软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2011,22(7):1488−1502 [doi: 10.3724/SP.J.1001.2011.03841]

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

基于事件的社会网络演化分析框架*

吴 斌, 王 柏, 杨胜琦+

(北京邮电大学 北京市智能通信软件与多媒体重点实验室,北京 100876)

Framework for Tracking the Event-Based Evolution in Social Networks

WU Bin, WANG Bai, YANG Sheng-Qi+

(Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing 100876, China)

+ Corresponding author: E-mail: sheng_qi.yang@yahoo.com.cn

Wu B, Wang B, Yang SQ. Framework for tracking the event-based evolution in social networks. *Journal of Software*, 2011,22(7):1488−1502. http://www.jos.org.cn/1000-9825/21/3841.htm

Abstract: This paper presents a fundamentally different framework for uncovering the intricate properties of evolutionary networks. Contrary to static snapshots methods, this paper first traces the timelines of the networks. Then, based on extracted smooth segments from the timelines, a graph approximation algorithm is applied to capture the frequent characteristics of the network and reduce the noise of interactions. Moreover, by employing the relationship among multi-attributes, an innovative community detection algorithm is proposed for a detailed analysis on the approximate graphs. To track these dynamic communities, this paper also introduces a community correlation and evaluation method. Finally, by applying this novel framework to several real-world networks, this paper demonstrates the critical relationship between event and social evolution, and reveals meaningful properties in actual dynamic behaviors.

Key words: social network; evolution; dynamic pattern; community detection

摘 要: 提出了一个全新的复杂网络分析框架来跟踪动态网络的演化规律,发现其在演化过程中的时间特性.不同于传统静态时间片的分析方法,整个框架首先利用有效而快速的方法发现网络的 timeline,然后利用图近似算法刻画 timeline 中的平稳演化段落,这样可以有效地降低个体行为的不确定性所带来的网络演化噪声.此外,综合考虑到网络中个体的多维属性,还提出一种高效的社团发现算法,用以发现动态网络中的社团结构.为了对社团进行演化分析,提出了社团演化的评价方法,以发现社团演化过程的动态特征.最后,为了示例该框架的有效性和实用性,整个框架被应用于多个实际的网络数据集,并且揭示了这些网络在演化过程中的时间特性及社团演化模式.

关键词: 社会网络;演化;动态模式;社团发现

中图法分类号: TP311 **文献标识码**: A

基于网络的复杂理论研究近几年引起了人们浓厚的兴趣,由此诞生的复杂网络研究方法被广泛应用于科技、经济和社会生活等领域,用以发现和刻画不同个体之间的交互关系(如呼叫关系、科研领域的合作关系等)

* 基金项目: 国家自然科学基金(90924029, 60905025); 国家科技支撑计划(2006BAH03B05)
收稿时间: 2009-02-19; 修改时间: 2009-07-21, 2009-11-04; 定稿时间: 2010-03-11

以及由个体所组成的社团结构特征(如朋友圈、生物圈等).在现实的网络当中,有些网络是相对静态的,如蛋白质网络、Internet 的物理层网络.而对大多数网络来说,它们的拓扑结构会随着时间而发生明显的变化,如每个人所处的朋友圈在不同时期会有所不同.传统方法[1−3]主要从整个网络或某个特定时间快照出发,关注网络的全局特征,因此往往忽略了在特殊时间点个体或社团的突发事件(emergencies)[4].

相比于其静态特征,目前,社会网络中的动态特性受到了广泛关注.许多基于网络演化的分析方法被应用于各种社会网络当中[1,2].在这些方法中,图被用来描述某一特定时间的网络快照.这样,基于这些图序列,网络的动态特性就能够被刻画出来[5−7].但是,这些方法忽略了社会网络中个体行为的随机性和突发性,而这些特性往往具有特殊的表现形式,并且会对最终分析结果产生一定的影响.简言之,基于这种“硬切分”的动态网络分析方法忽略了演化网络中两个重要的特性:噪声和事件.其中,噪声由具有社会化特征的个体行为的随机性和不确定性造成(如拨错电话号码而造成的无效通话).这种个体行为往往持续时间短,并且不具有网络扩散特性.噪声如果得不到有效的处理,有可能在演化分析过程中被放大,从而影响整个分析结果.而事件则是由个体或群体的异常行为所引起,并且具有一定的持续时间,往往也会具有扩散特性,从而造成局部或整个网络的异常性变化.本文所关注的是如何发现并利用网络演化过程中的事件,并对其产生的影响进行分析.

综合考虑这两种因素,本文提出了一种全新的基于事件发现的网络演化分析框架.从事件对网络演化所产生的影响来看,基于事件发生前后两个平稳时间段的对比研究不仅能够刻画网络在这两个时间段的特征,而且还能反映该事件对网络演化所产生的影响.综上,本文的主要贡献可总结如下:

- Timeline 可以有效地描述网络演化趋势.虽然以前的研究工作[1,8]已经提出过网络演化的 timeline 概念,但这些方法所得到的 timeline 只是其整个分析框架中的一个“副产品”,并且不具有实用性(受到算法复杂度限制).而本文提出的 timeline 发现算法同时注重了效率和效果,因此更加实用;
- 不同于以往基于单个稀疏图的图近似算法[9],文本提出的图近似算法基于图序列.算法的目的是用一个图来近似概括多个图的基本拓扑结构,这样就把对不同时间段的网络的分析简化成对这些时间段的近似图的分析.这样,不仅简化了分析工作,而且能够收到减小噪声的效果;
- 虽然目前有许多社团发现算法在社会网络分析中都取得了不错的实验效果,但大都不适用于动态网络.作为整个框架的一部分,本文提出了一种全新的社团发现算法.这种算法基于先前得到的加权近似图,能够快速而有效地获得动态网络中的社团结构;
- 动态网络的多角度分析最近也得到了广泛的关注[5,10].为了研究社团结构的演化模式,本文提出了一种社团关联及评价方法.其所得结果揭示了社团大小、结构和生命周期间的紧密联系.

整个框架的分析流程如图 1 所示.

