软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2016,27(11):2747−2762 [doi: 10.13328/j.cnki.jos.004910]

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

混合指标量子群智能社会网络事件检测方法*

胡文斌, 王 欢, 严丽平, 邱振宇, 肖 雷, 杜 博

(武汉大学 计算机学院,湖北 武汉 430072)

通讯作者: 胡文斌, E-mail: hwb@whu.edu.cn, http://www.whu.edu.cn

摘 要: 社会网络错综复杂,如果能够及时发现和预测当前网络可能发生的重大事件并采取有效的处置策略,将具有重大意义.链路预测的理论框架和评价方法为社会网络事件检测提供了一条有效途径.目前,链路预测的研究工作大多针对特定网络提出相似性指标,试图取得更高的链路预测精度.这些研究存在如下问题:(1) 不同的相似性指标适用于不同的网络,不具有普适性;(2) 独立的相似性指标无法全面反映网络演化的多样性和复杂性;(3) 链路预测时未考虑网络演化过程中可能出现波动,无法进行事件检测.基于上述问题,提出一种社会网络事件检测的混合指标群智能方法 IndexEvent,由最佳权重算法 OWA(optimal weight algorithm)和波动检测算法 FDA(fluctuation detection algorithm)组成,可以评价不同网络的演化波动,发现网络波动异常,进行事件检测.主要工作如下:(1) 提出了混合指标,并证明了基于混合指标的链路预测算法可以取得更高的预测精度;(2) 基于量子粒子群算法提出了最佳权重算法 OWA,以高效地确定不同网络的最佳混合指标;(3) 提出了一种网络波动检测算法 FDA,定量评价不同时段网络演化的波动程度,并在考虑微观因素的基础上进行改进.对不同特征的网络进行实验,结果表明,IndexEvent 方法能够准确地反映事件造成的网络演化波动,有效地检测事件.

关键词: 量子粒子群;事件检测;链路预测;社会网络;网络演化;网络波动性评价

中图法分类号: TP18

中文引用格式: 胡文斌,王欢,严丽平,邱振宇,肖雷,杜博.混合指标量子群智能社会网络事件检测方法.软件学报, 2016,27(11): 2747−2762. http://www.jos.org.cn/1000-9825/4910.htm

英文引用格式: Hu WB, Wang H, Yan LP, Qiu ZY, Xiao L, Du B. Hybrid quantum swarm intelligence indexing for event detection in social networks. Ruan Jian Xue Bao/Journal of Software, 2016,27(11):2747−2762 (in Chinese). http://www.jos.org.cn/1000-9825/4910.htm

Hybrid Quantum Swarm Intelligence Indexing for Event Detection in Social Networks

HU Wen-Bin, WANG Huan, YAN Li-Ping, QIU Zhen-Yu, XIAO Lei, DU Bo

(Computer School, Wuhan University, Wuhan 430072, China)

Abstract: In complicated social networks, discovering or predicting important events is significant. The theoretical framework and evaluation methods of link prediction offer an effective solution for detecting events in social networks. Most of the current research focuses on proposing different similarity indexes to achieve higherlink prediction accuracy. However this type of approach has following problems: (1) Because different similarity indexes are designed for different networks, they are not universal; (2) The independent similarity index is difficult to reflect diversity and complexity of real network evolutions; (3) Without considering the fluctuation in the network evolution, the link prediction cannot detect events. To solve these problems, this paper proposes a swarm intelligence method

* 基金项目: 国家重点基础研究发展计划(973)(2012CB719905); 国家自然科学基金(61572369, 61471274); 湖北省自然科学基金(2015CFB423); 武汉市重大科技计划项目(2015010101010023)

Foundation item: National Program on Key Basic Research Project of China (973) (2012CB719905); National Natural Science Foundation of China (61572369, 61471274); National Natural Science Foundation of Hubei Province (2015CFB423); Wuhan Major Science and Technology Program (2015010101010023)

收稿时间: 2015-06-03; 修改时间: 2015-08-11; 采用时间: 2015-08-24

based on mixed indexes (IndexEvent), which can evaluate fluctuations and detect events in social networks. The main work is as follow: (1) A proof is provided on the proposed mixed indexes that the link prediction algorithm based on mixed indexes can achieve a higher accuracy; (2) Based on the quantum-behaved particle swarm algorithm, an optimal weight algorithm (OWA) is developed to determine best mixed indexes for different networks efficiently; (3) A fluctuation detection algorithm (FDA) is designed to quantitatively estimates fluctuations in network evolutions at different periods. And micro factors are taken into account to improve FDA. The results of the experiments show that IndexEvent can effectively reflect evolution fluctuations and detect events.

Key words: quantum particle swarm; event detection; link prediction; social network; network evolution; network fluctuation evaluation

社会网络是指社会不同个体成员之间因互动而形成的相对稳定的关系体系.社会网络事件检测通常通过检测各个时段的网络演化的变化,分析出网络演化的异样,从而检测出当前网络发生的事件,提出干预和处置策略[1].社会网络事件检测具有广泛的应用场景和极大的实用价值,例如,它可以分析犯罪网络中核心头目的更替、预测公司人员结构调整影响、分析股票波动,进行舆情检测等[2].

在真实的社会网络中,很多事件的发生都有可能导致网络偏离正常的演化机制,产生异常的网络演化波动.如何基于当前复杂的社会网络,快速、准确地检测出当前网络发生的重大事件,评估不同事件产生的影响,并且提出有效的处置策略,是社会网络事件检测面临的重大挑战.目前,社会网络事件检测的主要方法如下:(1) 直接建立网络演化模型,调整模型参数,使产生的网络更加接近真实网络,仿真各个时间段的网络结构,发现异常网络结构.比较典型的模型有 Watts-Strogatz(WS)小世界模型[3]、Barabási-Albert(BA)无标度模型[4].但这些网络演化模型都处于理论探究阶段,尚不能应用在真实网络上.(2) 基于网络结构图的分析方法进行事件检测,涉及图形数据挖掘、数理统计、机器学习等理论,常见方法有图形模式识别[5]、图形相似度比较[6]、统计过程控制[7]、扫描统计[8].但是这些方法计算量巨大,适用范围有限,忽视了网络演化的动态性.

本文借助链路预测的理论框架和评价方法提出了一种新的社会网络事件检测思路,量化网络演化波动,发现异常波动,从而实现事件检测.链路预测是指利用网络的结构或节点的属性信息来预测未产生连边的两节点间产生连边的可能性.常见的链路预测方法有基于马尔可夫链或机器学习的方法[9]和基于似然分析的方法[10],但这两种方法都存在计算复杂度过高的问题.与这两类方法相比,基于相似性的链路预测方法更简单、高效,并且通常能够取得很好的预测效果.具体步骤:利用某种相似性指标来计算当前时刻不存在连边的两节点之间产生连边的可能性得分,然后根据每条边的得分值进行降序排列,选取排列靠前的一定数目的边作为预测结果输出.目前,基于相似性的链路预测研究侧重于提出新的相似性指标来取得更好链路预测精度[11].然而,真实网络的演化机制纷繁复杂[12],很难通过基于单一的相似性指标去准确刻画.现有的相似性指标都是针对具有特定拓扑结构性质的网络才可能有最佳效果,缺乏普适性.同时,现有的相似性指标之间的协作关系缺乏探究,每个相似性指标都被独立用于链路预测.

为了解决独立相似性指标的精度不足、缺乏普适性和协作的问题,本文提出了混合指标,并在此基础上提出了一种社会网络事件检测方法 IndicesEvent,包含最佳权重算法 OWA(optimal weight algorithm)和波动检测算法 FDA(fluctuation detection algorithm).OWA 可以自动高效地确定不同社会网络对应的最佳混合指标.同时,FDA 借助最佳混合指标对不同时刻网络演化的波动进行量化评价,检测重大事件的发生,并在考虑微观因素的基础上进行改进 FDA.综上,本文主要工作可以总结如下:

(1) 提出 OWA 来高效地确定混合指标中各单位指标的最佳权重,可以得到不同社会网络对应的最佳混合指标.解决现有链路预测的相似性指标缺乏普适性、相互之间独立无法协作预测等问题;
(2) 提出一种网络波动检测算法 FDA,并在考虑节点微观演化规律因素的基础上进行改进,解决链路预测无法检测网络演化波动异常的问题;
(3) 结合 OWA 和 FDA 提出一种通用的社会网络事件检测方法 IndexEvent,可以准确地检测出网络中发生的重大事件,并对事件影响进行量化评价.

本文第 1 节介绍相关研究工作.第 2 节详细介绍 IndexEvent 方法,并借助 WS 小世界模型的实例、BA 无标度模型的实例来提出最佳权重算法 OWA 和网络波动检测算法 FDA.第 3 节通过真实的通信网络和邮件网络来

确定验证 IndexEvent 方法的效果,并分析实验结果.第 4 节是结论及展望.

