软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2015,26(9):2339−2355 [doi: 10.13328/j.cnki.jos.004703]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于链路预测的社会网络事件检测方法*

胡文斌[1,2], 彭 超[1], 梁欢乐[1], 杜 博[1]

[1](武汉大学 计算机学院,湖北 武汉 430072)

[2](软件工程国家重点实验室(武汉大学),湖北 武汉 430072)

通讯作者: 胡文斌, E-mail: hwb@whu.edu.cn

摘 要: 网络演化分析与事件检测,是当前社会网络研究的热点和难点.现有的研究工作主要是针对网络提出不同的模型,并用网络特征指标对仿真结果进行评价.这些方法存在如下问题:(1) 每种方法仅针对特定网络,通用性不高;(2) 特征指标多种多样,不同模型的表现情况缺乏统一的评价标准;(3) 未考虑网络演化的时间特性,难以描述网络演化的波动性,无法检测事件.针对上述问题,提出一种基于链路预测的社会网络事件检测方法 LinkEvent(由相似性计算算法 SimC 和事件检测算法 EventD 组成),它可以对不同网络的波动性进行统一评价,并依此建立事件检测模型.主要工作包括:(1) 证明了链路预测可以反映网络演化机制,相同机制下的模型演化法和链路预测在分析网络演化上具有内在的一致性;(2) 基于链路预测,提出一种网络相似性计算算法 SimC(similar computing),并在考虑微观因素的基础上进行改进;(3) 利用相似性计算结果,提出一种事件检测算法 EventD(event detecting)检测出新事件.在不同特征的网络上进行实验,结果表明:所提出的 LinkEvent 方法能够较好地解决网络演化波动性问题,实现事件检测;同时也证明了利用链路预测技术进行网络演化分析的可行性以及相似性计算和事件检测算法的有效性.

关键词: 社会网络分析;事件检测;链路预测;网络演化分析;网络波动性分析

中图法分类号: TP311

中文引用格式: 胡文斌,彭超,梁欢乐,杜博.基于链路预测的社会网络事件检测方法.软件学报,2015,26(9):2339−2355. http://www.jos.org.cn/1000-9825/4703.htm

英文引用格式: Hu WB, Peng C, Liang HL, Du B. Event detection method based on link prediction for social network evolution. Ruan Jian Xue Bao/Journal of Software, 2015,26(9):2339−2355 (in Chinese). http://www.jos.org.cn/1000-9825/4703.htm

Event Detection Method Based on Link Prediction for Social Network Evolution

HU Wen-Bin[1,2], PENG Chao[1], LIANG Huan-Le[1], DU Bo[1]

[1](Computer School, Wuhan University, Wuhan 430072, China)

[2](State Key Laboratory of Software Engineering (Wuhan University), Wuhan 430072, China)

Abstract: Tracking the evolution and detecting events are popular and difficult problems in the field of social network analysis. Most of the research focuses on proposing different models to fit different network characteristics. This type of approach usually has three problems: (1) Each model is designed for one particular network and cannot well fit other networks; (2) There are many network statistics, so the evaluation of these network models lacks of unified platforms; (3) Without taking temporal information into account, these network models can hardly track the evolution and detect events. To solve these problems, this paper presents a method for event detection in social networks based on link prediction, which can evaluate the fluctuation of the networks and detect the events in social networks. The main work is as follow: (1) Demonstrates the method "modelling and evaluating" is in accord with link prediction on revealing the network evolution mechanism; (2) Proposes an algorithm similarity computing (SimC) to compute the similarity of networks and further improves this algorithm by taking micro factors into account; (3) Evaluates the fluctuation of the network evolution and proposes an event

* 基金项目: 国家自然科学基金(70901060, 61471274); 湖北省自然科学基金(2011CDB461); 软件工程国家重点实验室(武汉大学)开放基金(SKLSE 2010-08-15); 武汉市科技局青年晨光计划(201150431101); 武汉市科技重大计划项目(2015010101010023)

收稿时间: 2014-03-21; 修改时间: 2014-06-10; 定稿时间: 2014-08-01

detecting (EventD) algorithm to detect the events. The results of the experiment show that the presented method can effectively solve the problem of tracking the evolution and detecting events.

Key words: social network analysis; event detection; link prediction; network evolution analysis; network volatility analysis

社会网络是不断发展变化的,网络演化分析与事件检测是社会网络分析的重要组成部分.网络演化分析是指通过跟踪网络不同阶段的特征变化来描述其演化规律,进而分析网络增长、传播等行为,预测未来结构,甚至加以人为干预,以得到预期结果.网络演化分析技术已经随着社交网络的爆炸式发展而广泛应用于用户行为分析、消息传播引导等领域.然而不同社会网络的特征千差万别,演化机制纷繁复杂,如何高效地模拟真实网络的增长、传播等行为,已经成为当前面临的首要挑战.事件检测是网络演化分析技术的一项具体应用,一般是指在描述网络演化规律的基础上,通过分析网络各个阶段的差异,检测出网络中发生的事件并提出干预策略.事件检测在分析犯罪网络中核心头目的更替、判断公司邮件网络中组织架构的变迁等方面具有重要的指导意义.真实网络中,各种各样事件的发生都可能使网络偏离正常演化方向,从而呈现出不同的结构变化.如何定义并检测出这些事件、评估事件影响、提出相应策略进而干预,是事件检测研究的难点.

为了反映网络的特征变化、揭示其内在演化规律,学者们提出了许多演化模型,较为典型的有 Erdős-Renyi (E-R)随机图模型[1]、Watts-Strogatz(W-S)小世界模型[2]、Barabási-Albert(B-A)无标度模型[3].这些方法的一般步骤是:基于一种或多种演化机制构建网络模型,调整模型参数以适配真实网络,仿真得到各个时间段的网络;最后,通过度分布、平均聚集系数等网络统计特征来评价模型对真实网络的描述程度.此类模型演化方法(model evolution,简称 ME)的优点是实现简单,可根据网络特性调整参数来构建不同的网络.然而上述模型仅仅针对特定网络设计,很难兼顾各种统计特性,不同模型的表现情况缺乏统一的评价标准;同时,由于未考虑网络各时间段前后的相互关系,忽略了网络演化过程中的波动性,这些模型都难以描述网络演化的稳定程度,无法对网络中的事件进行检测.