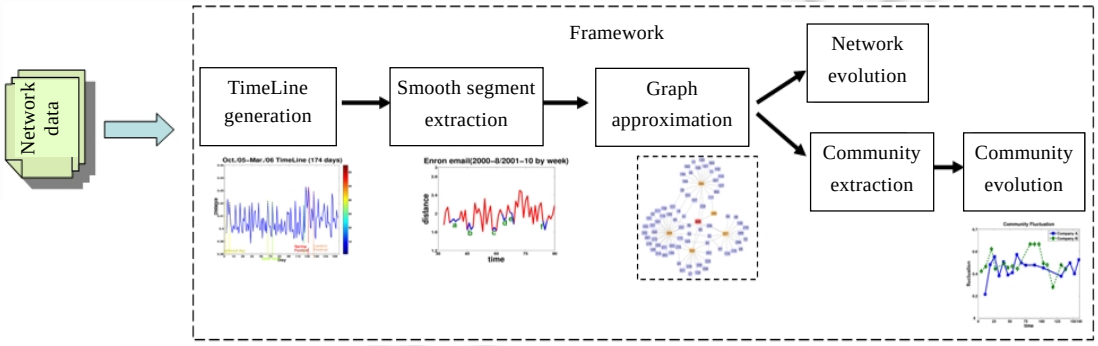


Fig.1 Event-Based social network evolution analysis framework

图 1 基于事件的社会网络演化分析框架

本文第 1 节介绍相关工作.第 2 节介绍文中用到的符号、定义以及相关理论.第 3 节详细介绍框架中的算法,并进行简要分析.第 4 节为实验及结果分析.第 5 节对本文进行总结,并提出下一步工作重点.

1 相关工作

网络演化分析作为一个新兴的领域,近年来受到了广泛的关注.从研究角度来讲,Jin 等人[11]针对增长型的网络结构,尤其是 WWW,提出了两种有效的描述模型.Leskovec 等人[7]则从网络的基本特征入手(如边点比、平均最短路径等),揭示了网络在演化过程中呈现出的与静态网络所不同的特性.Backstrom 等人[10]通过研究动态网络中社团的演化特征,发现个体加入社团与社团结构有着密切的关联.Tantipathananandh 等人[12]提出了一种基于染色方法的算法框架来进行社团演化分析.虽然这些研究者对网络演化较早地进行了关注,并做了大量开创性的工作,但其方法都存在着一些不足,如认为网络的演化只是一个不断增长的过程,而忽略了对网络随机性和突发性的考虑等.最近,一些研究者开始注意到传统方法的不足,其中与本文关联密切的研究有:Tong 等人[13]提出了 Colibri 方法集来处理静态和动态的网络分析;Lin 等人[8]提出了 FacetNet 框架,整合了社团发现和演化的分析方法.为了避免演化带来的噪声,该框架中的社团发现算法不仅依赖于当前的网络结构,还考虑了过去的网络特征;Asur 等人[2]认识到网络演化中事件发生的必然性,提出了分析个体和社团演化行为特征的框架.在该框架中,关键事件被用来预测社团的发展趋势;在 Sun 等人[1]提出的 GraphScope 框架中,二分图被用来描述整个网络,并用来进行社团划分.

2 预备知识

2.1 符号及定义

在复杂网络研究过程中,经常借助图的形式来刻画网络的拓扑结构.图中的点代表网络中的个体(如人、网页等).而图中的边代表网络中个体之间的联系(如人与人之间的通话关系、网页之间的链接关系等).

表 1 列出了本文中用到的基本符号.对于一个由 n 个时间片组成的演化网络 G,通常描述成如下的形式:

$$G=\{g^{(1)},g^{(2)},\ldots,g^{(n)}\} \tag{1}$$

Table 1 Symbols

表 1 符号表

Symbol	Definition
G	Whole graph
$g^{(t)}$	Network snapshots taken at time t;
$S^{(t)}$	Graph series at segment t
$T^{(t)}$	Graph approximation of $S^{(t)}$
$C_j^{(t)}$	Community i in $T^{(t)}$
$V(g^{(t)})$	Vertices set of $g^{(t)}$
$E(g^{(t)})$	Edges set of $g^{(t)}$
$adj^{(t)}(v)$	Neighbors of v at time t: $\{u\|(v,u)\in E(g^{(t)})\}$
$d^{(t)}(v)$	Degree of v at time t: $\|\{u\|(v,u)\in E(g^{(t)})\}\|$
$w^{(t)}(u,v)$	Weight of edge (u,v) at time t
$\delta(g^{(t)},g^{(t+1)})$ or $\delta(t,t+1)$	Distance between $g^{(t)}$and $g^{(t+1)}$
$\tilde{d}^{(t,t+1)}(v)$	Distance of v between time t and $t+1$

定义 1(graph distance,图距离). 图距离$\delta(t,t+1)$用来衡量两个图差别的大小.定义为

$$\delta(t,t+1)=\tilde{D}(g^{(t)}\,\|\,g^{(t+1)}) \tag{2}$$

其中, $\tilde{D}(g^{(t)}\,\|\,g^{(t+1)})$ 为相对熵的变形形式,其详细定义将在第 3 节给出介绍.对于演化网络 G,通过公式(2)计算得到其任意两个连续时间片的图距离,就能得到 G 的图距离序列$\{\delta(1,2),\delta(2,3),\ldots,\delta(n-1,n)\}$.本文中,这种图距离序列被称为 timeline.

定义 2(graph segment,网络段). 网络段是指由连续 n 个网络快照构成的图序列的集合.定义为

$$S^{(t)}=\{g^{(a)},g^{(a+1)},\ldots,g^{(a+n-1)}\} \tag{3}$$

其中,$n\geqslant 2$,t 为该网络段在所有网络段中的序号.对$\forall i,j$,有 $S^{(t_1)}\cap S^{(t_2)}=\varnothing$.

根据定义 2,演化网络 G 又可以表示为如下形式:

$$G=\{S^{(1)},S^{(2)},\ldots\} \tag{4}$$

定义 3(approximate graph,近似图). 近似图 $T^{(t)}$是图段 $S^{(t)}$的单图映射.它既保留了 $S^{(t)}$中各个图的共有性质,同时也有效降低了其中的噪声.

根据定义 3 及表达式(4),为了便于网络的演化分析,网络 G 将被转化为如下形式:

$$G=\{T^{(1)},T^{(2)},\ldots\} \tag{5}$$

定义 4(community,社团). 社团定义为网络中的一种子结构.在社团内部,点之间的联系比较紧密.相比之下,这些点和社团外部的点的联系则比较稀疏.

现实中,这种社团结构可能代表了某些社会组织或团体,如某种朋友圈、同事圈、亲友圈等.对这些社团的研究,尤其是演化分析具有重要的意义.