1 相关工作

在链路预测方面,Sarukkai[13]基于马尔可夫链进行网络的链路预测分析.Zhu 等人[14]率先在自适应性网站的预测中使用基于马尔可夫链的预测方法.Newman 等人[15]认为很多网络的连接可以反映内在的层次结构,提出了一种最大似然估计的算法进行链路预测,该方法在处理像草原食物链这样具有明显网络层次组织的网络时具有较好的精度.Guimera 等人[16]提出一种基于随机分块模型的链路预测方法,该模型中,节点被分为若干集合,两个节点间连接的概率只与相应的集合相关.该方法不仅可以预测缺失边,还可以预测网络的错误链接,例如纠正蛋白质相互作用网络中的错误链接.然而,基于最大似然的方法[15,16]计算复杂度太高,不适合于在规模大的网络中应用.Liben-Nowell 等人[17]提出了基于网络拓扑结构的相似性定义方法,把指标分为基于节点和基于路径两类,并分析了部分指标在社会合作网络中链路预测的效果.Kleinberg 等人[11]通过对比多种相似性指标在链路预测中的表现(共同邻居[18]、Jaccar 系数[19]、Adamic/Adar[20]、优先链接[21]等),系统地阐述了链路预测问题.Lichtenwalter 等人[22]提出了一种引入监督学习的链路预测方法,效果比无监督学习的方法提高了 30%以上.Symeonidis 等人[23]通过引入多条路径信息提出了多路谱聚类方法,提高了在社会网络和蛋白质作用网络上的链路预测精度.Huang 等人[24]在得到节点的直接相似性后,利用协同过滤技术对相似性指标进行一轮加权处理,取得了较好的效果.Rao 等人[25]为了将链路预测应用到大规模网络,实现了基于 MapReduce 的计算模型的链路预测算法.此外,文献[26–29]考虑两端节点度的影响,从不同角度提出了其他相似性指标,但是大多集中在链路预测自身机制探讨或提高预测精度上,缺乏实际应用研究.

在社会网络事件检测方面,Noble 等人[30]提出了通过迭代比较发现异常网络结构的方法,以及通过子结构条件熵来量化图形结构的异常程度.但是该方法主要注重于理论研究,实际计算非常复杂.Papadimitriou[6]面向较大规模的 Web 网络,充分考虑了节点和边的重要性,通过检测网络异常来判断服务器、爬虫程序等是否有异常发生.McCulloh 等人[7]提出了专门的社会网络变化检测方法,有效地屏蔽正常波动带来的干扰,将网络实质变化从正常波动中分离出来,但是这种检测方法要求网络参数满足正态分布,致使其应用范围受到极大的限制.Priebe 等人[31]采用扫描统计量方法对邮件网络进行检测,发现了网络中的“震荡”区域.Wan 等人[8]通过分别检测网络节点网络参数的变化和相对于社团通联结构的变化,发现邮件网络的异常事件.但 Priebe[31]和 Wan[8]的工作由于要计算的统计量过多,导致计算量巨大.Wu 等人[32]和 Baruah 等人[33]提出的网络相似性计算方法都由于没有统一各种因素的影响,导致最终结果不具参考性.Qiao 等人[34]对犯罪成员的邮件网络进行分析,挖掘出犯罪组织的主要成员,发现异常通信事件.

综上,现有的社会网络的事件检测问题仍缺乏有效的解决方案.本文提出通用的 OWA 来寻找最匹配当前网络的链路预测指标和 FDA 来精确检测网络波动性异常,并且结合 OWA 和 FDA 构建了一种高效的社会网络事件检测方法 IndexEvent.

2 IndexEvent 方法框架

真实的社会网络具有统一性、多样性和复杂性的特点,很难预知一个真实网络的演化机制,并构建合适的网络演化模型进行分析.链路预测和网络演化具有内在一致性[35],借助社会网络数据集得到的不同时段网络拓扑结构信息,利用链路预测可以评价不同时段网络演化的波动.本文提出的 IndexEvent 方法依次对各时间段的演化波动进行定量评价,通过评价结果去检测网络事件的发生.IndexEvent 方法框架如图 1 所示,它包含算法 OWA 和 FDA.

(1) 基于量子粒子群算法 QPSO(quantum-behaved particle swarm algorithm)的 OWA 能够高效地确定给定时段网络演化的最佳混合指标.

(2) FDA 算法借助最佳混合指标,量化不同时段的网络演化波动,得到事件检测序列.事件检测序列中事件检测值越低,其对应时段的网络演化波动越大,发生事件的可能性也就越大.

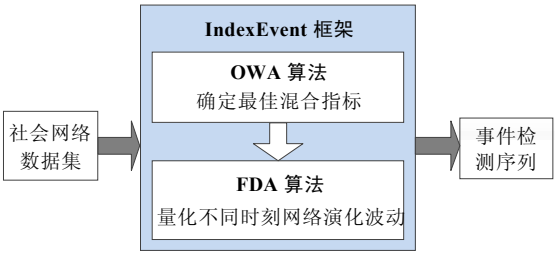


Fig.1 Framework of method IndexEvent

图 1 IndexEvent 方法框架

IndexEvent 方法的实现步骤如图 2 所示,以 OWA 和 OWA 为基础实现 IndexEvent 方法的事件检测.

输入:社会网络的数据集.
1. 执行算法 OWA,得到起始时段网络的最佳混合指标.
2. 执行算法 OWA,得到各时段的事件检测值,构成事件检测序列.
3. 对事件检测序列进行降序排序,输出排序靠前的事件检测值对应的时段.

Fig.2 Implementation steps of method IndexEvent

图 2 IndexEvent 方法的实现步骤

2.1 算法OWA

第 2.1.1 节提出了混合指标的概念.第 2.1.2 节详细介绍了 OWA 如何确定最佳混合指标.同时,在第 2.1.1 节和第 2.1.2 节中,我们都会结合 WS 小世界网络进行详细解释.

2.1.1 混合指标

基于相似性指标的链路预测方法是一种简约、高效的方法,目前已有许多节点相似指标被提出,常见相似性指标及其计算方法见表 1.

Table 1 Common similarity indexs

表 1 常见的相似性指标

名称	定义	名称	定义
共同邻居指标(CN)[18]	$S_{ij}=\|\Gamma(i)\cap\Gamma(j)\|$	Jaccard 指标(JA)[19]	$S_{ij}=\frac{\|\Gamma(i)\cap\Gamma(j)\|}{\|\Gamma(i)\cup\Gamma(j)\|}$
优先链接指标(PA)[21]	$S_{ij}=k(i)\times k(j)$	Sorenson 指标(SO)[27]	$S_{ij}=\frac{2\|\Gamma(i)\cap\Gamma(j)\|}{k(i)+k(j)}$
Adamic-Adar 指标(AA)[20]	$S_{ij}=\sum_{z\in\Gamma(i)\cap\Gamma(j)}\frac{1}{\lg k(z)}$	大度节点有利指标(HPI)[28]	$S_{ij}=\frac{\|\Gamma(i)\cap\Gamma(j)\|}{\min\{k(i),k(j)\}}$
Salton 指标(SA)[26]	$S_{ij}=\frac{\|\Gamma(i)\cap\Gamma(j)\|}{\sqrt{k(i)\times k(j)}}$	LNH-I 指标(LNH)[29]	$S_{ij}=\frac{\|\Gamma(i)\cap\Gamma(j)\|}{k(i)\times k(j)}$

通过相似性指标计算出的两个节点之间相似性得分越大,它们之间存在连边的可能性就越大.其中,$S(i,j)$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 的度为 $k(i)=|\Gamma(i)|$.一个真实网络的演化往往交杂多种网络演化机制,基于一种相似性指标的链路预测算法很难全面地刻画真实网络的演化.

