor diffusion over complex networks [J]. Information Sciences, 2021, 562: 240-258.
[16] VAN DER LINDEN S. Misinformation: susceptibility, spread, and interventions to immunize the public[J]. Nature Medicine, 2022, 28(3): 460-467.
[17] ZAREIE A, SAKELLARIOU R. Minimizing the spread of misinformation in online social networks: A survey[J]. Journal of Network and Computer Applications, 2021, 186: 103094.
[18] PENG S, ZHOU Y, CAO L, et al. Influence analysis in social networks: A survey[J]. Journal of Network and Computer Applications, 2018, 106: 17-32.
[19] WANG J, WANG Y C, HUANG M J. False information in social networks: Definition, detection and control[J]. Computer Science, 2021, 48: 263-277.
[20] ADAR E, ADAMIC L A. Tracking information epidemics in blogspace[C]// The 2005 IEEE/WIC/ACM International Conference on Web Intelligence(WI'05). IEEE, 2005: 207-214.
[21] GOMEZ-RODRIGUEZ M, LESKOVEC J, KRAUSE A. Inferring networks of diffusion and influence[J]. ACM Transactions on Knowledge Discovery from Data(TKDD), 2012, 5(4): 1-37.
[22] MYERS S, LESKOVEC J. On the convexity of latent social network inference[J]. Advances in Neural Information Processing Systems, 2010: 1741-1749.
[23] LI H, XIA C, WANG T, et al. Capturing dynamics of information diffusion in SNS: A survey of methodology and techniques [J]. ACM Computing Surveys(CSUR), 2021, 55(1): 1-51.
[24] GUILLE A, HACID H, FAVRE C, et al. Information diffusion in online social networks: A survey[J]. ACM Sigmod Record, 2013, 42(2): 17-28.
[25] LI M, WANG X, GAO K, et al. A survey on information diffusion in online social networks: Models and methods[J]. Information, 2017, 8(4): 118.
[26] PASTOR-SATORRAS R, VESPIGNANI A. Epidemic spreading in scale-free networks[J]. Physical Review Letters, 2001, 86(14): 3200.
[27] LIU D, YIN Y W, SONG M. Microblog information diffusion: Simulation based on sir model[J]. Journal of Beijing University of Posts and Telecommunications (Social Sciences Edition), 2014, 16(3): 28.
[28] NAZEMIAN A, TAGHIYAREH F. Influence maximization in independent cascade model with positive and negative word of mouth[C]// 6th International Symposium on Telecommunications(IST). IEEE, 2012: 854-860.
[29] LI S, KONG F, TANG K, et al. Online influence maximization under linear threshold model[J]. Advances in Neural Information Processing Systems, 2020, 33: 1192-1204.
[30] ZHOU F, JIAO J R, LEI B. A linear threshold-hurdle model for product adoption prediction incorporating social network effects [J]. Information Sciences, 2015, 307: 95-109.
[31] TRIVEDI N, SINGH A. Efficient influence maximization in social-networks under independent cascade model [J]. Procedia Computer Science, 2020, 173: 315-324.
[32] KEMPE D, KLEINBERG J, TARDOSÉ. Maximizing the spread of influence through a social network[C]// Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2003: 137-146.
[33] GRANOVETTER M. Threshold models of collective behavior [J]. American Journal of Sociology, 1978, 83(6): 1420-1443.
[34] SERRANO E, IGLESIAS C Á, GARIJO M. A novel agent-based rumor spreading model in twitter[C]// Proceedings of the 24th International Conference on World Wide Web. 2015: 811-814.
[35] RODRIGUEZ M G, SCHÖLKOPF B. Submodular inference of diffusion networks from multiple trees[J]. arXiv: 1205. 1671, 2012.
[36] GRAY C, MITCHIELL L, ROUGHAN M. Bayesian inference of network structure from information cascades [J]. IEEE Transactions on Signal and Information Processing over Networks, 2020, 6: 371-381.
[37] GHALEBI E, MIRZASOLEIMAN B, GROSU R, et al. Dynamic network model from partial observations[C]// Proceedings of the 32nd International Conference on Neural Information Processing Systems(NIPS'18). 2018: 9884-9894.
[38] DJOLONGA J, KRAUSE A. Learning implicit generative models using differentiable graph tests[J]. arXiv: 1709. 01006, 2017.
[39] HAO X, LI X. Network topology inference with estimated node

importance[J]. Europhysics Letters,2021,134(5):58001.