2.2 相关理论

相对熵通常被用来衡量两个分布之间的距离.在统计学中,它对应的是似然比的对数期望.其定义如下:

定义 5(relative entropy,相对熵). 两个概率密度函数为 $p(x)$和 $q(x)$之间的相对熵(或 Kullback-Leibler)距离定义为

$$D(p\,\|\,q)=\sum_{x\in\chi}p(x)\log\frac{p(x)}{q(x)}=E_p\log\frac{p(\chi)}{q(\chi)} \tag{6}$$

定义 6(type,型). 序列 $x_1,x_2,\ldots,x_n$的型 P_χ(或经验概率分布)是 X 中每个字符在该序列中出现次数的相对比例(对任意的 $a\in X,P_\chi(a)=N(a|x)/n$,其中,$N(a|x)$表示字符 a 在序列 $x\in X^n$ 中出现的次数).

本文引入离散随机变量 $X(v),v\in V(G)$来描述网络中的个体.简单起见,$X(v)$被认为是独立同分布的(*i.i.d*).网络中,每个时间快照 $g^{(t)}$均有 $X(v)$在其上的型映射 $X^{(t)}(v)$(需要说明的是,并不是对$\forall x,t$,均有 $v\in g^{(t)}$).通过计算两个时间快照 t 和 t+1 的型映射的相对熵 $D(\chi^{(t)}\|\chi^{(t+1)})$,进而可以获得两个图的图距离$\delta(t,t+1)$.本文中对公式(6)进行了必要的改进.

3 框架详述

3.1 Timeline生成

由于本文提出的框架基于 timeline,所以需要设计一种合理而高效的生成算法,以便用 timeline 来描述网络演化的趋势,发现演化过程中的重要事件.

Timeline 生成算法使用了一个大小为 ws 的滑动窗口来确定变化点(change spot)两侧网络的变化.图 2 示例了一个大小为 8 的滑动窗口,其中,较深色点代表变化点后不再出现的节点(离网节点),空心点代表变化点后新入网的节点(入网节点),最深色点代表变化点前后均出现过的节点(稳定节点),最浅色点代表只在某一时刻出现过的节点(噪声节点).如前所述,从 t 时刻到 t+1 时刻,网络的变化由$\delta(t,t+1)$衡量,其值是该时刻离网节点、入网节点和稳定节点的变化的累计(对于噪声节点,由于其所占比例较小,并且对整个网络的演化产生的影响也较小,所以这里忽略了其在变化点处对网络产生的影响).对单个节点来讲,其在该时刻的变化定义为 $\tilde{d}_{t,t+1}(v)$:

$$\tilde{d}^{(t,t+1)}(v)=\begin{cases}\left|\log\dfrac{d^{(t)}(v)+1}{1}\right|, & v\in V(\text{离网节点})\\ \left|\log\dfrac{1}{d^{(t+1)}(v)+1}\right|, & v\in V(\text{入网节点})\\ \left|\log\dfrac{d^{(t)}(v)}{d^{(t+1)}(v)}\right|+\left|\log\dfrac{adj^{(t)}(v)\cap adj^{(t+1)}(v)}{adj^{(t)}(v)\cup adj^{(t+1)}(v)}\right|, & v\in V(\text{稳定节点})\end{cases} \tag{7}$$

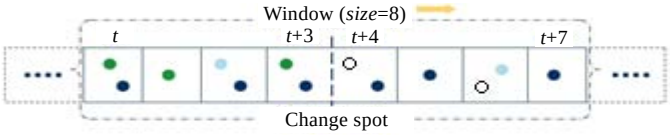


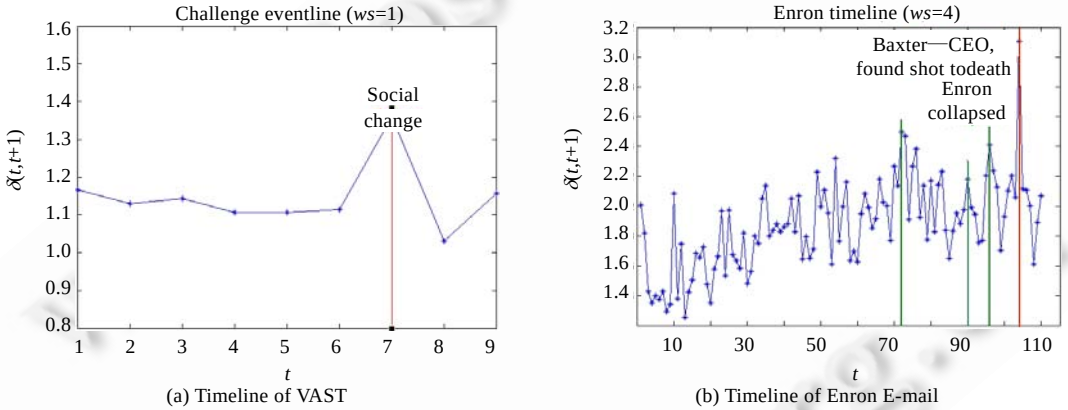
Fig.2 Moving window with *size*=8

图 2 一个大小为 8 的移动窗口

从公式(7)可以看出,对于入网节点和离网节点,只需考虑其度的变化情况.而对于稳定节点,既考虑了其度(表示其活跃度)的变化,又考虑了其邻居(环境)的变化.综合考虑这 3 种节点,整个网络从 t 时刻到 t+1 时刻的变化 $\delta(t,t+1)$定义为

$$\delta(t,t+1)=\frac{\sum_{\forall v\in V(dead)}\tilde{d}^{(t,t+1)}(v)+\sum_{\forall v\in V(born)}\tilde{d}^{(t,t+1)}(v)+\sum_{\forall v\in V(stable)}\tilde{d}^{(t,t+1)}(v)}{|V(g^{(t)})\cup V(g^{(t+1)})|} \tag{8}$$

为了避免由于图的大小给 $\delta(t,t+1)$带来的影响,公式(8)在对 3 种节点的变化值 $\tilde{d}_{t,t+1}(v)$ 进行累加后,除以了 t 和 t+1 时刻图的节点并集大小.通过公式(8),对演化网络中每个变化点求得的距离值所组成的坐标集 $\{(t,\delta(t,t+1))\}$就形成了整个网络的 timeline.图 3(a)刻画了 VAST 数据集(说明详见第 4.1 节)的 timeline.可以看出,网络变化的最大值为 $\delta(7,8)$,即时刻 7 和时刻 8 之间网络发生了较大的事件(或变化).图 3(b)刻画了 Enron Email 数据集的网络演化 timeline,可以看出其更加复杂的演化过程.