为了更全面地刻画真实网络的演化,本文基于现有的节点相似性指标提出了一种混合指标的概念,称为 MixSimIndex,用公式(1)表示.

$$MixSimIndex=F(w_1,...,w_m,...,w_n,SimIndex_1,...,SimIndex_m,...,SimIndex_n),SimIndex_m\in\phi,\sum_{m=1}^{n}w_m=1 \quad (1)$$

其中,$SimIndex_m$ 是特定的相似性指标,称为 MixSimIndex 的单位指标.w_m 为对应的单位指标 $SimIndex_m$ 的权重.n 是所选定的相似性指标的数目,即单位指标的数目.ϕ代表相似性指标集合,以后研究中提出的新的相似性指标

也可以不断加入该集合中.单位指标都源于集合ϕ.F 表示混合指标 MixSimIndex,MixSimIndex 和各单位指标及权重之间的函数关系,增强了混合指标形式的灵活性.

$$MixSimIndex=\sum_{m=1}^{n}w_m\times SimIndex_m, SimIndex_m\in\phi, \sum_{m=1}^{n}w_m=1 \tag{2}$$

本文采用公式(2)表示 MixSimIndex,集合ϕ只考虑表 1 中指出的 8 种相似性指标.例如,*MixSimIndex*=0.7×*AA*+0.1×*SA*+0.2×*CN*.基于 MixSimIndex 可计算节点对 i 和 j 的相似性得分为

$$S(i,j)=0.7\times\sum_{z\in\Gamma(i)\cap\Gamma(j)}\frac{1}{\lg^{k(z)}}+0.1\times\frac{|\Gamma(i)\cap\Gamma(j)|}{\sqrt{k(i)\times k(j)}}+0.2\times|\Gamma(i)\cap\Gamma(j)|.$$

同时,由公式(2)可知,独立的相似性指标(如表 1 中的 CN,JA,PA,SO,AA,HPI,SA,LNH 等)属于 MixSimIndex 的特例.例如,集合ϕ中只有 CN 被选作为单位指标,则 CN 的权重一定是 1,*MixSimIndex*=*CN*,此时,混合指标就等同于独立的相似性指标.

链路预测算法精度的衡量指标主要有 AUC[36]、精确度(precision)[37]和排序分(ranking score)[38].它们对精度判定的侧重点不同.Precision 只考虑排在前 L 位的边是否准确预测,而 Ranking Score 更多地考虑了所预测的边的排序.AUC 是最常用的一种衡量指标,它从整体上衡量算法的精度.

本文选取 AUC 作为衡量指标,其定义用公式(3)表示:

$$AUC=\frac{n'+0.5n''}{n} \tag{3}$$

其中,n 表示比较的次数,n'表示从测试集中随机选择边的得分大于从不存在边构成集合中随机选择边的得分的次数,n''表示相等的次数.AUC 反映了链路预测算法的预测精度,值越大,说明所对应的指标越好,越符合当前的网络演化机制.如果所有节点对的得分是随机产生的,则理想情况下,*AUC*=0.5.当 *AUC*>0.5 时,才能表明相似性指标的有效性,所以我们规定ϕ中满足 AUC 大于 0.5 的相似性指标才能被选为单位指标.

Watts 等人提出了著名的 WS 小世界网络[3],它介于规则网络和随机网络之间.为了更好地解释混合指标,我们构造 WS 小世界网络实例:生成 200 个节点,每个节点有 4 个近邻居,以 0.3 的概率随机化重连边.所构造实例特性见表 2.

Table 2 Structure of WS small world network characteristics

表 2 构造的 WS 小世界网络特性

节点	边	群聚系数	直径	平均距离
200	400	0.140 38	8	4.361 6

利用表 2 中的相似性指标对所构造的 WS 小世界网络实例进行链路预测,当前网络存在的所有的边作为测试集合,当前网络所有节点间不存在的边构成不存在边集合.同时,由于所构建的 WS 小世界网规模较小,为了更好地说明混合指标的有效性,本节将测试集和不存在边集合中的所有边都进行一一比较,防止边选取时的随机性导致 AUC 的波动.基于表 2 中各相似性指标链路预测算法对应的 AUC 值比较见表 3.

Table 3 AUC values of common similarity indexs

表 3 常见相似性指标 AUC 值

Index	CN	PA	AA	SA	JA	SO	HPI	LNH
AUC	0.633 0	0.601 3	0.638 4	0.636 7	0.633 4	0.636 8	0.636 5	0.636 6

由于表 3 中的 8 种相似性指标的 AUC 值都大于 0.5,所以我们可以灵活地选取一定数目的指标作为混合指标的单位指标.以单位指标组合 AA 和 HPI,CN 和 PA 为例,如表 4 所示,当 MixSimIndex 为 0.5×*AA*+0.5×*HPI* 和 0.9×*CN*+0.1×*PA* 时,它们得到的 AUC 值都高于表 3 中的 8 个独立指标;但是当 MixSimIndex 为 0.1×*AA*+0.9×*HPI* 时,它们得到的 AUC 值低于单独相似性指标 AA 的 AUC 值;当 MixSimIndex 为 0.1×*CN*+0.9×*PA* 时,它们得到的 AUC 值低于表 3 中的 8 个单独相似性指标.由表 4 可知,混合指标的存在是有意义的.但并不是所有的混合指标都能取得比常见独立相似性指标更高的链路预测精度,只有当混合指标中的单位指标被赋予合适权重的情况

下,才能取得比独立相似性指标更高的链路预测精度.

Table 4 Example of hybrid index AUC value

表 4 混合指标 AUC 值举例

MixSimIndex	0.5×*AA*+0.5×*HPI*	0.1×*AA*+0.9×*HPI*	0.9×*CN*+0.1×*PA*	0.1×*CN*+0.9×*PA*
AUC	0.638 7	0.637 6	0.690 0	0.620 3

2.1.2 OWA 的实现

上一节已经论证了混合指标的可行性,混合指标的优劣与其各单位指标的权重息息相关.如果各单位指标权重确定后的混合指标的链路预测效果最佳,则称其为最佳混合指标,即最符合当前网络演化机制的混合指标.如何快速确定最佳混合指标,将是本节致力解决的问题.

量子信息的基本存储单元是量子比特,$|0\rangle$和$|1\rangle$表示一个量子比特的两种极化状态[39].量子比特状态可表示为 $P_{ic}|0\rangle+P_{is}\beta|1\rangle$,$P_{ic}$和$P_{is}$分别表示量子位状态$|0\rangle$和$|1\rangle$的概率幅.Tang 等人[40]将量子机制融入到原始粒子群优化算法(particle swarms optimization,简称 PSO)[41]中,提出了量子粒子群优化算法(quantum-behaved particle swarm optimization,简称 QPSO).与原始 PSO 相比,QPSO 需要设置的参数减少,并且有更强的寻优能力.本文提出基于 QPSO 的算法 OWA 快速确定混合指标中各单位指标的最佳权重,生成最佳混合指标.

假设已确定最佳混合指标中的有 n 单位指标 $SimIndex_1,SimIndex_2,...,SimIndex_n$,它们各自对应的权重为 $w_1,w_2,...,w_n$,可组成权重数组 $W=(w_1,w_2,...,w_n)$.$fitness(MixSimIndex(W))$表示权重数组 W 对应的混合指标的适应度值,即在权重数组 $W=(w_1,w_2,...,w_n)$对应的混合指标 $MixSimIndex=\sum_{m=1}^{n}w_m\times SimIndex_m$ 上进行链路预测最终得到的衡量指标值.适应度值越大,则表示权重数组 W 生成的混合指标越优秀.算法 OWA 具体可分为 3 个步骤:

(1) 产生携带权重数组的初始量子粒子群:量子粒子群中每个量子态粒子编码方式如公式(4)所示:

$$P_i=\begin{bmatrix}\cos(\theta_{i1})&\cos(\theta_{i2})&...&\cos(\theta_{in})\\ \sin(\theta_{i1})&\sin(\theta_{i2})&...&\sin(\theta_{in})\end{bmatrix} \tag{4}$$

其中,$\theta_{ij}=2\pi\times rnd$,$rnd$ 为(0,1)之间的随机数;$i=1,2,\ldots,m$,$j=1,2,\ldots,n$,m 是量子粒子群中粒子数目,粒子数目越多,越能增加初始化时权重数组的差异性;n 表示最佳混合指标中的有 n 单位指标.每个量子态粒子占据的两个位置分别对应于概率幅 P_{is} 和 P_{ic},可用公式(5)和公式(6)表示:

$$P_{is}=(\sin(\theta_{i1}),\sin(\theta_{i2}),\ldots,\sin(\theta_{in})) \tag{5}$$

$$P_{ic}=(\cos(\theta_{i1}),\cos(\theta_{i2}),\ldots,\cos(\theta_{in})) \tag{6}$$

每个量子态粒子的概率幅 P_{is} 和 P_{ic} 可以通过公式(7)和公式(8)转化为对应的权重数组 W_{is} 和 W_{ic},W 可以取值为 W_{is} 或 W_{ic}:

$$W_{is}=\left(\frac{\sin(\theta_{i1})}{\sum_{m=1}^{n}\sin(\theta_{im})},\frac{\sin(\theta_{i2})}{\sum_{m=1}^{n}\sin(\theta_{im})},...,\frac{\sin(\theta_{in})}{\sum_{m=1}^{n}\sin(\theta_{im})}\right) \tag{7}$$

$$W_{ic}=\left(\frac{\cos(\theta_{i1})}{\sum_{m=1}^{n}\cos(\theta_{im})},\frac{\cos(\theta_{i2})}{\sum_{m=1}^{n}\cos(\theta_{im})},...,\frac{\cos(\theta_{in})}{\sum_{m=1}^{n}\cos(\theta_{im})}\right) \tag{8}$$