链路预测(link prediction,简称 LP)[4]是指在给定网络当前时间段拓扑结构(点与点之间的链接关系)的前提下,如何准确预测下一时间段新出现的边.其具体步骤为:按照某种指标计算当前时间段内所有点对的得分,删除已存在的点对(即,网络中已经存在的边),将剩余点对按照得分降序排列,根据评价指标选取前 L 个点对作为预测结果输出.与模型演化方法不同,链路预测充分利用了当前时间段的既有信息,采用不同演化机制构建的各种指标对下一时间段的网络结构进行预测;同时,由于具有统一的评价标准,各种指标的预测效果可以比较.

那么,ME 与 LP 之间是否存在必然的内在联系?通过相关实验发现:在某种演化机制构建的网络模型上进行链路预测,基于同种机制构建的指标,其表现远远优于其他机制构建的指标.这说明 LP 可以反映网络的内部演化机制,针对不同特性的网络,应用不同机制的链路预测指标跟踪其波动性,并进行事件检测是可行的.

本文首先在宏观上通过链路预测推断出该网络的内部演化机制,然后在微观上考虑每个点与其邻居在何种程度上符合这种演化机制,以此提出一种基于链路预测的网络相似性计算算法 SimC.利用 SimC 可以得出网络各个时间段之间的相似性,从而对平稳网络段和事件发生段进行区分.在此基础上,本文构建了事件检测算法 EventD,利用部分事件发生点的先验信息来设定事件阈值,检测出新事件的发生.

综上,本文主要贡献总结如下:

(1) 证明了链路预测可以反映网络演化机制,相同机制下的 ME 和 LP 在分析网络演化上具有内在的一致性,进而说明采用链路预测跟踪网络波动性以及事件检测是可行的;

(2) 基于链路预测技术,提出一种网络相似性计算算法 SimC,并在考虑微观因素基础上对 SimC 算法进行改进,从而对平稳网络段与事件发生段进行有效区分;

(3) 利用相似性计算结果对网络演化的波动程度进行评价,提出一种事件检测算法 EventD,检测新事件的发生.

本文第 1 节介绍相关研究工作.第 2 节通过实验证明网络模型与链路预测的内在一致性.第 3 节介绍相似性计算算法 SimC 和事件检测算法 EventD.第 4 节为实验分析.第 5 节是结论及展望.

1 相关工作

网络演化分析作为一个热点领域,一直受到国内外学者的广泛关注.目前,许多工作集中在网络演化的统计特性研究上,包括度分布、聚集系数、密度、社团等[5].通过研究统计特性的变化来发现网络的演化规律,是较为重要的一种手段.Leskovec 和 Kleinberg[11]发现:随着网络的长期演化,网络变得更加稠密,节点之间的平均距离反而变小.

通过一系列的研究,社会网络演化分析有两个广泛接受的演化机制:三元闭包[12]与优先链接[3].模型演化方法正是基于这些机制而展开的.宏观方面,较为著名的演化模型包括 E-R 模型[1]、W-S 模型[2]、B-A 无标度模型[3]、Marsili-Vega-Slanina 模型[13]等;微观方面,Leskovec 等人[14]分析了微观个体行为,建立了基于极大似然估计的评估模型.然而,这些模型都只是在适应某种网络特性的基础上来模仿真实的网络行为,并不能反映真实网络的波动性.

在链路预测方面,Kleinberg 等人[4]系统地提出了链路预测问题,并对比了多种相似性指标(共同邻居[12]、Jaccar 系数[15]、Adamic/Adar[16]、优先链接[3]等)在链路预测中的表现.Sarukkai[17]提出了基于马尔可夫链的链路预测方法.Newman 等人[18]发现:很多复杂网络具有层次结构,对层次结构的分析可以预测已经丢失的链接.Lichtenwalter 等人[19]考虑影响分类的因素提出了一种监督学习的链路预测平台,效果比无监督学习提高 30%以上.Symeonidis 等人[20]引入多条路径信息,提出了多路谱聚类方法(multi-way spectral clustering method),有效地提高了蛋白质作用网络和社会网络上的链路预测精度.Kunegis 等人[21]从网络谱特征的变化角度出发,提出了两种新的链路预测方法.Rao 等人[39]实现了基于 MapReduce 计算模型的链路预测算法,将链路预测应用于大规模网络.Dong 等人[22]研究了异质网络上的链路预测和推荐问题.此外,文献[23−26]各自提出了其他的相似性指标.该类文献多集中于链路预测自身机制的探讨或是提高预测的准确性上,缺乏具体的应用研究.

在网络波动性和本地事件(local events)研究方面,Albert 和 Barabási[27]基于边的重连事件提出了改进的 B-A 模型.O'Madadhain 等人[28]提出了社会网络分为持续关系(persistent relationship)和离散事件(discrete events)两种类型,并针对事件网络提出了链路预测和节点排名算法.Chundi 等人[29]针对电子邮件数据提出了一种时间分析方法,抽取出了隐藏的社会结构:个人中心交流模式(egocentric communication patterns)和社会化交流模式(sociocentric communication patterns).Alexandridis 等人[30]研究了语义网演化过程中的节点崩溃与自组织在社会知识结构方面的影响.针对事件检测,Wu 等人[31]和 Baruah 等人[32]提出的网络相似性计算方法均没有统一各种因素的影响,导致结果不具有参考性.Qiao 等人[33]基于个性特征对犯罪成员的邮件网络进行分析,挖掘出犯罪网络的核心成员,并发现异常通信事件.

综上,网络波动性描述与事件检测问题仍缺乏有效的解决方案.本文开创性地将链路预测技术引入到网络波动性研究上来,提出了一种高效的事件检测方法.

2 网络演化与链路预测

相对于模型演化方法,链路预测可以充分利用当前时间段的拓扑结构信息,通过某种反映网络演化机制的指标对下一时间段的网络结构进行预测,从而揭示了各个时间段的相互关系.为了证明模型演化与链路预测之间的内在联系,本文进行了如下实验.

Barabási 与 Albert 基于优先链接的网络演化机制提出了著名的无标度模型,在该模型算法的基础上,本文采用如下方式构造 B-A 网络:初始网络为空,每次加入 1 个新节点,每个新节点引入 1 条边;迭代 150 次后停止.实验构造的网络特性参数见表 1.