(a) Timeline of VAST

(a) VAST 数据集的 timeline

(b) Timeline of Enron E-mail

(b) Enron E-mail 数据集的 timeline

Fig.3 Timeline examples

图 3 Timeline 示例

从效率上讲,由于每次窗口只向前移动一步,且对 3 种节点的状态是在原有窗口的基础上加以更新,所以其时间复杂度为 $O(n)$(假设$|V(g^{(t)})|$为 n).同时,公式(8)的时间复杂度为 $O(\rho n)$,其中,$\rho=2|E|/|V|$,ρ为节点的平均度.由于现实的社会网络一般呈现出稀疏特性,一般有$\rho<10$.所以,整个 timeline 发现算法的时间复杂度可以控制在 $O(mn)$,其中,m 为演化网络的有效变化点,近似为网络时间片数目(由于采用滑动窗口的策略,网络中前 $\frac{ws}{2}-1$ 个变化点和最后 $\frac{ws}{2}-1$ 个变化点并不能被有效计算,所以有效变化点为 $m-ws+2$ 个.实际中,$m>>ws$).

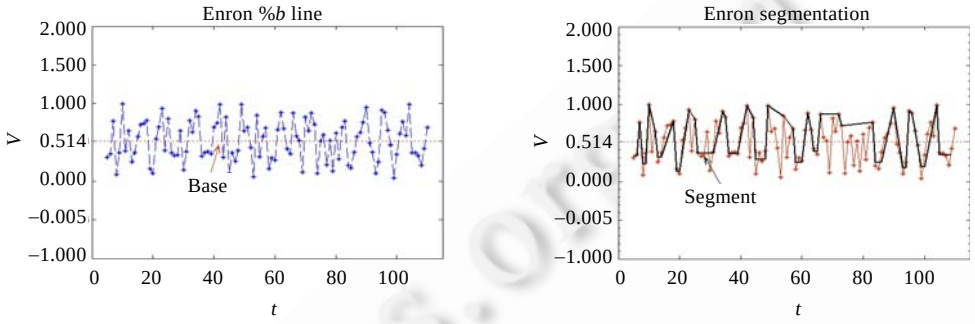
3.2 网络段抽取

在某个时间段内,网络演化呈现出一种平稳的(smooth)或相对多变的(eventful)的趋势.对这两种情况分段处理,往往能够简化被分析的演化网络.先前的网络分段方法[1]采用了增量的方式,这种方法的一个缺陷是增量

过程中的噪声会被放大,从而影响对整个演化段描述的准确性.一般认为,网络演化过程经常呈现出变化与平稳交替出现的一个过程.所以,直接分析平稳演化段不仅能够容易得到这些时间段内网络所共有的性质,而且还能通过事件前后两个平稳演化段的对比得到事件对网络演化的影响.这里,我们提出了一种两步自动分段方法:

(1) 首先使用 Bollinger Bands[18]对原始的 timeline 进行归一化处理(其值被归一到[0,1]),并产生一个线性分段的基准.通过该步骤,图 3(b)所示的 Enron 数据集的 timeline 即可转化为如图 4(a)所示的形式;

(2) 基于步骤 1 的%b 线,使用滑动窗口策略找出相对于基准的较高值和较低值.裁剪较高值并平滑所有的较低值,从而得到 timeline 的平稳演化段落.如图 4 所示,基于图 4(a)中的%b 线,经过步骤 2 的滑动窗口方法,可以自动提取出 Enron 数据集 timeline 中的平稳演化段落(如图 4(b)中黑线所示).



(a) %b line
(a) %b 线

(b) Segmentation
(b) 分段

Fig.4 Automatic segmentation using Bollinger Bands

图 4 基于 Bollinger Bands 的自动分段

上述两步分段方法的具体描述可参见文献[18].从图 4(b)可以看出,基于 Bollinger Bands 的 timeline 平稳演化段抽取方法不仅较好地刻画了网络演化规律,而且该过程不需要任何输入参数.

3.3 图近似

图近似的目的是:(a) 从网络平稳演化段 $S^{(t)}$抽象出 $T^{(t)}$以简化下一步的分析;(b) $T^{(t)}$能够在尽量保留 $S^{(t)}$中各个时间片的共性的同时,减小噪声.

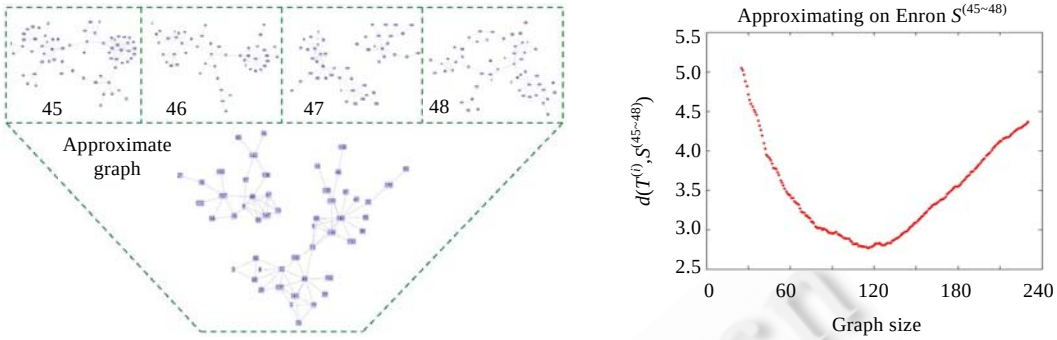
一般认为,$T^{(t)}$应该仅仅保留 $S^{(t)}$中各时间快照的共有结构.但是,这种方法并不足以刻画 $S^{(t)}$.例如,$S^{(t)}$由 5 个时间快照组成$\{g^{(a)},\ldots,g^{(a+4)}\}$,如果边$(u,v)$出现在了除 $g^{(a+2)}$以外的其他 4 个快照中,按照这种策略,(u,v)并不会被包含到 $T^{(t)}$中.但是很显然,(u,v)能够代表 $S^{(t)}$中个体 u 和个体 v 之间的关联关系,应该被包含到 $T^{(t)}$中.为了表示 $T^{(t)}$对 $S^{(t)}$的描述程度,这里引入一种网络段描述偏差的定义,即 $d(T^{(t)},\tilde{S}^{(t)})$.对于网络段 $S^{(t)}$,其近似图 $T^{(t)}$对它的描述能力定义为

$$d(T^{(t)},\tilde{S}^{(t)})=\sqrt{\sum_{g^{(a)}\in\tilde{S}^{(t)}} w_i\tilde{D}(T^{(t)}\parallel\tilde{S}^{(t)})^2} \tag{9}$$

其中,w_i 为权重.即使是对平稳演化段,其内部各个时间片之间也会有小的变化.这里定义 $S^{(t)}$中$\delta(t,t+1)$最小时 w_t=1,w_{t+1}=1,其他权重值定义为 $\dfrac{\delta(j,j+1)}{\delta(t,t+1)}$. $d(T^{(t)},\tilde{S}^{(t)})$ 越小,说明 $T^{(t)}$与 $S^{(t)}$越“相似”.