(2) 权重数组更新:权重数组的更新是由概率幅 P_{is} 和 P_{ic} 的更新实现的.每次迭代,都将 P_{is} 和 P_{ic} 通过公式(9)得到 $\overline{P}_{is}$ 和 $\overline{P}_{ic}$,然后 $P_{is}=\overline{P}_{is}$,$P_{ic}=\overline{P}_{ic}$,实现更新,继续进行下次迭代:

$$\begin{cases}\tilde{P}_{ic}=(\cos(\theta_{i1}(t)+\Delta\theta_{i1}(t+1)),...,\cos(\theta_{in}(t)+\Delta\theta_{in}(t+1)))\\ \tilde{P}_{is}=(\sin(\theta_{i1}(t)+\Delta\theta_{i1}(t+1)),...,\sin(\theta_{in}(t)+\Delta\theta_{in}(t+1)))\end{cases} \tag{9}$$

其中,

$$\Delta\theta_{ij}(t+1)=w\Delta\theta_{ij}(t)+c_1r_1(\Delta\theta_l)+c_2r_2(\Delta\theta_g),$$

$$\Delta\theta_l=\begin{cases}2\pi+\theta_{ilj}+\theta_{ij}(\theta_{ilj}-\theta_{ij}<-\pi)\\ \theta_{ilj}-\theta_{ij}(-\pi\leqslant\theta_{ilj}-\theta_{ij}\leqslant-\pi),\\ \theta_{ilj}-\theta_{ij}-2\pi(\theta_{ilj}-\theta_{ij}>\pi)\end{cases}$$

$$\Delta\theta_g=\begin{cases}2\pi+\theta_{gj}+\theta_{ij}(\theta_{gj}-\theta_{ij}<-\pi)\\ \theta_{gj}-\theta_{ij}(-\pi\leqslant\theta_{gj}-\theta_{ij}\leqslant-\pi).\\ \theta_{gj}-\theta_{ij}-2\pi(\theta_{gj}-\theta_{ij}>\pi)\end{cases}$$

(3) 权重数组变异处理:原始的 PSO 算法易陷入混合指标局部最优,主要原因在于搜索过程中权重数组的多样性的丢失.算法 OWA 借助量子非门实现变异操作[42]来避免多样性丢失.设权重数组变异概率为 P_m,在(0,1)之间随机生成 rnd_i,如果 $rnd_i<P_m$,则随机选择该量子态粒子上$\lceil n/2\rceil$个量子比特,通过公式(10)进行变异操作.

$$\begin{bmatrix}0 & 1\\ 0 & 1\end{bmatrix}\begin{bmatrix}\cos(\theta_{ij})\\ \sin(\theta_{ij})\end{bmatrix}=\begin{bmatrix}\sin(\theta_{ij})\\ \cos(\theta_{ij})\end{bmatrix}=\begin{bmatrix}\cos\left(\frac{\pi}{2}-\theta_{ij}\right)\\ \sin\left(\frac{\pi}{2}-\theta_{ij}\right)\end{bmatrix}\tag{10}$$

设有 m 个量子态粒子共经历了 $g_{\max}$ 次迭代优化.P_{il} 对应的权重数组 W_{il},为粒子 i 当前搜索到的适应度值最高的权重数组.P_g 对应的权重数组为 W_g,为整个粒子群当前搜索到的适应度值最大的权重数组.

OWA 具体实现如下:

算法 1. OWA.

输入:社会网络数据集.

1. 选定衡量指标.
2. 确定最佳混合指标的单位指标.
3. 通过公式(3)初始化,生成 m 个携带权重数组的量子态粒子.
4. For g=1 to $g_{\max}$:
   For i=1 to m:
5. 量子态粒子概率幅 P_{ic} 和 P_{is} 通过公式(7)和公式(8)对应的权重数组 W_{is} 和 W_{ic}.
   If $fitness(MixSimIndex(W_{ic}))>fitness(MixSimIndex(W_{il}))$ then $P_{il}=P_{ic}$ End
   If $fitness(MixSimIndex(W_{is}))>fitness(MixSimIndex(W_{il}))$ then $P_{il}=P_{is}$ End
   If $fitness(MixSimIndex(W_{il}))>fitness(MixSimIndex(W_g))$ then $P_g=P_{il}$ End
6. 通过公式(9)更新权重数组, $P_{is}=\overline{P}_{is}, P_{ic}=\overline{P}_{ic}$.
7. 通过公式(10)对权重数组进行变异处理.
   End
   End
8. P_g 通过公式(7)或者公式(8)转换为对应的权重数组,确定最佳混合指标并输出.

算法 OWA 可以自动计算各单位指标的合适权重,快速、高效地确定最佳的混合指标,避免了考虑单位指标数量级的差异.对于本文采用的衡量指标 AUC,最佳混合指标的单位指标只需满足 AUC 值大于 0.5.当有多个相似性指标都满足单位指标的要求时,可以把所有满足条件的指标都作为 MixSimIndex 的单位指标.在 OWA 迭代足够多次数的情况下,不利于 AUC 值提升的单位指标的权重会被确定为 0.但实际操作中,可适当选取一定数目满足条件的相似性指标作为单位指标,比如,简单选取 AUC 值最大的两个相似性指标,时间复杂度可由 $O(n\times m\times g_{\max})$降低到 $O(2\times m\times g_{\max})$,这会有效地提升 OWA 的效率.

为了更好地解释算法 OWA,我们继续探讨表 2 中构造的 WS 小世界网络.如果 n=2,仅仅选定 PA 和 CN 两个指标作为 MixSimIndex 单位指标,设其对应的权重分别为 w_1,w_2.设量子粒子群规模为 150,通过 OWA 算法在 200 次迭代后,可得到 w_1=0.97,w_2=0.03.此时的 AUC 值可高达 0.708 5,远高于独立的相似性指标(表 3)和没有优

化的一般混合权重指标(见表 4).同时,如表 5 所示,在相同的实验设置下,基本的 PSO[40]得到的最佳混合指标 AUC 值为 0.691 2,低于 OWA 的表现,也证明了 OWA 的高效性.

Table 5 Algorithm OWA and basic PSO efficiency correspondenc

表 5 算法 OWA 和基本的 PSO 效率对应

	OWA	PSO
粒子群规模	150	150
迭代次数	200	200
最佳混合指标 AUC 值	0.708 5	0.691 2

2.2 算法FDA

第 2.2.1 节在 BA 无标度网络实例上验证了链路预测衡量指标值变动与网络演化波动的一致性.第 2.2.2 节详细介绍了算法 FDA,利用链路预测衡量指标量化不同时刻的网络演化波动,发现异常波动,进行事件检测.同时,在第 2.2.1 节和第 2.2.2 节中,我们都会结合 BA 无标度网络进行详细解释.

2.2.1 一致性分析

给定的网络 G 在 t 时刻的网络快照可用 g^t 表示,假设现有网络 G 的 n 个快照构成集合$\{g^1,g^2,\ldots,g^t,\ldots,g^n\}$,每两个相邻快照之间的时间间隔是一致的.第 2.1 节中的 OWA 已能在给定网络快照的情况下,高效确定当前时段的最佳混合指标.基于最佳混合指标的链路预测算法,能够取得最好的链路预测效果(最高的衡量指标值).因为网络演化机制与链路预测算法具有内在的一致性[35,43],所以基于最佳混合指标的链路预测算法最符合当前网络的演化机制.正常情况下,当前网络的演化机制会在一段时间内保持一定程度的稳定性,所以基于当前最佳混合指标的链路预测算法会在一段时间内取得较高的衡量指标值.当某时刻最佳混合指标的衡量指标值发生了显著的下降,可以猜测发生了网络事件,扰乱了网络的内在演化机制,造成衡量指标值的下降.

比如,基于网络快照 g^t 得到 t 时刻的最佳混合指标,表示为 $BMixSimIndex_t$.如果没有网络事件发生,那么基于 $BmixSimIndex_t$ 的链路预测算法应该在后续的 t+1 和 t+2 时刻对应的网络快照 g^{t+1},g^{t+2} 上仍然取得较高的 AUC 值.如果基于 $BMixSimIndex_t$ 的链路预测算法在网络快照 g^{t+1} 上保持着较高的 AUC 值,在网络快照 g^{t+2} 上 AUC 值明显下降,那么可以推测 t+1 到 t+2 时段发生了网络事件,扰乱了网络内在的演化机制,导致 $BMixSimIndex_t$ 不再符合 t+1 到 t+2 时段的网络演化机制,所以基于 $BMixSimIndex_t$ 的链路预测算法不能在网络快照 g^{t+2} 取得较高的 AUC 值.