Table 1 Statistics of the generated B-A network

表 1 构造的 B-A 网络特性

节点	边	平均度	平均路径长度
150	149	1.987	4.650

基于网络拓扑结构信息的节点相似性研究是链路预测的主流,本文选用 8 种节点相似性指标进行链路预测,指标定义见表 2.

Table 2 Eight similarity indexes based on neighbor nodes

表 2 8 种节点相似性指标

名称	定义	名称	定义
共同邻居指标(CN)[12]	$S_{ij}=\|\Gamma(i)\cap\Gamma(j)\|$	Jaccard 指标(JA)[15]	$S_{ij}=\frac{\|\Gamma(i)\cap\Gamma(j)\|}{\|\Gamma(i)\cup\Gamma(j)\|}$
优先链接指标(PA)[3]	$S_{ij}=k(i)\times k(j)$	Sorenson 指标(SO)[24]	$S_{ij}=\frac{2\|\Gamma(i)\cap\Gamma(j)\|}{k(i)+k(j)}$
Adamic-Adar 指标(AA)[16]	$S_{ij}=\sum_{z\in\Gamma(i)\cap\Gamma(j)}\frac{1}{\log^{k(z)}}$	大度节点有利指标(HPI)[25]	$S_{ij}=\frac{\|\Gamma(i)\cap\Gamma(j)\|}{\min\{k(i),k(j)\}}$
Salton 指标(SA)[23]	$S_{ij}=\frac{\|\Gamma(i)\cap\Gamma(j)\|}{\sqrt{k(i)\times k(j)}}$	LNH-I 指标(LNH)[26]	$S_{ij}=\frac{\|\Gamma(i)\cap\Gamma(j)\|}{k(i)\times k(j)}$

其中,S_{ij} 表示节点 i 与节点 j 的相似性得分,$\Gamma(i)$表示节点 i 的邻居所组成的集合,节点 i 在网络中的度为 $k(i)=|\Gamma(i)|$.

对构造的网络采用表 2 中的相似性指标进行链路预测,评价指标采用 AUC[34].AUC 作为衡量链路预测算法精确度的主要指标,具体定义为

$$AUC=\frac{n'+0.5n''}{n} \tag{1}$$

其中,n 表示比较次数,n'表示测试集中边的分数值大于随机选择不存在边的分数值的次数,n''表示相等的次数.AUC 反映了链路预测指标的预测精度,越大说明指标越好.各相似性指标 AUC 表现比较见表 3.

Table 3 Comparison of prediction accuracy of each index on the generated B-A network

表 3 B-A 网络预测精度比较

Index	CN	PA	AA	SA	JA	SO	HPI	LNH
AUC	0.467	**0.893**	0.468	0.468	0.464	0.459	0.464	0.457

实验结果表明,PA 指标的表现远优于其他指标.这并非巧合,实际上,PA 指标与 B-A 网络所基于的演化机制均为优先链接,这恰恰证明了链路预测可以反映网络演化机制,相同机制下的 ME 和 LP 在分析网络演化上具有内在的一致性.现实中的网络是复杂且难以描述的,很多网络并不能在预知其演化机制的前提下进行分析.由上述实验可知:对于未知演化机制的网络,首先可以对其进行各种指标的链路预测,通过对比各个指标的表现来推测其演化机制,进而根据网络各个时间段是否符合该种演化机制来定义网络的波动程度以及事件.这说明了利用链路预测技术研究网络波动性和事件检测是可行的.

由此,本文提出了 LinkEvent 方法对社会网络事件进行检测.

3 基于链路预测的社会网络事件检测方法

本文提出的 LinkEvent 方法基于对网络演化序列的相似性分析,利用链路预测的相关指标设计出高效合理的算法,从而描述网络的演化趋势,检测网络中的事件.

3.1 整体框架

图 1 描述了 LinkEvent 方法进行事件检测的整体框架,输入数据集包括通话网络、邮件网络等社会网络.框架包括如下两个部分:

(1) 对输入数据采用算法 SimC 计算网络各个时间段的相似性,并根据计算结果得出网络演化序列 GraphS;

(2) 在 GraphS 上结合阈值 T,采用算法 EventD,输出事件序列 EventS.

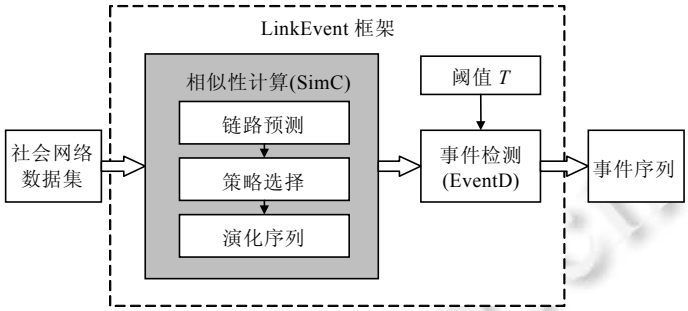

Fig.1 Framework of LinkEvent

图 1 LinkEvent 框架

LinkEvent 描述见算法 1.

算法 1. *LinkEvent*.

Input:网络 $G=\{g^1,g^2,g^3,\ldots,g^n\}$($n$ 为网络的时间跨度,g^t 为网络在 t 时刻的快照);

Output:事件序列 EventS.

1: 算法 SimC(见第 3.2 节)

1.1: 采用链路预测推测网络 G 的演化机制

1.2: 计算节点相似性

1.3: 根据选择策略计算节点微观权重

1.4: 计算改进的节点相似性

1.5: 输出演化序列 GraphS

2: 算法 EventD(见第 3.3 节)

2.1: 根据既有事件计算事件发生值区间

2.2: 确定事件发生阈值 T

2.3: 遍历 GraphS,输出事件序列 EventS

Step 1.2 中,计算节点之间相似性时,可能会出现结果无限大和部分节点变化却未计算的问题.为此,本文引入虚拟点 $V_{virtual}$ 予以解决,这将在第 3.2.3 节介绍.原始的 SimC 算法未考虑节点的微观演化机制因素,改进的 SimC 算法在 Step 1.3~Step 1.4 中根据多种策略来确定权重,这将在第 3.2.4 节重点讨论.为了提高算法的灵活性,Step 2.2 中,事件发生阈值 T 是在事件发生区间基础上人为确定的,这将在第 3.3 节讨论.