生成 $T^{(t)}$时,首先需要构造一个空图,然后向 $T^{(t)}$加入 $S^{(t)}$中的边,直到使 $d(T^{(t)},\tilde{S}^{(t)})$ 达到最小值为止.这时,$T^{(t)}$即所求 $S^{(t)}$的近似图.然而,这种最优化的结果是由加边顺序决定的,为 NP 问题.所以,这里使用了一种启发式算法.把边集 $E(S^{(t)})$按降序排列(权值为边在 $S^{(t)}$中出现的次数),并加入到 $T^{(t)}$,同时计算 $d(T^{(t)},\tilde{S}^{(t)})$ 值,直到该值达到最小.图 5 示例了图 4(b)中平稳段的近似过程.可以看出,该网络演化段由 4 个时间快照组成(如图 5(a)上部所示).应用上述算法,当 $d(T^{(t)},\tilde{S}^{(t)})$ 达到最小时,获得如图 5(a)下部所示的近似图 T.图 5(b)为算法运行过程中

$d(T^{(i)},\tilde{S}^{(i)})$ 的变化情况.



(a) Approximate graph of segment in Fig.3(c)

(a) 图 3(c)中 segment 的近似图

(b) Value of $d(T^{(b)},\tilde{S})$

(b) $d(T^{(b)},\tilde{S})$ 的变化情况

Fig.5 Process of graph approximation on $S^{(45\sim48)}$ in Fig.3(c)

图 5 对图 3(c)中 $S^{(45\sim48)}$的近似过程

在 $T^{(t)}$构造的开始阶段,随着 $S^{(t)}$中边的加入,$d(T^{(t)},\tilde{S}^{(t)})$ 的值将是一个逐渐减少的过程.当达到某种临界值后,$d(T^{(t)},\tilde{S}^{(t)})$ 的值将开始增加. $d(T^{(t)},\tilde{S}^{(t)})$ 变化的整个过程将会呈现出“V”字形(如图 5(b)所示).因此,在实际求解近似图的过程中,可以通过每次加入 m 条边到 $T^{(t)}$使 $d(T^{(t)},\tilde{S}^{(t)})$ 快速收敛到最优值附近,然后再变为每次加入 1 条边,直到达到 $d(T^{(t)},\tilde{S}^{(t)})$ 最小值的方法来实现.这种方法可以快速提高算法的效率,尤其适用于 $S^{(t)}$子图数目较多、规模较大的情况.因此,实践中,该算法的时间复杂度往往能达到 $O(\log(|E(S^{(t)})|))$.

3.4 社团发现

虽然传统的基于模块度优化的社团划分方法[3,14]能够在静态图上产生不错的结果,但其不适用于多维度的动态网络.对于演化网络,社团发现应该考虑网络结构的时间特性.作为整个框架的一部分,这一节介绍的社团发现算法将基于带有多维度边点属性的近似图,融合了 k-clique 社团发现算法.所谓 k-clique,是一个由 k 个节点构成的完全子图.如果两个 k-clique 有公共的 k–1 个节点,那么这两个 k-clique 被认为是相邻的.一个 k-clique 社团是一个所有相邻 k-clique 的集合.k-clique 社团的显著特点是重叠性(即两个社团间有公共点)和非完全覆盖(社团所涉及的点并不能覆盖整个网络).CPM(clique percolation method)[15]是一种被广泛应用的社团发现及分析算法[5],其核心是一个基于 k-clique 的合并过程.然而在实际应用中,该算法有两个主要的不足:(1) 由于其在进行 clique 合并前要建立 clique 间的关联关系(是否具有 k–1 个公共点),而这种关系的建立需要进行多次 clique 间的比较.当图中 clique 结构较多且关联较紧密时,会极大地影响该算法的效率;(2) 需要保存大量 clique 间的关系,从而造成资源的大量开销.

为了克服上述两点不足,本节提出一种新的基于 clique 合并的社团发现算法——CBCD(clique- based community detection).该算法能够把 clique 合并的时间复杂度从 CPM 的 $O(n\log^2 n)$提高到 $O(n\log n)$.此外,由于 CBCD 采用立即合并策略,所以对空间开销也比较小.具体而言,CBCD 算法可被描述为如下步骤:

(1) 对于给定近似图 $T^{(t)}$,使用文献[16]中所提算法找出所有 clique.对任意两个有公共点的 clique,如果其公共点个数达到这两个 clique 中较小的一个 clique 的 $size$–1,那么这两个 clique 就进行合并.该步骤迭代运行,直至没有 clique 合并再次发生;

(2) 步骤 1 获得的社团具有重叠特性,为了得到非重叠社团,需要把重叠的点划分给其中某个社团.例如,节点 v 所处的社团为 $\{C_i^{(t)},C_j^{(t)},...\}$,定义 v 与其中某个社团联系的紧密程度为

$$w^{(t)}(v,C_i^{(t)})=\sqrt{\sum_{\substack{\forall(v,u)\in E(C_i^{(t)})\\ u\in V(C_j^{(t)})}} w^{(t)}(v,u)} \tag{10}$$

通过计算公式(10),找出使其达到最大的社团 $C_i^{(t)}$,把点 v 划入这个社团,并把 v 从其他社团中删除;

(3) 由于基于 clique 的社团划分方法点覆盖率较低,这一步需要一个点吸收的过程,即把原先不在这个社团中的点再吸收进来.如果某一点 v(不属于任何一个社团),但与之有关联的社团有 $\{C_i^{(t)},C_j^{(t)},...\}$,通过使 $w^{(t)}(v,C_i^{(t)})$ 达到最大,把 v 加入到社团 $C_i^{(t)}$ 中;

(4) 经过步骤 3 后,如果两个相关联的社团变得更加紧密,则把它们合并起来.根据用户定义的阈值 Q_c,与

$$w^{(t)}(C_i^{(t)},C_j^{(t)})=\frac{C_i^{(t)}与C_j^{(t)}之间的边权重}{C_i^{(t)}内部的边权重}或w^{(t)}(C_i^{(t)},C_j^{(t)})=\frac{C_i^{(t)}与C_j^{(t)}之间的边权重}{C_j^{(t)}内部的边权重} \tag{11}$$

相比较.如果公式(11)中任意一个值大于 Q_c(通常取大于 0.5 的值),则 $C_i^{(t)}$ 与 $C_j^{(t)}$ 可进行合并.