[40] XU S,SMITH D. Contrastive training for models of information cascades[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2018.

[41] DOU P,SONG G,ZHAO T. Network topology inference from incomplete observation data[J]. Science China Information Sciences,2018,61(2):028102:1-028102:3.

[42] KONG L,GAO C,PENG S. Clustering-Based Network Inference with Submodular Maximization[C]// Pacific Rim International Conference on Artificial Intelligence. 2022:118-131.

[43] RODRIGUEZ M G,BALDUZZI D,SCHÖLKOPF B. Uncovering the temporal dynamics of diffusion networks[J]. arXiv:1105.0697,2011.

[44] GOMEZ-RODRIGUEZ M,SONG L,DANESHMAND H,et al. Estimating diffusion networks: Recovery conditions, sample complexity & soft-thresholding algorithm[J]. The Journal of Machine Learning Research,2016,17(1):3092-3120.

[45] GOMEZ-RODRIGUEZ M,LESKOVVEC J,SCHÖLKOPF B. Structure and dynamics of information pathways in online media[C]// Proceedings of the Sixth ACM International Conference on Web Search and Data Mining. 2013:23-32.

[46] DU N,SONG L,YUAN M,et al. Learning networks of heterogeneous influence[C]// Proceedings of the 25th International Conference on Neural Information Processing Systems(NIPS'12). 2012:2780-2788.

[47] WANG S,HU X,YU P S,et al. MMRate:inferring multi-aspect diffusion networks with multi-pattern cascades[C]// Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2014:1246-1255.

[48] TAN Q,LIU Y,LIU J. Motif-aware diffusion network inference[J]. International Journal of Data Science and Analytics,2020,9(4):375-387.

[49] ZHAO T,SONG G,HE X. Inferring diffusion networks with life stage heterogeneity[J]. Science China Information Sciences,2018,61:1-16.

[50] GAO C,WANG Y,WANG Z,et al. Pairwise-interactions-based Bayesian Inference of Network Structure from Information Cascades[C]// Proceedings of the ACM Web Conference 2023. 2023:102-110.

[51] ESLAMI M,RABIEE H R,SALEHI M. Dne:A method for extracting cascaded diffusion networks from social networks[C]// 2011 IEEE Third International Conference on Privacy,Security,Risk and Trust and 2011 IEEE Third International Conference on Social Computing. IEEE,2011:41-48.

[52] RAMEZANI M,RABIEE H R,TAHANI M,et al. Dani:A fast diffusion aware network inference algorithm[J]. arXiv:1706.00941,2017.

[53] FOROUTAIN N,HAMZEH A. Discovering the hidden structure of a social network:A semi supervised approach[J]. IEEE Transactions on Computational Social Systems, 2017,4(1):14-25.

[54] CRAWFORD F W. Hidden network reconstruction from information diffusion[C]// 2015 18th International Conference on Information Fusion(Fusion). IEEE,2015:180-185.

[55] TAHANI M,HEMMATYAR A M A,RABIEE H R,et al. Inferring dynamic diffusion networks in online media[J]. ACM Transactions on Knowledge Discovery from Data(TKDD),2016,10(4):1-22.

[56] BOURIGAULT S,LAGNIER C,LAMPRIER S,et al. Learning social network embeddings for predicting information diffusion[C]// Proceedings of the 7th ACM International Conference on Web Search and Data Mining. 2014:393-402.

[57] HU Q,XIE S,LIN S,et al. Clustering embedded approaches for efficient information network inference[J]. Data Science and Engineering,2016,1(1):29-40.