3.2 相似性计算算法SimC及改进

对于给定的网络 $G=\{g^1,g^2,g^3,\ldots,g^n\}$,$t$ 时刻的网络快照可用图 g^t 表示,g^t 与 g^{t+1} 之间的相似程度受到如下 3 个因素影响:

(1) 相对于 g^t,在 g^{t+1} 中新点的增加以及因此带来的新边的引入;

(2) 相对于 g^t,在 g^{t+1} 中旧点的消失以及相应边的消失;

(3) g^t 和 g^{t+1} 中,点保持稳定,与之关联边的单纯增加或减少.

以上 3 个因素相互叠加,各个节点及其关联的边随时间不断变化,在宏观上就表现为网络的整体波动.如何描述网络的波动程度,为事件检测提供分析基础,成为现在要解决的首要问题.为此,本文提出了 SimC 算法,并针对相关问题进行了探讨.

3.2.1 网络波动性的描述

在 G 的网络快照图 g^t 与 g^{t+1} 中,节点 i 的相似性定义为节点 i 在 g^t, g^{t+1} 中保持稳定的程度,用 $s(v_i^t,v_i^{t+1})$ 表示,其计算方法将在第 3.2.2 节讨论.

图 g^t,g^{t+1} 的相似性是图中各个节点相似性叠加的宏观表现,用 $S(g^t,g^{t+1})$表示,定义为

$$S(g^t,g^{t+1})=\sum_{i\in U_{t,t+1}} s(v_i^t,v_i^{t+1})\times\frac{1}{|U_{t,t+1}|} \tag{2}$$

其中,$U_{t,t+1}=g^t\cup g^{t+1}$.

g^t,g^{t+1} 的相似性反映的是两个图之间的近似程度,其值越大,表示网络在[t,t+1]时间段内变化越小,网络的波动程度越小.[t,t+1]时间段内,网络的波动性用 $\hat{D}(g^{t+1}\parallel g^t)$ 表示,定义为

$$\hat{D}(g^{t+1}\parallel g^t)=\frac{1}{S(g^t,g^{t+1})} \tag{3}$$

网络演化序列 GraphS 定义为各个时间段波动性的集合,如公式(4):

$$GraphS=\{\hat{D}(g^2\parallel g^1),\hat{D}(g^3\parallel g^2),...,\hat{D}(g^n\parallel g^{n-1})\} \tag{4}$$

3.2.2 节点相似性的计算

链路预测中,节点相似性指标是衡量图中两个不同节点的相似程度.核心思想是,两个节点的相似性取决于其拓扑结构信息(包括共同邻居数量、度的大小等).借鉴这一思想,网络 G 中的节点 i 在 g^t,g^{t+1} 中可看做两个不同的节点 v_i^t,v_i^{t+1},两者的相似性也可以用 v_i^t,v_i^{t+1} 拓扑结构来描述.

例如,链路预测中的 Jaccard 指标 $S_{ij}=\frac{|\Gamma(i)\cap\Gamma(j)|}{|\Gamma(i)\cup\Gamma(j)|}$[15],表示 v_i 与 v_j 的相似性由他们共同的邻居决定.相应地,v_i^t 与 v_i^{t+1} 的相似性可用公式(5)描述,记为 JAS 指标:

$$s(v_i^t,v_i^{t+1})=\frac{|\Gamma(v_i^t)\cap\Gamma(v_i^{t+1})|}{|\Gamma(v_i^t)\cup\Gamma(v_i^{t+1})|} \tag{5}$$

按此方式,链路预测中的 PA 指标应用到 v_i^t,v_i^{t+1} 相似性计算中,可得公式(6),记为 PAS 指标:

$$s(v_i^t,v_i^{t+1})=|\Gamma(v_i^t)|\times|\Gamma(v_i^{t+1})| \tag{6}$$

结合公式(5)、公式(2)或公式(6)、公式(2),可以计算出图 g^t,g^{t+1} 的相似性.

相较于本文根据节点相似性累加得出图的相似性,文献[35]给出了社区演化中同一社区前后状态的相似重叠度(relative overlap)计算方式.应用到 g^t,g^{t+1} 中,也可以描述 g^t,g^{t+1} 的相似程度,如公式(7),记为 ROS:

$$S(g^t,g^{t+1})=\frac{|A(g^t)\cap A(g^{t+1})|}{|A(g^t)\cup A(g^{t+1})|} \tag{7}$$

其中,$A(g^t)$表示 g^t 中所有节点的集合.

为了说明本文提出的相似性计算方法的优越性,举例分析如下.

图 2 示例了一个简单网络从 t 到 t+3 时间段内的演化过程,每 1 步均只增加 1 个节点,时间窗口设定为 1.

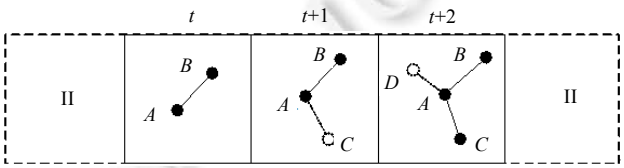

Fig.2 An example of network evolution

图 2 一个简单的网络演化示例图

分别利用 JAS,PAS 以及 ROS 计算 g^t,g^{t+1} 以及 g^{t+1},g^{t+2} 的相似性,结果见表 4.

为了提高事件检测的敏感性,网络波动性的变化幅度应尽可能地大,也即 $S(g^t,g^{t+1})$变化应尽可能地明显.由表 4 可知,JAS 表现优于 ROS,PAS 表现优于 JAS.实际上,ROS 只在宏观上考虑了节点的变化,并没有考虑边的变化,故效果最差.JAS 虽然具体到每一个节点及其关联边的拓扑结构变化,却并不能体现网络的演化规律.第 2 节已经证明了采用合适的链路预测指标更能反映网络的演化规律.图 2 所示的网络按照优先链接方式演化,因此

基于优先连接机制的 PAS 效果也最好.当然,以上分析只是在理想数据集上进行的模糊对比,每种指标的优劣性评价将在第 3.3.3 节介绍,真实网络的实验对比将在第 4.3 节介绍.