Barabási 等人基于优先连接的网络演化机制提出了著名的 BA 无标度模型[4].为了更好地解释链路预测衡量指标变动与网络演化波动的一致性,我们构建 BA 网络实例 $N1$ 和加入事件的 BA 网络实例 $N2$.$N1$ 构造步骤如下:(1) 初始网络为空,第 1 个时间步加入一个节点;(2) 从第 2 个时间步开始,每次加入 1 个新节点,并且每个新节点优先与现有的度最大的节点构成一条边;(3) 迭代 200 个时间步后停止.$N2$ 构造步骤如下:

(1) 正常网络演化阶段:第 1 个时间步加入一个节点.从第 2 个时间步开始,每次加入 1 个新节点,并且每个新节点优先与现有的度最大的节点构成一条边.迭代 120 个时间步后停止.

(2) 网络事件发生阶段:从第 121 时间步开始到第 159 时间步,每次加入 1 个新节点,并且每个新节点随机的与现有节点构成一条边.

(3) 网络演化恢复正常阶段:从第 160 时间步开始到第 200 时间步,每次加入 1 个新节点,每个新节点再次优先与现有的度最大的节点构成一条边.

网络 $N1$ 和 $N2$ 的区别在于,$N2$ 从第 121 时间步到第 159 时间步中采用了不同于 $N1$ 的演化机制来模仿网络事件的发生.AUC^t 表示基于 t 时刻的网络快照 g^t 进行链路预测得到的 AUC 值.

$$AUC=\frac{n'+0.5n''}{n}.$$

t 时刻网络快照 g^t 相对于 t−1 时刻网络快照 g^{t-1} 新增加的边构成测试边集合,t 时刻节点之间不存在的边构成不存在边集合.

为了检测出从第 121 时间步到第 159 时间步发生的网络事件,对网络 $N1$ 和 $N2$ 作如下探究:

(1) 在正常网络演化阶段第 120 时间步结束时,$N1$ 和 $N2$ 的网络快照一样,其各相似性指标的 AUC^{120} 值相同,见表 6.

Table 6 Common similarity index AUC^{120}

表 6 常见相似性指标 AUC^{120}

Index	CN	PA	AA	SA	JA	SO	HPI	LNH
AUC	0.462 2	0.709 6	0.466 4	0.466 4	0.462 2	0.466 4	0.466 4	0.466 4

表 6 中,所有相似性指标中只有 PA 的 AUC 值超过 0.5,所以只能选取 PA 作为最佳混合指标的单位指标,则 PA 的权重一定为 1,第 120 时间步结束时当前网络的最佳混合指标 $BMixSimIndex_{120}=PA$.

(2) 每隔 20 个时间步,利用衡量指标 AUC 评价基于 $BMixSimIndex_{120}=PA$ 和表 1 中各相似性指标的链路预测算法效果,其 AUC 值如图 3 所示.由于 SA,JA,SO,HPI,LNH 与 CN,AA 有相似的曲线,为了让图 3 更清晰,只有 CN,AA 指标被选取展示.PA 在无事件发生的网络 $N1$ 和发生事件的网络 $N2$ 上的变化都在图 3 中进行了展示.

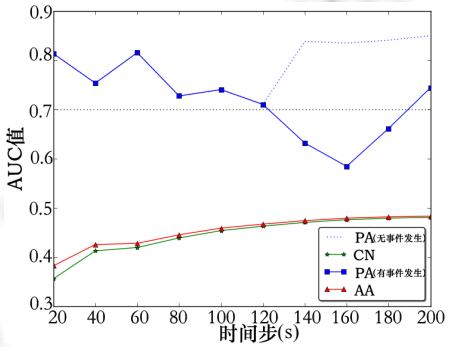


Fig.3 Similarity index corresponding to the index value of the time step change map

图 3 相似性指标对应的衡量指标值随时间步变化图

从图 3 可得出如下结论:

(1) 当选取正确的相似性指标(AUC 值大于 0.5)时,链路预测衡量指标变动与网络演化波动的具有一致性.对于上述 $N2$ 网络的例子,$PA(BMixSimIndex_{120})$可以很好地通过其 AUC 值变动来反映网络演化波动.但是 SA,JA,SO,HPI,LNH,CN 和 AA 一直没有太大的变化,无法反映网络演化波动.PA 指标对应的 AUC 值在正常网络演化阶段(第 1 时间步~第 120 时间步)一直高于 0.7,但是在网络事件发生阶段(第 121 时间步~第 159 时间步),由于网络内在演化机制发生了改变(每个新节点优先与现有的度最大的节点构成一条边变成每个新节点随机的与现有节点构成一条边),PA 对应的 AUC 值迅速下降,远低于 0.7.在网络演化恢复正常阶段(第 160 时间步~第 200 时间步),PA 对应的 AUC 值才逐渐提升.对于没有事件发生的 $N2$ 而言,它的最佳混合指标 $BMixSimIndex_{120}=PA$ 的衡量指标值一直保持高于 0.7.

(2) 通过最佳混合指标对应的衡量指标,可以很好地进行网络事件检测.

基于最佳混合指标 $BMixSimIndex_{120}=PA$,可以很好地反映 $N2$ 的网络事件发生阶段和正常网络演化阶段的差异.

2.2.2 FDA 的实现

第 2.2.1 节验证了链路预测衡量指标变动与网络演化波动的一致性.本节进一步提出了算法 FDA,以有效地避免非事件引起的网络正常波动带来的干扰,发现事件引起的异常网络波动,进行事件检测.假设现有网络快照集$\{g^1,g^2,\ldots,g^t,\ldots,g^n\}$,$M^t$ 代表 t 时刻的网络演化波动的评价值,用公式(11)表示:

$$M^t=\begin{cases}AUC^t-\dfrac{\sum_{T=t-\Delta t}^{t-1}AUC^{T-1}}{\Delta t}, & 0<\Delta t<t,\Delta t\text{为整数}\\ AUC^t, & \Delta t=0\end{cases} \tag{11}$$

其中,Δt 表示记忆时间.t 时刻的网络演化波动值 M^t 是由 AUC^t 与 AUC^{t-1} 至 $AUC^{t-\Delta t}$ 平均值的差值的决定.当多个事件引起连续的网络波动时,Δt 应设置较小值来避免事件间的相互干扰.当Δt=0 时,此时 M^t=AUC^t,相当于 t 时刻网络演化波动的评价值直接由 t 时刻链路预测的 AUC 值来表示.

MD 为事件检测阈值,可以根据需求来灵活设置.当 M^t>MD 时,则认为 t 时刻的网络演化发生了显著异常波动,很可能发生了事件.设置放大系数 A,通过公式(12)对网络演化波动序列(M^1,M^2,…,M^t,…,M^n)中的异常波动进行放大处理,得到事件检测序列(F^1,F^2,…,F^t,…,F^T).F^t 是通过对 M^t 放大处理后得到的 t 时刻的网络事件检测值.F^t 可以为负数值,值越小,表明此时的网络演化波动越大,发生事件的可能性就越大.事件检测序列(F^1,F^2,…,F^t,…,F^T)相对于网络演化波动序列(M^1,M^2,…,M^t,…,M^T),可以更好地区分非事件引起的网络正常波动和事件引起的异常波动.

$$F^t=\begin{cases}A\times M^t, & M^t<MD\\ M^t, & M^t\geqslant MD\end{cases} \tag{12}$$

定义时刻 TO 为网络检测稳定点,一般可选取前后网络演化没有明显波动的时刻作为 TO.事件检测阈值 MD,网络检测稳定点 TO,放大系数 A 可人为设置,增强了算法的灵活性.一种基于 AUC 的算法 FDA 表示如下:

算法 2. 基于 AUC 的 FDA.

输入:事件检测阈值 MD,网络检测稳定点 TO,放大系数 A,记忆时间Δt,网络快照集合{g^1,g^2,…,g^t,…,g^n}.

1. 从网络演化稳定的时刻中选取 TO.
2. 基于 TO 时的网络快照,通过算法 OWA 得到 TO 时最佳混合指标 $BMixSimIndex_{TO}$.
3. For t=1 to n
   执行公式(14),得到 M^t.
   执行公式(15),得到 F^t.
   End
4. 输出事件检测序列(F^1,F^2,…,F^t,…,F^T).

公式(11)是基于衡量指标 AUC 提出的,但衡量指标 AUC 只是宏观上评价整体网络演化波动,并未考虑微观上每个节点演化的差异.实际上,如果节点周围的拓扑结构变化符合网络演化规律,则可以看做是正常演化,其对网络演化波动影响较小.如果节点周围拓扑结构变化不符合演化规律,则极有可能是事件发生导致内在演化规律被打破.考虑节点的微观演化,有助于更精准、全面地量化网络演化波动,进一步避免非事件引起的网络正常波动的干扰.于是,我们进一步提出一种考虑了节点微观演化的衡量指标 $mAUC$,对宏观的 AUC 值进行调整.