算法 CBCD 形式化地描述了上述 4 个步骤.需要说明的是,该算法的第 2 行~第 5 行对任意两个 clique 进行了比较,以判断能否进行合并.当 clique 数目比较大时,这种实现方式的效率比较低.实际中,我们采用了 inverted index 方法,为每个点建立其所关联的 clique,这样就只需对有重叠点的两个 clique 进行比较合并,极大地提高了算法效率.从整个算法的复杂度来看,对于一个具有 n 个节点的无向图,Rossman[19]证明了 clique 抽取时间复杂度的下界为 $\omega(n^{k/4})$.幸运的是,对于大多数社会网络来讲,由于其所具有的稀疏特性,往往能够达到这个下界.此外,在对社会网络抽取 k-clique 时,k 值一般取 3~6 之间.所以,对于一个大小为 n 的网络,clique 的抽取一般能达到 $O(n)$的复杂度.对于抽取出的 m 个 clique,应用优化了的 CPM 算法,生成 clique 社团的时间复杂度为 $O(m\log m)$.假设算法产生的社团个数为 l,上述算法描述的步骤 2 和步骤 3 的复杂度分别为 $O(l)$和 $O(n)$,最后一步为 $O(l\log l)$.对整个算法来讲,复杂度为 $O(n+n+l+m\log m+l\log l)$.对于稀疏图,有 $n>m>l$.所以,这个算法的时间复杂度为 $O(n\log n)$.

算法. CBCD.

输入:近似图 $T^{(t)}$;

输出:社团 $\{C_1^{(t)},...,C_m^{(t)}\}$.

1. 根据文献[16]中所提出的算法找到 $T^{(t)}$中所有 clique,$\{C_1^{(t)},...,C_n^{(t)}\}$; //步骤 1
2. FOR i=1 TO n–1
3. FOR j=i+1 TO n
4. IF $C_i^{(t)}$和$C_j^{(t)}$ 之间的公共点达到 $\min(V(C_i^{(t)}),V(C_j^{(t)}))-1$
5. 把 $C_j^{(t)}$ 并入 $C_i^{(t)}$
6. FOR 节点 v 为重叠点 DO //步骤 2
7. 找出使公式(10)达到最大的社团 $C_a^{(t)}$, $C_a^{(t)}=C_a^{(t)}\cup\{v\}$,并把 v 从其他社团中删除;
8. FOR 社团外部节点 v DO //步骤 3
9. FOR { $C_i^{(t)}$ |v is adjacent to $C_i^{(t)}$ }
10. 计算 $w^{(t)}(v,C_i^{(t)})$
11. $C_b^{(t)}=C_b^{(t)}\cup\{v\}$,其中,$C_b^{(t)}$ 使得 $w^{(t)}(v,C_b^{(t)})$ 最大
12. REPEAT //步骤 4
13. FOR 社团 $C_i^{(t)}$ 和 $C_i^{(t)}$ DO
14. IF $w^{(t)}(C_i^{(t)},C_j^{(t)})\geqslant Q_c$ 或者 $w^{(t)}(C_j^{(t)},C_i^{(t)})\geqslant Q_c$ THEN $C_i^{(t)}=C_i^{(t)}\cup C_j^{(t)}$;
15. UNTIL 无合并发生;

RETURN $\{C_1^{(t)},...,C_m^{(t)}\}$

3.5 社团演化

社团演化追踪是演化分析的一个重要部分.对于不同时刻发现的社团,如何把它们关联起来(即找到某一时

刻该社团的前继和后继)?如何判断一个社团是新诞生的,还是消亡了的,分裂了还是合并了?对社会网络来讲,这些问题都有着重要的意义.传统的社团演化分析主要有两种方法:(1) 点重合度[6],即前后两个时刻的社团之间点的重合度达到某个阈值,就认为这两个社团之间存在演化关系;(2) 结构相似[5].然而,这两种方法均存在一定的缺陷.基于此,这里提出一种新的社团演化关联度定义,即前后两个时刻的两个社团 $C_i^{(t)}$ 与 $C_j^{(t+1)}$ 之间的关联度为

$$k=\max(NCor(C_i^{(t)},C_j^{(t+1)})\times ECor(C_i^{(t)},C_j^{(t+1)})) \tag{12}$$

其中, $NCor(C_i^{(t)},C_j^{(t+1)})$和$ECor(C_i^{(t)},C_j^{(t+1)})$ 分别定义为两个社团节点和边的 Jaccard 系数.同时,公式(12)不单独追求 $NCor(C_i^{(t)},C_j^{(t+1)})$或$ECor(C_i^{(t)},C_j^{(t+1)})$ 的最大化,而是使两个值均达到一个合理的匹配程度.

为了评价网络演化过程中社团的波动性,本文提出一种社团波动定义,用来衡量演化过程中前后两个时刻网络中社团的波动大小.对图 $g^{(t)}$和 $g^{(t+1)}$,分别有 m 和 n 个社团,那么这两个图之间的社团关联度定义为

$$CCor^{(t,t+1)}=\sum_{\substack{C_i^{(t)}\\1\leqslant i\leqslant m}}\left(\frac{|N(C_i^{(t)})|}{|N(g^{(t)})|}\sum_{\substack{C_j^{(t+1)}\\1\leqslant j\leqslant n}}NCor(C_i^{(t)},C_j^{(t+1)})ECor(C_i^{(t)},C_j^{(t+1)})\right) \tag{13}$$

根据公式(13),波动性 $CFlu(g^{(t)},g^{(t+1)})$定义为

$$CFlu^{(t,t+1)}=1-CCor^{(t,t+1)} \tag{14}$$

此外,如何定量地衡量一个社团的性质也具有重要意义.先前的研究主要从稳定性[5]、波动性[5]和健壮性的角度来衡量社团演化模式,本文将从社团结构特性出发,提出一种社团评价标准,以便于研究社团结构和演化之间的关系.如图 6 所示:左图描述了一个社团结构 C(各边上的数字代表边的权重);右边是一个相同大小的完全图,其每条边的权重是左边社团总权重的 $2/n(n-1)$.这里把这种结构作为图 6 左图社团的评价结构,记为 $std(C)$.对于社团 C 来讲,其评价值可定义为

$$Eva(C)=D(C,std(C))=\sum_{\forall(v,u)\in E(std(C))}\left|\log\frac{w_C(v,u)}{w_{std(C)}(v,u)}\right|\Big/|E(C)| \tag{15}$$

公式(15)既考虑了社团结构的紧密程度,同时又衡量了社团内部各点之间边权重分布的均匀程度,其值也可认为是社团 C 与 $std(C)$的差别.这个定义主要用来分析社团结构与演化之间的关系.