Table 4 Results of different similarity indexes

表 4 3 种相似性指标计算结果

相似性计算指标	$S(g^t,g^{t+1})$	$S(g^{t+1},g^{t+2})$
JAS	1/2	2/3
PAS	1	2
ROS	2/3	3/4

3.2.3 虚拟点的引入

从第 3.2.2 节的举例可以看出:采用 PAS,JAS 计算时,相对于 g^t,在 g^{t+1} 中新增节点的相似性计算值为 0,消失节点的相似性计算值也为 0.按此计算方式,网络中孤立节点(非连通网络的子图)的整体消失或者增加对相似性计算没有任何影响,这明显不符合实际情况.

为了解决这一问题,本文在网络中引入了一个虚拟节点 $V_{virtual}$,也称为观察者,可理解为从该节点的视角来观察整个网络的变化.$V_{virtual}$ 在网络中与所有的点都存在一条虚拟边,如图 3 所示.

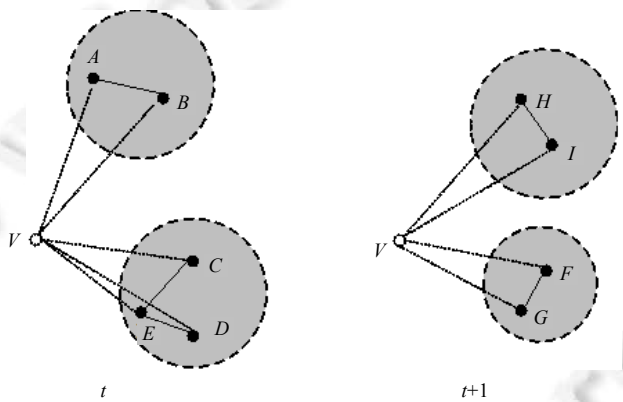

Fig.3 Introduction of virtual node

图 3 非连通网络中虚拟节点的引入

图 3 所示为一个非连通网络的演化情况.在 t+1 时刻,原来存在的子图 CDE 消失,而子图 FG 产生.若没有引入 $V_{virtual}$,则计算相似性时,这些子图的消失和产生对于网络的影响均无法体现.引入 $V_{virtual}$ 后,则可以通过 $V_{virtual}$ 的相似性计算体现出来.

引入虚拟点后,节点相似性计算可以完整地描述网络波动性.按照第 3.2.2 节的方式,将链路预测中的 8 种指标应用到相似性计算中来,可得出 8 种节点相似性计算指标,见表 5.

第 3.2.2 节举例可知,某种演化机制的网络采用相同机制的指标计算表现优于其他指标.对于未知演化机制的网络,应首先利用链路预测推断其演化机制,再采用对应机制的指标来计算相似性.对网络 G 进行链路预测,选取表现最好的 AUC 所对应的指标为最优指标(也称演化指标)O.指标 O 即反映了网络的演化机制.

根据以上分析,本文提出相似性计算算法 SimC,具体见算法 2.

算法 2. SimC.

Input:网络 $G=\{g^1,g^2,g^3,\ldots,g^n\}$;

Output:演化序列 GraphS.

1: 在 G 上进行链路预测,选取最优指标 O

2: for t in $n-1$:

for i in $U_{t,t+1}$:

根据选取 O 对应指标计算节点相似性 $s(v_i^t, v_i^{t+1})$

计算 $S_{t,t+1}$, $\hat{D}(g^{t+1} \| g^t)$

3: return GraphS

Table 5 Similarity computing with virtual node

表 5 引入虚拟点的节点相似性计算

名称	节点的相似性	名称	节点的相似性
共同邻居指标(CNS)	$s(v_i^t, v_i^{t+1}) = \lvert \Gamma(v_i^t) \cap \Gamma(v_i^{t+1}) \rvert + 1$	Jaccard 指标(JAS)	$s(v_i^t, v_i^{t+1}) = \frac{\lvert \Gamma(v_i^t) \cap \Gamma(v_i^{t+1}) \rvert + 1}{\lvert \Gamma(v_i^t) \cup \Gamma(v_i^{t+1}) \rvert + 1}$
优先链接指标(PAS)	$s(v_i^t, v_i^{t+1}) = (k(v_i^t)+1) \times (k(v_i^{t+1})+1)$	Sorenson 指标(SOS)	$s(v_i^t, v_i^{t+1}) = \frac{2(\lvert \Gamma(v_i^t) \cap \Gamma(v_i^{t+1}) \rvert + 1)}{k(v_i^t) + k(v_i^{t+1}) + 2}$
Adamic-Adar 指标(AAS)	$s(v_i^t, v_i^{t+1}) = \sum_{z \in \Gamma(v_i^t) \cap \Gamma(v_j^t)} \frac{1}{\lg \frac{k(v_z^t)+k(v_z^{t+1})}{2}}$	大度节点有利指标(HPIS)	$s(v_i^t, v_i^{t+1}) = \frac{\lvert \Gamma(v_i^t) \cap \Gamma(v_i^{t+1}) \rvert + 1}{\min\{k(v_i^t)+1, k(v_i^{t+1})+1\}}$
Salton 指标(SAS)	$s(v_i^t, v_i^{t+1}) = \frac{\lvert \Gamma(v_i^t) \cap \Gamma(v_i^{t+1}) \rvert + 1}{\sqrt{(k(v_i^t)+1) \times (k(v_i^{t+1})+1)}}$	LNH-I 指标(LNHS)	$s(v_i^t, v_i^{t+1}) = \frac{\lvert \Gamma(v_i^t) \cap \Gamma(v_i^{t+1}) \rvert + 1}{(k(v_i^t)+1) \times (k(v_i^{t+1})+1)}$

3.2.4 SimC 算法改进

原始 SimC 算法将所有节点平等看待,节点相似性直接累加得出图的相似性,并没有考虑节点的微观差异.实际上,某个节点及其周围拓扑结构变化若符合网络演化规律,可以看做是正常的演化,其对网络的波动性影响较小;而节点的变化若不符合演化规律,极有可能是事件的发生导致内在演化原则被打破,对网络的波动性影响是较大的.因此,不同节点在计算相似性时应该区别对待.为此,本文引入节点演化权重的概念.