[58] BOURIGAULT S,LAMPRIER S,GALLINARI P. Representation learning for information diffusion through social networks:an embedded cascade model[C]// Proceedings of the Ninth ACM International Conference on Web Search and Data Mining. 2016:573-582.

[59] KURASHIMA T,IWATA T,TAKAYA N,et al. Probabilistic latent network visualization:Inferring and embedding diffusion networks[C]// Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2014:1236-1245.

[60] ZHANG Y,LYU T,ZHANG Y. Cosine:Community-preserving social network embedding from information diffusion cascades[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2018,32(1).

[61] ZHUO W,ZHAO Y,ZHAN Q,et al. DiffusionGAN:Network embedding for information diffusion prediction with generative adversarial nets[C]// 2019 IEEE International Conference on Parallel & Distributed Processing with Applications,Big Data & Cloud Computing,Sustainable Computing & Communications,Social Computing & Networking(ISPA/BDCloud/SocialCom/SustainCom). IEEE,2019:808-816.

[62] RONG Y,ZHU Q,CHENG H. A model-free approach to infer the diffusion network from event cascade[C]// Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. 2016:1653-1662.

[63] HU S,CAUTIS B,CHEN Z,et al. Model-free inference of diffusion networks using RKHS embeddings[J]. Data Mining and Knowledge Discovery,2019,33(2):499-525.

[64] LAMPRIER S,BOURIGAULT S,GALLINARI P. Extracting diffusion channels from real-world social data:a delay-agnostic learning of transmission probabilities[C]// Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015. 2015:178-185.

[65] GRIPON V,RANNAT M. Reconstructing a graph from path traces[C]// 2013 IEEE International Symposium on Information Theory. IEEE,2013:2488-2492.

[66] AMIN K,HEIDARI H,KEARNS M. Learning from contagion(without timestamps)[C]// International Conference on Machine Learning. PMLR,2014:1845-1853.

[67] HUANG H,YAN Q,GAN T,et al. Learning diffusions without timestamps[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019:582-589.

[68] HUANG H,YAN Q,CHEN L,et al. Statistical inference of diffusion networks[J]. IEEE Transactions on Knowledge and Data Engineering,2019,33(2):742-753.