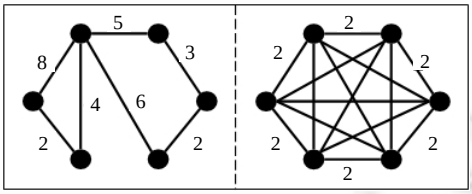


Fig.6 Community and its corresponding complete structure

图 6 社团和与其对应的评价结构

4 实验及讨论

4.1 数据集介绍

VAST 数据集来自 IEEE VAST 2008,是一个开放的竞赛项目的数据集.它包括了一组涉及 400 人左右的 10 天通话数据,已知在这 10 天中,这个小型的社会结构发生了一次社会变革.Enron Email 数据集来自 Enron 公司内部的邮件联系网络,涉及的大约 150 名用户大都为公司高层.其时间跨度为 111 周(1999/12~2002/03).在这段时间内,有许多标志性事件发生,其中包括 CEO 的变动、公司破产等.Cond-mat 来自 Cornell e-Print Cond-mat 数据库,时间跨度为 70 周(2001/03~2006/12).在这个数据集中,所有记录均代表论文中作者的合作关系.Calls *A* & *B*

均来自同一个公司内部员工的通话数据(既包括普通员工,也包括公司高层),*A* 跨越了 187 天(2005~2006),*B* 跨越了 152 天(2007~2008).在 *B* 时间段,该公司发生了一些重要的人事变动情况,尤其是高层领导的变动.Calls *C* 来自某运营商在某一城市 2005~2006 的通话记录,该记录包括了比较详细的通话信息,如通话时长、通话双方的年龄、通话所涉及的基站等.Calls *A*,*B*,*C* 这 3 组通话数据均来自国内某一移动运营商.需要声明的是,为了保护用户的隐私,这里使用的所有通话数据的电话号码均被一种唯一标识的 ID 所替换.此项研究工作并不会针对个人用户.

表 2 列出了这 6 个数据的一些基本信息.从 $\bar{\rho}$ 值可以看出:对于一个相对封闭的社会环境,人们的通信对像更为集中;而对于非封闭性网络(如第 6 个数据集),通信对象则较为分散.

Table 2 Datasets

表 2 数据集

Datasets	$\|V(G)\|$	$\|E(G)\|$	$\|V(g^{(i)})\|$	$\|E(g^{(i)})\|$	$\bar{\rho}^*$	Time span
VAST	400	9,834	372	983	2.64	10 (d)
Enron	150	24k	60	219	3.34	111 (w)
Calls *A*	265	113k	167	812	4.83	118 (d)
Calls *B*	352	54k	196	436	2.23	102 (d)
Cond-mat	52k	280k	1k	4k	3.95	117 (m)
Calls *C*	64k	1,090k	7.4k	10.8k	1.5	174 (d)

$\bar{\rho}$: Ratio of edge and vertices, $|E(g^{(i)})|/|V(g^{(i)})|$; Time span: d—day, w—week, m—month

4.2 网络演化追踪

应用本文提到的 Timeline 发现算法,可以快速而有效地得到 6 个网络的重要事件和演化趋势.在图 3(a)中,很容易看到第 7 天和第 8 天之间的明显变化.实际上,在这两天中,这个网络所描述的社会结构发生了重大变革.对于 Enron Email,图 3(b)也很明显地展示了一些重要的时刻,如公司倒闭、前 CEO 自杀等事件.由通话数据可以看出,Calls *A*,*B*,*C* 均呈现出某种周期性,且周期长度大约为 7.通过观察发现,造成这种现象的原因是人们在工作日和周末的通话对象和通话行为有着明显的不同:对于开放性网络,一些局部性的事件较难发现,但可以发现显著的社会性事件,例如图中标记出的国庆、春节、元宵节等(如图 7(c)所示).同样地,数据集 *C* 也呈现出这种现象.从图 7(d)可以看出,Cond-mat 科研合作网呈现出一种不断上升的演化发展趋势,这表明,该研究领域(condensed matter)在逐渐发展,并吸引了越来越多的科研工作者.

基于以上 Timeline 抽取的结果,这里对各个数据集按第 3.2 节介绍的方法进行平稳段抽取,所得结果见表 3(由于 VAST 数据情况较为简单,仅按时间点 7 分为两段,故这里未再列出).基于这些平稳段,应用算法 2 可以进一步获得其近似图.

Table 3 Segmentation

表 3 网络段抽取

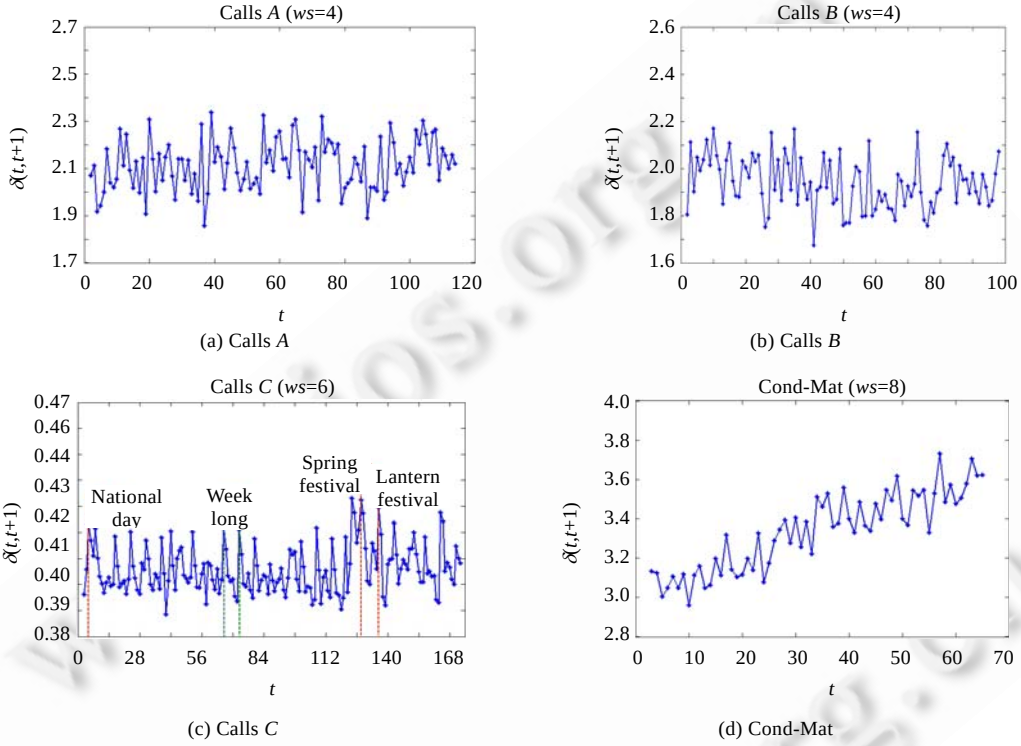
	Enron	Cond-mat	Calls *A*	Calls *B*	Calls *C*
$\|\{S^{(i)}\}\|$	12	15	18	18	16
Avg.$\|S^{(i)}\|$	4	4	5	4	5