定义节点的演化权重 w 为该节点及其周围拓扑结构变化与网络演化规律的契合程度,w 越大,表示节点的变化越符合演化规律. $w(v_i^t, v_i^{t+1})$ 表示节点 v_i 在[t,t+1]时间段的演化权重.

将链路预测精度 $AUC = \frac{n' + 0.5n''}{n}$ 具体到微观层面分析,假设网络 G 通过链路预测推断其最优指标(也即演化指标)为 O,g^t 中节点 v_i,v_j 不存在边,g^{t+1} 中两者之间产生了一条边 e_{ij}.把 e_{ij} 作为测试集,g^t 中 v_i 与其他节点不存在的边作为随机选择集,每次将 e_{ij} 在 O 指标下的分数值与随机选择集中的边进行比较.比较 n 次后,所得 AUC 定义为边 e_{ij} 的链路预测精度 $EAUC_e^{t,t+1}$.$EAUC_e^{t,t+1}$ 表示边 e_{ij} 的产生与演化规律的契合度,其值越大,说明 e_{ij} 的产生越符合 O 所对应的演化机制.

进一步,定义节点 v_i 的链路预测精度 $VAUC_i^{t,t+1}$ 为 v_i 在 g^{t+1} 中新增加边的链路预测精度的平均值,见公式(8):

$$VAUC_i^{t,t+1} = \begin{cases} \dfrac{\sum_{e \in NE_i^{t,t+1}} EAUC_e^{t,t+1}}{\lvert NE_i^{t,t+1} \rvert}, & \lvert NE_i^{t,t+1} \rvert > 0 \\ 0.5, & \lvert NE_i^{t,t+1} \rvert = 0 \end{cases} \tag{8}$$

其中, $NE_i^{t,t+1}$ 表示 v_i 在 g^{t+1} 中新增加的边的集合.$VAUC_i^{t,t+1}$ 反映了节点变化与演化规律的契合程度,其值越大,表明节点的变化越符合演化规律.

根据 $VAUC_i^{t,t+1}$ 可以计算节点的演化权重 $w(v_i^t, v_i^{t+1})$,本文提出了两种权重策略:

- 第 1 种策略: $w(v_i^t, v_i^{t+1})$ 由 $VAUC_i^{t,t+1}$ 与随机演化值 0.5 的比值确定,见公式(9):

$$w(v_i^t, v_i^{t+1}) = \frac{VAUC_i^{t,t+1}}{0.5} \tag{9}$$

- 第 2 种策略:将节点的链路预测精度分级, $VAUC_i^{t,t+1} = 0.5$ 表示不符合演化规律, $VAUC_i^{t,t+1} = 0.8$ 表示较符合, $VAUC_i^{t,t+1} = 1.0$ 表示完全符合.每个等级对应的演化权重计算方式见公式(10):

$$w(v_i^t,v_i^{t+1})=\begin{cases}\dfrac{VAUC_i^{t,t+1}}{\alpha}, & VAUC_i^{t,t+1}<0.5\\ VAUC_i^{t,t+1}, & 0.5\leqslant VAUC_i^{t,t+1}<0.8\\ \alpha\times VAUC_i^{t,t+1}, & 0.8\leqslant VAUC_i^{t,t+1}\leqslant 1.0\end{cases} \tag{10}$$

其中,α表示缩放因子,旨在更明显地区分 3 个等级,本文设置为 2.

考虑节点演化权重后,得到改进的 wSimC 算法见算法 3.

算法 3. wSimC.

Input:网络 $G=\{g^1,g^2,g^3,\ldots,g^n\}$;

Output:演化序列 GraphS.

1: 在 G 上进行链路预测,选取最优指标 O

2: for t in $n-1$:

for i in $U_{t,t+1}$:

根据 O 对应指标计算节点相似性 $s(v_i^t,v_i^{t+1})$

for e in $NE_i^{t,t+1}$:

计算 $EAUC_e^{t,t+1}$

计算 $VAUC_i^{t,t+1}$

选择权重策略,计算 $w(v_i^t,v_i^{t+1})$

计算改进后的节点相似性 $s'(v_i^t,v_i^{t+1})=w(v_i^t,v_i^{t+1})\times s(v_i^t,v_i^{t+1})$

计算 $S_{t,t+1}$, $\hat{D}(g^{t+1}\,\|\,g^t)$

3: return GraphS

两个策略的效果对比将在第 4.4 节的实验中予以说明.

原始的 SimC 算法中,对于网络 G(取 n 个时间快照,平均节点数为 N,平均度为 k),需要计算相似性的节点数为$(n-1)N$.改进的 wSimC 算法中,在计算某一节点的相似性时,还需要计算该点所有边的链路预测精度,时间复杂度为 $O((N-k)k)$,故整个算法的时间复杂度为 $O((n-1)N*(N-k)k)$.Barabási 与 Albert[3]在研究真实网络时发现:真实网络其大多遵守幂律分布,大部分是稀疏图,其节点的度都比较小,故 k 可认为是一个常数.由于数据集采集原因,网络的时间快照数 n 也可认为是一个常数.因此,对于稀疏的大型网络而言,wSimC 算法的时间复杂度为 $O(N^2)$.

3.3 事件检测

通过对网络进行相似性计算,可以得到网络演化序列 GraphS.GraphS 描述了网络的演化趋势,如平稳、发展、衰减等.分析序列的各个阶段,基于已经发生的事件信息,可以检测出新事件的发生.

3.3.1 网络平稳与事件

网络演化情况可按照波动性分为 3 种状态:

(1) 若网络 G 在 t 和 $t+1$ 时刻的网络完全相同,则称 G 在时间段$[t,t+1]$处于绝对平稳状态.此时,规定 $\hat{D}(g^{t+1}\,\|\,g^t)=0$.绝对平稳状态是一种理想状态,在真实网络中几乎不存在;

(2) 若网络 G 在 $t,t+1$ 时刻的波动性小于阈值 T,则称 G 在时间段$[t,t+1]$处于相对平稳状态,时间段$[t,t+1]$称为平稳段.相对平稳状态表明网络在演化规律作用下正常波动;

(3) 若网络 G 在 $t,t+1$ 时刻的波动性超出阈值 T,则称网络 G 在时间段$[t,t+1]$处于事件状态,t 称为事件点,时间段$[t,t+1]$称为事件段.事件可定义为干扰网络正常演化的事情,它通过改变具体点或边的拓扑结构来影响网络演化.