[69] HAN K, TIAN Y, ZHANG Y, et al. Statistical estimation of diffusion network topologies[C]// 2020 IEEE 36th International Conference on Data Engineering(ICDE). IEEE, 2020: 625-636.
[70] GAN T, HAN K, HUANG H, et al. Diffusion Network Inference from Partial Observations[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(9): 7493-7500.
[71] SUN Y, ZHANG Y, YAN Q, et al. Fast inference algorithm of diffusion networks without infection temporal information[J]. Journal of Frontiers of Computer Science & Technology, 2019, 13(4): 541.
[72] WANG L, ERMON S, HOPCROFT J E. Feature-enhanced probabilistic models for diffusion network inference[C]// Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Berlin, Heidelberg: Springer, 2012: 499-514.
[73] DU N, SONG L, WOO H, et al. Uncover topic-sensitive information diffusion networks[C]// Artificial Intelligence and Statistics. PMLR, 2013: 229-237.
[74] ZHOU D H, HAN W B, WANG Y J, et al. Information diffusion network inferring and pathway tracking[J]. Science China Information Sciences, 2015, 58(9): 1-15.
[75] MANCO G, RITACCO E, BARBIERI N. A Factorization Approach for Survival Analysis on Diffusion Networks[J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 33(1): 1-13.
[76] ULLMAN J B, BENTLER P M. Structural equation modeling [M]// Handbook of Psychology, Second Edition, 2012, 2.
[77] BAINGANA B, MATEOS G, GIANNAKIS G B. Proximal-gradient algorithms for tracking cascades over social networks[J]. IEEE Journal of Selected Topics in Signal Processing, 2014, 8(4): 563-575.
[78] BAINGANA B, GIANNAKIS G B. Tracking switched dynamic network topologies from information cascades[J]. IEEE Transactions on Signal Processing, 2016, 65(4): 985-997.
[79] SHEN Y, BAINGAGA B, GIANNAKIS G B. Tensor decompositions for identifying directed graph topologies and tracking dynamic networks[J]. IEEE Transactions on Signal Processing, 2017, 65(14): 3675-3687.
[80] GAO S, PANG H, GALLINAR P, et al. A novel embedding method for information diffusion prediction in social network big data[J]. IEEE Transactions on Industrial Informatics, 2017, 13(4): 2097-2105.
[81] WANG D, ZHOU W, ZHENG J X, et al. Who spread to whom? Inferring online social networks with user features[C]// 2018 IEEE International Conference on Communications (ICC). IEEE, 2018: 1-6.
[82] YANG X, DONG M, CHEN X, et al. Recommender system-based diffusion inferring for open social networks[J]. IEEE Transactions on Computational Social Systems, 2019, 7(1): 24-34.
[83] ERDŐS P, RÉNYI A. On the evolution of random graphs[J]. Publications of the Mathematical Institute of the Hungarian Academy of Sciences, 1960, 5(1): 17-60.
[84] BORGATII S P, EVERETT M G. Models of core/periphery structures[J]. Social Networks, 2000, 21(4): 375-395.
[85] CLAUSET A, MOORE C, NEWMAN M E J. Hierarchical structure and the prediction of missing links in networks[J]. Nature, 2008, 453(7191): 98-101.
[86] LESKOVEC J, KLEINBERG J, FALOUTSOS C. Graphs over time: densification laws, shrinking diameters and possible explanations[C]// Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining. 2005: 177-187.
[87] LANCICHINETTI A, FORTUNATO S, RADICCHI F. Benchmark graphs for testing community detection algorithms[J]. Physical Review E, 2008, 78(4): 046110.
[88] LESKOVEC J, FALOUTSOS C. Scalable modeling of real graphs using kronecker multiplication[C]// Proceedings of the 24th International Conference on Machine Learning. 2007: 497-504.
[89] LESKOVEC J, KLEINBERG J, FALOUTSO C. Graph evolution: Densification and shrinking diameters[J]. ACM transactions on Knowledge Discovery from Data(TKDD), 2007, 1(1): 2-es.
[90] CHEN W, YUAN Y, ZHANG L. Scalable influence maximization in social networks under the linear threshold model[C]// 2010 IEEE International Conference on Data Mining. IEEE, 2010: 88-97.
[91] PEEL L, PEIXOTO T P, DE DOMENICO M. Statistical inference links data and theory in network science[J]. Nature Communications, 2022, 13(1): 1-15.
[92] PEIXOTO T P. Reconstructing networks withunknown and heterogeneous errors[J]. Physical Review X, 2018, 8(4): 041011.
[93] ZHOU F, XU X, TRAJCEVSKI G, et al. A survey of information cascade analysis: Models, predictions, and recent advances [J]. ACM Computing Surveys(CSUR), 2021, 54(2): 1-36.
[94] GAO X, CAO Z, LI S, et al. Taxonomy and evaluation for microblog popularity prediction[J]. ACM Transactions on Knowledge Discovery from Data(TKDD), 2019, 13(2): 1-40.
[95] LI Y, FAN J, WANG Y, et al. Influence maximization on social graphs: A survey[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(10): 1852-1872.
[96] JIANG J, WEN S, YU S, et al. Identifying propagation sources in networks: State-of-the-art and comparative studies[J]. IEEE Communications Surveys & Tutorials, 2016, 19(1): 465-481.

WANG Yuchen, born in 1999, postgraduate. His main research interests include data mining and complex network analysis.

GAO Chao, born in 1980, Ph.D, professor, Ph.D supervisor. His main research interests include artificial intelligence theory, social computing, system simulation, big data analysis, etc.

(责任编辑:柯颖)