将链路预测衡量指标 $AUC=\dfrac{n'+0.5n''}{n}$ 引入到微观节点层面得到 AUC_i^t,求解 AUC_i^t 时,把 t 时刻节点 i 与其他节点相对于 t–1 时刻新增的连边作为测试集合,节点 i 与 t 时刻网络中其他节点之间不存在的连边构成不存在边集合.$\Gamma(t)$表示 t 时刻网络节点集合,N_t 表示$\Gamma(t)$中的节点数目.

用公式(13)和公式(14)表示如下:

$$mAUC^t=\left(\frac{\sum_{i\in\Gamma(T)}AUC_i^t}{N_t}\right)\times AUC^t \tag{13}$$

$$M^t=\begin{cases}mAUC^t-\dfrac{\sum_{T=t-\Delta t}^{t-1}mAUC^{T-1}}{\Delta t}, & 0<\Delta t<t,\Delta t\text{为整数}\\ mAUC^t, & \Delta t=0\end{cases} \tag{14}$$

算法 3. 基于 mAUC 的 FDA.

输入:事件检测阈值 MD,网络检测稳定点 TO,放大系数 A,记忆时间Δt,网络快照集合$\{g^1,g^2,\ldots,g^t,\ldots,g^n\}$.

1. 基于 TO 时的网络快照,通过算法 OWA 得到 TO 时最佳混合指标 $BMixSimIndex_{TO}$.
2. For t=1 to n

   执行公式(16)和公式(17),得到 M^t.

   执行公式(15),得到 F^t.

   End
3. 输出事件检测序列$(F^1,F^2,\ldots,F^t,\ldots,F^T)$,较小 F 值对应的时刻即为潜在事件发生点.

为了更好地解释算法 FDA,我们继续用第 2.2.1 节中 $N2$ 的例子进行说明.设置 MD=–0.1,TO=120,A=10,Δt=1,分别基于 AUC 和 mAUC 的 FDA,得到的事件检测序列如图 4 所示.

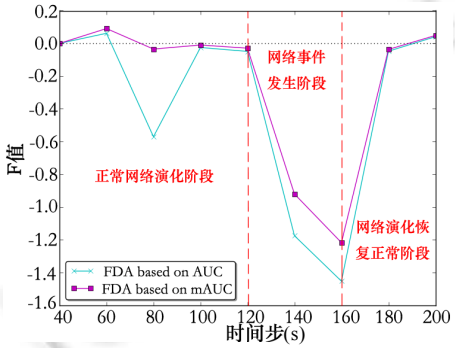


Fig.4 Network $N2$ event detection sequence change map

图 4 网络 $N2$ 上的事件检测序列变化图

从图 4 可以得出如下结论:

(1) 基于 AUC 和 mAUC 的 FDA 都可以很好地检测出网络事件.在网络事件发生阶段(第 121 时间步~第 159 时间步),网络事件检测值都大幅度降低.正常网络演化阶段(第 40 时间步~第 120 时间步)和网络演化恢复正常阶段(第 160 时间步~第 200 个时间步),网络事件检测值一直相对较高.同时,经过 FDA 的处理,网络事件检测值比衡量指标本身(如图 3 所示)具有更显著的事件检测效果.

(2) 基于 mAUC 的 FDA 通过考虑节点的微观演化对宏观衡量指标 AUC 值进行纠正,能够更真实地反映网络演化的波动,有效避免非事件造成的 AUC 值变动.从第 40 时间步~第 80 时间步,网络 $N2$ 正在构建中,还没有达到稳定状态,但是微观上,每个节点是符合网络的内在演化机制,所以基于 mAUC 的 FDA 仍能保持相对较高的网络事件检测值,很好地避免了非事件引起的正常网络波动的干扰.

3 实验分析

本节通过实验来验证 IndexEvent 方法中的关键理论.第 3.1 节介绍了实验中使用的真实的社会网络数据集.第 3.2 节利用 OWA 确定各真实数据集上的最佳混合指标,并与表 1 中提到的常见相似性指标对比来表明混合指标的有效性.基于第 3.2 节得出的最佳混合指标,第 3.3 节通过 FDA 量化各数据集上的不同时段的网络演化波动,进行事件检测.并通过与真实的事件进行比较,证明了 IndexEvent 方法的准确性.

3.1 数据集描述

为了验证本文提出的 IndexEvent 方法的有效性,通信网络(VAST)和邮件网络(Enron)被用作实验数据集.VAST 数据集源自 IEEE VAST 2008[44],包含 400 人在 10 天内的通话数据网络,并且已知在第 7 天与第 8 天之间发生了一次高层变动,引起了网络波动.Enron 数据集源自 Enron 公司的内部邮件联系网络[42],包含 150 人在 111

周内的通信数据,本节选择代表性的连续的 40 周,期间包括公司被收购、破产等多个事件.VAST 数据集的特点是发生的事件单一,事件发生的时间和原因确定,有助于检测特定事件,并分析事件前后网络的变化情况. Enron 数据集的特点是多个事件相继发生,有利于我们分析多个事件对网络的连续影响,分析不同事件所起的作用.

3.2 最佳混合指标确定

通过在真实网络 VAST 和 Enron 的仿真实验,本节验证混合指标的在真实网络应用中的可行性和算法 OWA 确定最佳混合指标的高效性.

- 对于 VAST 真实通话网络

(1) 以所选 10 天里的第 1 天作为基准时刻 TO,TO 时刻基于表 1 的 8 种不同相似性指标的链路预测算法的 AUC 值见表 7.

Table 7 AUC value of the link prediction algorithm with different similarity indexes for VAST dataset at TO time

表 7 TO 时刻,VAST 数据集上,不同相似性指标链路预测算法的 AUC 值

Index	CN	**PA**	**AA**	SA	JA	SO	HPI	LNH
AUC	0.504 73	**0.514 93**	**0.504 78**	0.503 24	0.504 21	0.504 26	0.504 43	0.504 24

(2) 基于表 7 所示结果,选取 AUC 值靠前的 PA 和 AA 作为混合指标的单位指标,利用 OWA 算法确定 PA 和 AA 对应的权重,可得其最佳混合指标 $BMixSimIndex_1$=0.273PA+0.727AA.在 TO 时刻,其对应的 AUC 值为 0.516 13,略高于基于它的单位指标 PA 和 AA 各自对应的链路预测精度;同时远高于表 1 的 8 种其他相似性指标的链路预测精度.

(3) 由于 SA,JA,SO,HPI,LNH,CN 和 AA 有相似的曲线,所以为了让图 5 更清晰,只有 PA,AA 指标被选取展示.$BMixSimIndex_1$ 和 PA 可以比其他相似性指标更好地反映网络在第 7 天与第 8 天之间发生的高层变动事件.

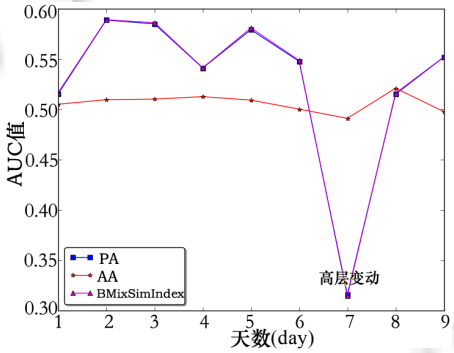


Fig.5 Change of AUC values of $BMixSimIndex_1$, PA, AA with the number of days in VAST network

图 5 在 VAST 网络中,$BMixSimIndex_1$,PA,AA 的 AUC 值随天数的变化趋势

- 对于 Enron 真实通信网络

(1) 以所选 40 周里的第 2 周作为基准时刻 TO,不同相似性指标的链路预测算法的 AUC^2 值见表 8.

Table 8 AUC^2 value of the link prediction algorithm with different similarity indexes for Enron dataset at TO time

表 8 TO 时刻,Enron 数据集上,不同相似性指标链路预测算法的 AUC^2 值

Index	**CN**	**PA**	AA	SA	JA	SO	HPI	LNH
AUC	**0.577 96**	**0.625 66**	0.576 00	0.575 09	0.575 05	0.362 10	0.074 49	0.575 11

(2) 基于表 8 所示结果,选取 AUC 值靠前的 PA 和 CN 作为单位指标,利用算法 OWA 确定 PA 和 CN 对应的权重,可得其最佳混合指标为 $BMixSimIndex_2$=0.014PA+0.986CN.在 TO 时刻,其对应的 AUC 值为 0.638 75 远高于 8 种独立相似性指标的链路预测算法精度.

综述,通过分别在 VAST 和 Enron 上进行混合指标的讨论,证明了混合指标在真实网络中的可行性和 OWA 的有效性.

(1) 当最佳混合指标中出现某个单位指标权重为 1、其他指标权重均为 0 的特例情况时,此时该网络最佳混合指标和这个权重为 1 的独立相似性指标的链路预测度精度一样.其他情况下的网络均可通过 OWA 找到一个链路预测精度优于独立相似性指标的最佳混合指标.