真实网络演化时,长期处于相对平稳状态,事件的发生导致其进入事件状态,事件影响消失后又恢复到新的相对平稳状态,依次交替.

3.3.2 事件检测算法

基于已有事件点组成的集合,分析 GraphS,即可判定出网络演化状态,检测事件的发生.本文提出的事件检测算法 EventD 基本思想如下:

根据已经发生的事件点组成的事件序列 $EventO=\{k|t=k$ 时发生事件,$k\in[1,m],m\leqslant n\}$,分析演化序列 GraphS,学习得到事件发生值区间$[L,H]$(L 为事件发生下边界,H 为上边界),选取 $T\in[L,H]$为发生阈值.为了提高时间检测的灵活性,阈值 T 由人工确定.在 $k=t$ 时,若 $\hat{D}(g^{k+1}\,\|\,g^{k})>T$,则网络在时间段$[k,k+1]$处于事件状态;反之,则为相对平稳状态.分析完毕,最终输出事件序列 EventS,见公式(11):

$$EventS=\{k\,|\,\hat{D}(g^{k+1}\,\|\,g^{k})>T,k\in[1,n-1]\} \tag{11}$$

具体过程见算法 4.

算法 4. EventD.

Input: $GraphED=\{\hat{D}(g^{2}\,\|\,g^{1}),\hat{D}(g^{3}\,\|\,g^{2}),...,\hat{D}(g^{n}\,\|\,g^{n-1})\}$,$EventO=\{k|t=k$ 时发生事件,$k\in[1,m],m\leqslant n\}$;

Output:事件序列 EventS.

1: for i in $EventO$:

$H=\max\{\hat{D}(g^{i+1}\,\|\,g^{i})\}$

$L=\min\{\hat{D}(g^{i+1}\,\|\,g^{i})\}$

2: 人工选取 $T\in[L,H]$

3: for k in $n-1$:

if $\hat{D}(g^{k+1}\,\|\,g^{k})>T$:

$k\in EventS$

4: return EventS

3.3.3 事件检测方法的评价

第 3.2.2 节中举例分析了多种指标下节点的相似性计算表现,符合网络演化机制的指标表现更好,但并没有用量化的数值来体现这种优劣性.第 3.2.4 节节点的演化权重计算时提出了两种选择策略,具体应用时,需要确定哪种策略的表现更好.为了实现以上目的,本文提出了一种简单的事件检测方法评价标准.

假定 k_1,k_2 为网络 G 中紧邻的两个事件点,$k_1+1<k_2$,G 在$[k_1+1,k_2]$时间段处于相对平稳状态,在$[k_2,k_2+1]$时间段处于事件状态,定义事件敏感表现:

$$Per=\frac{\hat{D}(g^{k_2+1}\,\|\,g^{k_2})-\dfrac{\sum_{i=k_1+1}^{k_2-1}\hat{D}(g^{i+1}\,\|\,g^{i})}{k_2-k_1-2}}{\dfrac{\sum_{i=k_1+1}^{k_2-1}\hat{D}(g^{i+1}\,\|\,g^{i})}{k_2-k_1-2}} \tag{12}$$

其中, $\hat{D}(g^{k_2+1}\,\|\,g^{k_2})>\max\{\hat{D}(g^{i+1}\,\|\,g^{i})\big|i=k_1+1,...,k_2-1\}$,这是因为事件段的波动性必然大于平稳段.

事件敏感表现 Per 是网络事件段波动性与平稳段平均波动性的比值,比值越大,表明事件越易被检测出,可用来评价事件检测方法的表现.在实际应用中,可针对相似性指标、权重策略等参数设计不同的事件检测算法,根据 Per 的评价结果,选取最优的参数配置.

4 实验分析

本节通过实验来分析框架中的关键理论,验证算法的表现情况.第 4.1 节介绍了实验中使用的社会网络数据集;第 4.2 节对比了真实数据集上链路预测各个指标的表现,推断出真实数据集的演化机制;基于第 4.2 节的分析结果,第 4.3 节对比了不同相似性计算指标下的算法表现,验证了 SimC 的有效性;第 4.4 节比较了多种权重策略下的算法表现,证明了改进 SimC 的有效性;第 4.5 节将算法检测的事件与真实的事件进行比较,表明了整个框

架的有效性.

4.1 数据集描述

为了验证本文提出方法的有效性,采用通信网络(VAST)和邮件网络(Enron)作为实验测试数据集.VAST 数据集来自 IEEE VAST 2008[36],它涉及 400 人组成的社会网络在 10 天内的通话数据,并已知在这期间发生了一次导致网络变革的事件.Enron 数据集来自 Enron 公司的内部邮件联系网络[37],它涉及 150 人的通信数据,时间跨度 111 周,本文选择有代表性的一段时间,约 47 周,期间包括公司破产等多次事件.

VAST 数据集的特点是单一的事件发生,其发生原因和时间确定,这有利于本文分析事件发生前后网络的变化;Enron 数据集的特点是多次事件连续发生,这有助于我们进行事件检测和分析事件对网络的影响.

4.2 链路预测指标分析

在第 2 节中,本文通过仿真实验说明了链路预测可以推断网络的演化机制.本节对真实网络使用不同链路预测指标进行预测,从而发现网络更为真实的演化机制.

本节实验比较了 PA,CN,AA,JA,SA,HPI,SO,LNH 等链路预测指标对 VAST 和 Enron 网络的预测能力.在 VAST 数据集实验中:训练集是 t 时刻前 3 天(t–3,t–2,t–1)的网络;测试集则是 t 时刻的网络,其链路预测精度 AUC 结果见表 6.其中,AUC^t 表示 t 时刻的链路预测精度.