(2) 真实网络的演化机制越复杂,通过 OWA 得到最佳混合指标的优势越明显.因为独立指标很难反映多种演化机制混合的真实网络,而混合指标则能更全面地反映.在 Enron 真实网络中,多个事件连续发生,网络的内在机制发生了重大变化,该网络对应的最佳混合指标精度比表现最佳的独立指标 PA 高 0.13.而在单个事件发生的 VAST 中,网络的演化机制变化较小,链路预测的精度仅仅提高了 0.001 2.

3.3 事件检测分析

通过在真实网络 VAST 和 Enron 上实验,本节验证了链路预测衡量指标值变动与网络演化波动的一致性关系,并验证了 FDA 进行事件检测的准确性.

对 VAST 通话数据集,设置 MD=–0.02,TO=1,A=10,Δt=1,VAST 上的事件检测序列如图 6 所示,可得如下结论:

(1) VAST 通话网络在第 7 天与第 8 天之间发生了一次高层变动,基于 AUC 和基于 mAUC 的 FDA 都能显著地检测到这次事件的发生,第 6 天到第 7 天的事件检测值大幅度下降.当到了第 8 天时,事件检测值恢复到正常水平,高层变动产生的影响慢慢减弱.

(2) 基于 AUC 的 FDA 对网络演化波动更加敏感,在高层变动事件发生第 7 天,事件检测值大幅度降低,在第 2 天和第 3 天也出现了一定的波动.基于 mAUC 的 FDA 有效地避免了基于 AUC 的 FDA 在第 2 天和第 3 天产生的非事件引起的网络波动干扰,能够精准地检测出第 7 天事件发生引起的网络波动.

对于 Enron 通信数据集,设置 MD=0,TO=1,A=1,Δt=0,Enron 上的事件检测序列如图 7 所示,表 9 列出了 Enron 网络检测出的事件与真实事件对应关系.可得如下结论:

(1) Enron 网络中多个事件连续发生,基于 AUC 和基于 mAUC 的 FDA 都很好地检测出多个网络事件.在网络事件发生时,事件检测值都会降低.

(2) 基于 AUC 的 FDA 对网络演化波动更加敏感,连续事件发生时,事件检测值变动较大,并且在没有发生事件是也会有一定波动.基于 mAUC 的 FDA 只在事件发生时才出现事件检测值降低,避免了正常波动带来的干扰.同时,可判断事件 2(美国证券交易委员会要求提交交易内容)严重扰乱了网络内在演化机制,引起了较大的网络演化波动,导致事件检测值迅速下降.

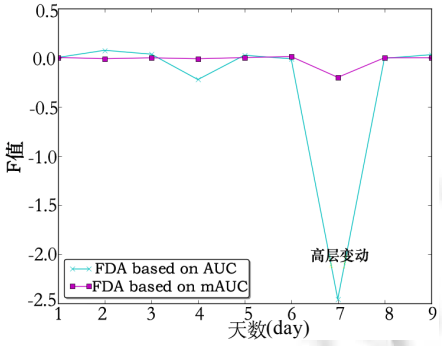


Fig.6 Sequence of event detection on VAST

图 6 VAST 上的事件检测序列图

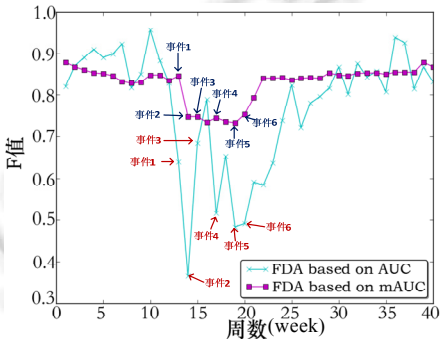


Fig.7 Event detection sequence on Enron

图 7 Enron 上的事件检测序列

Table 9 Corresponding relationship between the event and the real event detected by the Enron network

表 9 Enron 网络检测出的事件与真实事件对应关系

事件编号	发生时间	具体事件描述
事件 1	2001 年 10 月 16 日	安然发布 2001 年第二季财务报表
事件 2	2001 年 10 月 22 日	美国证券交易委员会要求提交交易内容
事件 3	2001 年 10 月 31 日	美国证券交易委员会开始对安然进行正式调查,次日安然抵押了部分资产
事件 4	2001 年 11 月 8 日	安然承认做了假账,次日迪诺基公司宣布准备收购安然
事件 5	2001 年 11 月 28 日	标准普尔将安然调低至“垃圾债券”,30 日,安然股价跌至 0.26 美元
事件 6	2001 年 12 年 2 日	安然申请破产保护

综上,通过对 VAST 和 Enron 真实网络上事件检测分析,证明了基于 AUC 和基于 mAUC 的 FDA 在单个事件发生的网络和多个事件发生的网络上都能进行有效的事件检测,并且可以对不同事件对网络的影响进行定量评估.基于 AUC 的 FDA 对网络波动更加敏感,容易受到非事件引起的网络波动的干扰.基于 mAUC 的 FDA 由于考虑了节点的微观演化,能够更精准地检测出事件引起的波动,避免无效波动的干扰.

4 结论及展望

为了检测网络事件,量化事件对社会网络演化产生的影响,本文提出了一种混合指标群智能方法 IndexEvent:利用最佳权重算法 OWA 来确定当前时段网络的最佳混合指标;然后,通过基于 AUC 或基于 mAUC 的网络波动检测算法 FDA 检测事件.为了验证 IndexEvent 方法的有效性,本文基于 WS 小世界网、BA 无标度网络、VAST 和 Enron 真实网络进行了大量实验探讨,并得出以下结论:

(1) 对于特定真实网络,在最佳混合指标为某个单位指标权重为 1、其他指标权重均为 0 的特殊情况下,最佳混合指标和这个权重为 1 的独立相似性指标的链路预测度精度一样.其他情况均可以通过 OWA 找到一个链路预测精度优于独立相似性指标的混合指标;并且,真实网络的演化机制越复杂,通过 OWA 得到最佳混合指标的优势越明显.

(2) 基于 AUC 的 FDA 对网络演化波动更加敏感,非事件引起的正常网络波动容易对事件检测结果造成干扰.基于 mAUC 的 FDA 由于考虑了节点的微观演化,能够更好地避免正常网络波动的干扰,精准地检测出事件引起的网络异常波动.

(3) 无论是在 WS 小世界网络和 BA 无标度网络的实例,还是在 VAST 和 Enron 的真实网络,IndexEvent 方法检测出的事件与真实事件有很好的匹配关系,证明了 IndexEvent 的准确性高,具有很好的实用性.

进一步的研究仍然需要在以下两个方面继续:

(1) 尝试利用链路预测去确定真实网络中不同的网络演化机制所占的具体比重;

(2) 进一步探讨各种相似性指标及最大似然估计指标之间的关系.

致谢 我们真诚地向对本文的工作给予支持和建议的审稿人、主编、编辑、同行、老师和同学表示感谢.

References:

[1] Reuter T, Bielefeld U, Papadopoulos S, Petkos G, Vries CD, Mezaris V. Social event detection at mediaeval 2013: Challenges, datasets, and evaluation. In: Proc. of the MediaEval Workshop. 2013. 18–19.

[2] Hidalgo CA, Rodriguez-Sickert C. The dynamics of a mobile phone network. Physica A Statistical Mechanics & Its Applications, 2008,387(12):3017–3024. [doi: 10.1016/j.physa.2008.01.073]

[3] Watts DJ, Strogatz SH. Collective dynamics of “small-world” networks. Nature, 1998,393(6684):440–442. [doi: 10.1038/30918]

[4] Zhang Z, Wu B. Pfaffian orientations and perfect matchings of scale-free networks. Theoretical Computer Science, 2015,570(C): 55–69. [doi: 10.1016/j.tcs.2014.12.024]

[5] Washio T, Motoda H. State of the art of graph-based data mining. AcmSigkdd Explorations Newsletter Homepage, 2003,15(1): 59–68. [doi: 10.1145/959242.959249]

[6] Papadimitriou P, Dasdan A, Garcia-Molina H. Web graph similarity for anomaly detection. Journal of Internet Services & Applications, 2010,1(1):19–30. [doi: 10.1007/s13174-010-0003-x]

[7] Mcculloh IA, Carley KM, Mcculloh IA, Carley KM. Social network change detection. Technical Report, CMU-ISR-08-116, Institute for Software Research, School of Computer Science, Carnegie Mellon University, 2008. [doi: 10.2139/ssrn.2726799]
[8] Wan X, Milios E, Kalyaniwalla N, Janssen J. Link-Based event detection in email communication networks. In: Proc. of the SAC ACM Symp. on Applied Computing. 2009. [doi: 10.1145/1529282.1529618]
[9] Taskar B, Abbeel P, Koller D. Discriminative probabilistic models for relational data. Eprint Arxiv, 2012. 485−492.
[10] Aaron C, Cristopher M, Newman MEJ. Hierarchical structure and the prediction of missing links in networks. Nature, 2008,453(7191):98−101. [doi: 10.1038/nature06830]
[11] Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks. Journal of the American Society for Information Science and Technology, 2007,58(7):1019−1031. [doi: 10.1002/asi.20591]
[12] Kossinets G, Watts DJ. Reports empirical analysis of an evolving social network. Science, 2006,311(5757):88−90. [doi: 10.1126/science.1116869]
[13] Sarukkai RR. Link prediction and path analysis using Markov chains. Computer Networks, 2000,33(1):377−386. [doi: 10.1016/S1389-1286(00)00044-X]
[14] Zhu J, Hong J, Hughes JG. Using Markov chains for link prediction in adaptive Web sites. In: Proc. of the ACMSigWeb Hypertext. 2002. 60−73.
[15] Clauset A, Moore C, Newman MEJ. Hierarchical structure and the prediction of missing links in networks. Nature, 2008,453(7191):98−101. [doi: 10.1038/nature06830]
[16] Guimera R, Sales-Pardo M. Missing and spurious interactions and the reconstruction of complex networks. Proc. of the National Academy of Sciences, 2010,106(52):22073−22078. [doi: 10.1073/pnas.0908366106]
[17] Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks. Journal of the American Society for Information Science & Technology, 2007,58(7):1019−1031. [doi: 10.1002/asi.20591]
[18] Rapoport A. Spread of information through a population with socio-structural bias: I. Assumption of transitivity. The Bulletin of Mathematical Biophysics, 1953,15(4):523−533. [doi: 10.1007/BF02476440]
[19] Jaccard P. Etude Comparative de la distribution florale dans une portion des Alpes et du Jura. Bulletin de La Societe Vaudoise Des Sciences Naturelles, 1990,37(142):547−579.
[20] Adamic LA, Adar E. Friends and neighbors on the Web. Social Networks, 2003,25(3):211−230. [doi: 10.1016/S0378-8733(03)00009-1]
[21] Barabasi AAR. Emergence of scaling in random networks. Science, 1999,286(5439):509−512. [doi: 10.1126/science.286.5439.509]
[22] Lichtenwalter RN, Lussier JT, Chawla NV. New perspectives and methods in link prediction. In: Proc. of the ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Washington, 2010. 243−252. [doi: 10.1145/1835804.1835837]
[23] Symeonidis P, Iakovidou N, Mantas N, Manolopoulos Y. From biological to social networks: Link prediction based on multi-way spectral clustering. Data and Knowledge Engineering, 2013,87(4):226-242. [doi: 10.1016/j.datak.2013.05.008]
[24] Huang Z, Li X, Chen H. Link prediction approach to collaborative filtering. In: Proc. of the Joint Conf. on Digital Libraries. ACM Press, 2005. 7−11. [doi: 10.1145/1065385.1065415]
[25] Rao J, Wu B, Dong YX. Parallel link prediction in complex network using MapReduce. Ruan Jian Xue Bao/Journal of Software, 2012,23(12):3175−3186 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4206.htm [doi: 10.3724/SP.J.1001.2012.04206]
[26] Salton G, McGill MH. Introduction to Modern Information Retrieval. New York: McGraw-H Hill, 1983.
[27] Sørensen T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. Biol Skr, 1948,5:1−34.
[28] Ravasz E, Somera AL, Mongru DA, Oltvai ZN，Barabási AL. Hierarchical organization of modularity in metabolic networks. Science, 2002,297(5586):1551−1555. [doi: 10.1126/science.1073374]
[29] Leicht EA, Holme P, Newman MEJ. Vertex similarity in networks. Physical Review E, 2006,73(2):026120. [doi: 10.1103/PhysRevE.73.026120]
[30] Noble CC, Cook DJ. Graph-Based Anomaly Detection. In: Proc. of the ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Washington, 2003. 631−636. [doi: 10.1145/956750.956831]
[31] Priebe CE, Conroy JM, Marchette DJ. Scan statistics on enron graphs. Computational & Mathematical Organization Theory, 2005, 11(3):229−247. [doi: 10.1007/s10588-005-5378-z]
[32] Wu B, Wang B, Yang SQ. Framework for tracking the event-based evolution in social networks. Ruan Jian Xue Bao/Journal of Software, 2011,22(7):1488−1502 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3841.htm [doi: 10.3724/SP.J.1001.2011.03841]
[33] Baruah RD, Angelov P. Evolving social network analysis: A case study on mobile phone data. In: Proc. of the 2012 IEEE Conf. on Evolving and Adaptive Intelligent Systems (EAIS). IEEE, 2012. 114−120. [doi: 10.1109/EAIS.2012.6232815]

[34] Qiao SJ, Tang CJ, Peng J, Liu W, Wen FL, Qiu JT. Mining key members of crime networks based on personality trait simulation email analysis system. Chinese Journal of Computer, 2008,31(10):1795−1803 (in Chinese with English abstract). [doi: 10.3321/j.issn:0254-4164.2008.10. 014]

[35] Kashima H, Abe N. A parameterized probabilistic model of network evolution for supervised link prediction. Trans. of the Japanese Society for Artificial Intelligence, 2006,22(2):340−349. [doi: 10.1109/ICDM.2006.8]

[36] Hanley JA, Mcneil BJ. The meaning and use of the area under a receiver operating chracteristic (roc) curve. Radiology, 1982, 143(1):29−36. [doi: 10.1148/radiology.143.1.7063747]

[37] Herlocker JL, Konstan JA, Terveen LG, Riedl JT. Evaluating collaborative filtering recommender systems. ACM Trans. on Information Systems, 2004,22(1):5−53. [doi: 10.1145/963770.963772]

[38] Zhou T, Ren J, Medo M, Zhang YC. Bipartite network projection and personal recommendation. Physical Review E (Statistical Nonlinear and Soft Matter Physics), 2007,76(4):70−80.

[39] Bouwmeester D, Ekert A, Zeilinger A. The physics of quantum information. In: Proc. of the Quantum Cryptography Quantum Teleportation Quantum Computation. 2010.

[40] Tang D, Cai Y, Zhao J, Xue Y. A quantum-behaved particle swarm optimization with memetic algorithm and memory for continuous non-linear large scale problems. Information Sciences, 2014,289(24):162−189. [doi: 10.1016/j.ins.2014.08.030]

[41] Schutte JF, Groenwold A. A study of global optimization using particle swarms. Journal of Global Optimization, 2004,31(31):93−108. [doi: 10.1007/s10898-003-6454-x]

[42] http://www.cs.cmu.edu/~enron/2012/8/1

[43] Liu HK, Lü LY, Zhou T. Uncovering the network evolution mechanism by link prediction. Scientia Sinica PhysMechAstron, 2011, 41(7):816−823 (in Chinese with English abstract). [doi: 10.1360/132010-922]

[44] http://www.cs.umd.edu/hcil/VASTchallenge08/2012/10/8

附中文参考文献:

[25] 饶君,吴斌,东昱晓.MapReduce 环境下的并行复杂网络链路预测.软件学报,2012,23(12):3175−3186. http://www.jos.org.cn/1000-9825/4206.htm [doi: 10.3724/SP.J.1001.2012.04206]

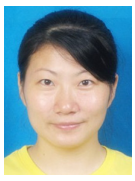
[32] 吴斌,王柏,杨胜琦.基于事件的社会网络演化分析框架.软件学报,2011,22(7):1488−1502. http://www.jos.org.cn/1000-9825/3841.htm [doi: 10.3724/SP.J.1001.2011.03841]

[34] 乔少杰,唐常杰,彭京,等.基于个性特征仿真邮件分析系统挖掘犯罪网络核心.计算机学报,2008,31(10):1795−1803. [doi: 10.3321/j.issn:0254-4164.2008.10.014]

[43] 刘宏鲲,吕琳媛,周涛.利用链路预测推断网络演化机制.中国科学:物理学力学天文学,2011,41(7):816−823. [doi: 10.1360/132010-922]

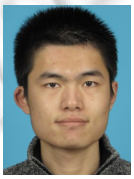
胡文斌(1976－),男,湖北武汉人,博士,副教授,博士生导师,主要研究领域为复杂网络,人工智能,调度优化.

王欢(1989－),男,博士生,主要研究领域为社会网络分析,数据挖掘.



严丽平(1980－),女,博士生,主要研究领域为社会网络分析,数据挖掘

邱振宇(1992－),男,硕士生,主要研究领域为复杂网络,社会网络分析.



肖雷(1992－),男,硕士生,主要研究领域为社会网络分析,数据挖掘.

杜博(1983－),男,博士,教授,博士生导师,CCF 专业会员,主要研究领域为复杂网络,社会网络分析.