Table 6 AUC comparison of VAST network

表 6 VAST 链路预测精度比较

Method	AUC^4	AUC^5	AUC^6	AUC^7	AUC^8	AUC^9	AUC^{10}
CN	0.5655	0.589	0.559	0.544	0.4875	0.589	0.505
PA	**0.8275**	**0.839**	**0.915**	**0.837**	**0.507**	**0.856**	**0.829**
AA	0.5815	0.608	0.5635	0.5555	0.494	0.5685	0.5025
SA	0.561	0.573	0.5565	0.5365	0.4925	0.564	0.504
JA	0.5705	0.588	0.551	0.548	0.4865	0.565	0.503
SO	0.56	0.598	0.5555	0.5415	0.453	0.5685	0.493
HPI	0.564	0.575	0.5505	0.523	0.487	0.563	0.5035
LNH	0.5605	0.589	0.5575	0.548	0.488	0.572	0.505

VAST 通信网络在第 7 天与第 8 天之间发生了一次高层变动,影响了网络结构.对表 6 分析可以明显看到:各项链路预测指标在第 8 天的链路预测准确率均在 0.5 左右,其中,AUC=0.5 可以理解为无规律的随机预测.除第 8 天外,网络无明显事件发生,处于相对平稳状态,PA 指标在这期间的链路预测 AUC 表现均在 0.8 以上,远高于其他指标.据此可以得到以下两个结论:

(1) 网络平稳段内,PA 指标预测精度更高,更符合网络的演化规律;

(2) 网络平稳段内的预测精度要高于事件发生时候的网络预测精度.

在接下来做 VAST 网络的相似性计算实验时,将重点分析 PAS 指标在相似性计算时是否真的比其他指标优异.

在 Enron 数据集实验中,使用(2001-05-07~2002-03-30)共 47 周的数据,时间窗口设置为 7,训练集为相对于 t 时刻前 3 周的网络,测试集为 t 时刻的网络,其链路预测精度 AUC 结果如图 4 所示,其中,横坐标标注的时间间隔为 4 周.

如图 4 所示,不同指标的 AUC 表现在不同时刻互有高低,并不能发现哪个指标在全局的表现更好.为了对网络做一个整体的演化规律分析,需要改变训练集和测试集的定义,现在将时间窗口设置为 47,对网络整体分析,训练集为 80%的边,测试集为 20%的边,其链路预测精度 AUC 结果见表 7.

从结果中可以看出,JA 的链路预测精度最高.因此在接下来对 Enron 网络的相似性计算中,将选择 JA 指标作为最优指标 O.

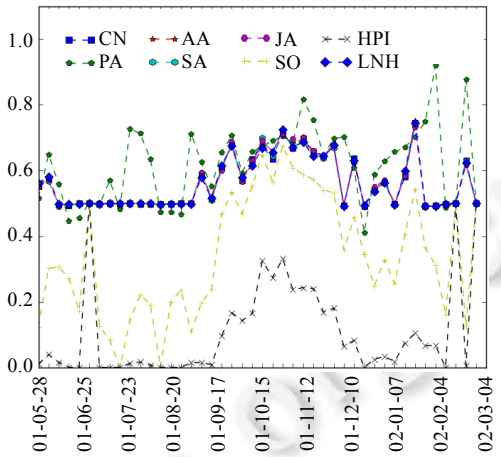

Fig.4 Link prediction precision comparison on Enron dataset

图 4 Enron 链路预测精度对比图

Table 7 Each kinds index results of link prediction on Enron dataset

表 7 Enron 链路预测各指标结果

Index	CN	PA	AA	SA	**JA**	SO	HPI	LNH
AUC	0.858 5	0.713	0.86	0.856	**0.871 5**	0.865 5	0.86	0.844 5

4.3 SimC的有效性验证

第 4.2 节利用链路预测推断了 VAST 数据集的演化机制,选取了 PAS 作为最优指标,但在实际进行相似性计算时,PAS 与其他指标具体表现如何呢?PAS 是否一定比其他指标更加优秀?另外,文献[31,32]分别提出 EI,TD 两种相似性计算方法,SimC 较这两种方法效果如何呢?为此,本文设计了对比实验来加以验证.

在 VAST 数据集上,分别采用各种指标下的 SimC 与 EI,TD,得到演化序列 GraphS,见表 8.

Table 8 GraphS of VAST network under different indexes

表 8 各种指标下的 VAST 演化序列

Index	$\hat{D}(g^2 \| g^1)$	$\hat{D}(g^3 \| g^2)$	$\hat{D}(g^4 \| g^3)$	$\hat{D}(g^5 \| g^4)$	$\hat{D}(g^6 \| g^5)$	$\hat{D}(g^7 \| g^6)$	$\hat{D}(g^8 \| g^7)$	$\hat{D}(g^9 \| g^8)$	$\hat{D}(g^{10} \| g^9)$
CNS	0.324	0.329	0.327	0.323	0.321	0.320	0.335	0.305	0.326
PAS	0.048	0.049	0.047	0.048	0.048	0.049	0.061	0.045	0.048
AAS	0.693	0.691	0.689	0.679	0.673	0.675	0.726	0.633	0.687
SAS	1.135	1.130	1.137	1.125	1.111	1.102	1.121	1.074	1.132
JAS	1.346	1.331	1.347	1.325	1.309	1.295	1.355	1.243	1.338
SOS	1.261	1.250	1.259	1.244	1.239	1.235	1.303	1.204	1.255
HPIS	1.095	1.088	1.088	1.086	1.078	1.084	1.118	1.061	1.087
LNHS	3.719	3.657	3.702	3.677	3.624	3.604	3.806	3.590	3.707
EI	0.154	0.147	0.149	0.145	0.144	0.146	0.175	0.129	0.148
TD	0.854	0.823	0.838	0.810	0.801	0.803	0.913	0.731	0.851

为了统一对比,将各个指标的演化序列进行归一化处理,结果如图 5 所示,其中,横坐标标注了网络变化的时间点.在图 5 中可见,PAS 在 1~7 天的网络平稳段内变化幅度最小,在第 7 天~第 8 天的值相较之前值变化最大,表现最好.为了量化,应用第 3.3.3 节的事件敏感表现 *Per* 评价各个指标的表现,结果见表 9.

通过表 9 可见,PAS 指标的 *Per* 值远远高于其他.这恰好验证了 PAS 作为最优指标是正确的,同时也表明 SimC 较之 EI,TD 更加高效.

同样,第 4.2 节在 Enron 网络中选取 JA 指标作为最优指标,采用 JAS 下的 SimC,计算其网络演化序列,并与 EI,TD 两种指标进行比较,如图 6 所示.由于 Per 表现是针对一次事件来计算的,Enron 网络含有多个事件,故此时