$|\{S^{(i)}\}|$: The number of smooth segmentations

Avg.$|S^{(i)}|$: The average number of snapshots in each smooth segment

如前所述,基于事件点两侧近似图的比较能够反映这个事件对网络产生的影响,这里通过列举一些关键事件前后网络的平稳段近似图来说明通过 Timeline 发现的事件所带来的网络结构的真实变化.对于 VAST 数据集,这里取出了时间点 7 前后两个平稳段近似图的主要结构(如图 8(a)、图 8(b)所示).通过对比发现,在这个结构中,高层人员发生了巨大的变化.例如,高层领导从 200 变化到了 300,并且伴有中层领导的明显变化.而从普通角色的个体来看,相比领导层的消失和替换,它们并没有发生太大的变化.对 Enron 数据来讲,图 8(c)、图 8(d)展示了在事件点 96(即 2001/12/2,Enron 公司提出破产保护申请)前后近似图的主要结构.事件发生之前,Louise Kitchen 为公司 CEO(如图 8(c)所示,位于中心用深色标出),其外围均为该公司中层;事件发生之后,虽然 Louise

Kitchen 仍然为公司 CEO,但其外围除公司中层外,增加了一些其他人员,如 Rick Buy(危机处理部门主管),Mark Haedicke(法律部门主管)和一些律师等,这可以从另外一个层面反映公司发生的危机事件.此外,近似图往往能够展示出与 Timeline 相似的演化趋势.如图 9(a)所示,从 Cond-mat 近似图大小的变化趋势可以看出,其演化也呈现出一个上升的过程.对于 Calls C 来讲(如图 9(b)所示),其演化过程要平稳得多.一个例外是 $T^{(8)}$,这一段恰为农历新年时期,可以推断,由于处于假期中,人们与其他人之间,尤其是同事的联系比较少,所以这个近似图的大小要明显小于其他图.本节实验说明,对基于时间的演化网络进行分段并对其中平稳段进行近似,不仅能够描述网络的演化趋势,如网络的上升、平稳等趋势,还能够通过对某些事件两侧近似图的对比研究发现关键事件对整个网络产生的影响.

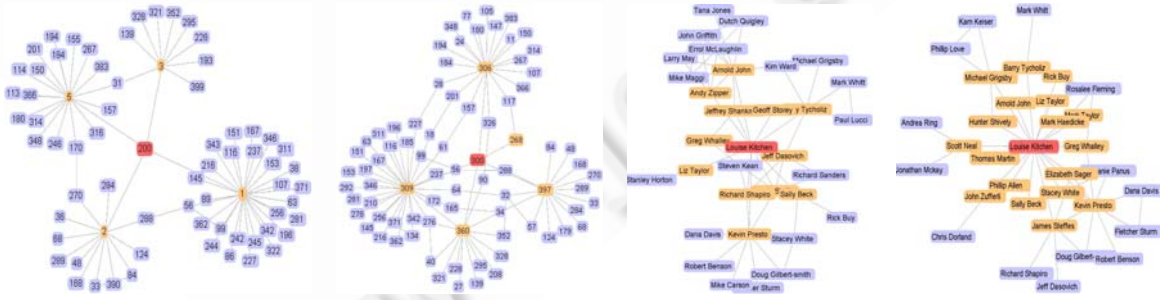


(a) Calls A (b) Calls B

(c) Calls C (d) Cond-Mat

Fig.7 Timelines

图 7 Timelines

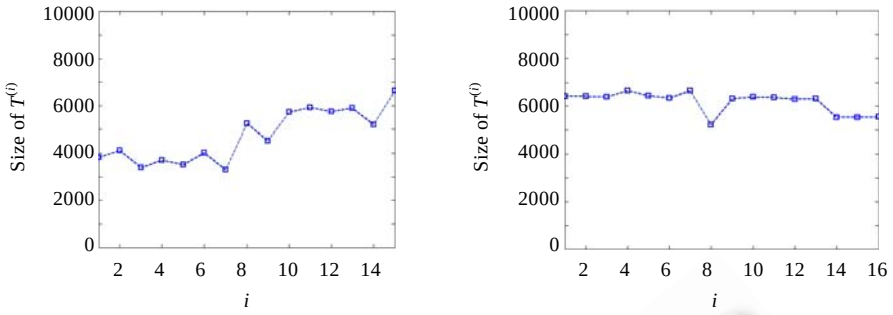


(a) Before the event in VAST (b) After the event in VAST (c) Before the event in Enron (d) After the event in Enron

(a) VAST 数据集事件前 (b) VAST 数据集事件后 (c) Enron 数据集事件前 (d) Enron 数据集事件后

Fig.8 Tracking the main structure of the network

图 8 对网络的主要结构的追踪



(a) Cond-Mat (b) Calls *C*

Fig.9 Size of approximate graphs ($T^{(i)}$)

图 9 近似图 $T^{(i)}$的大小

4.3 社团发现及演化追踪

除了对整个网络演化的分析,社团演化分析在实际应用中有着重要的意义.这一节将从社团角度讨论本文框架,并分析社团的演化特点.

通过与其他两种经典算法比较,表 4 从 4 个方面列出了 3 个通话数据集基于近似图的社团发现结果(CBCD: clique-based community detection,即第 3.4 节介绍的社团发现算法.运行参数 k=3,Q_v=0.6;表 4 中的"*"表示运行环境为 Intel Xeon CPU 2.60GHz 双核,2G 内存.由于 CPM 算法仅仅涉及 clique 的合并,所以 CBCD 的记录时间仅包括第 3.4 节算法描述的前两步骤).由于本文提出框架的主要目的是发现演化网络的动态特征,所以本文提出的社团发现算法并未强调网络节点的覆盖率.从该表可以看出,CBCD 运行所得到的社团数量和大小要小于其他两种算法.从运行效率来讲,CBCD 的运行时间要远小于 GN 算法,与 FAST 算法相当.从合理性的角度来讲,可以说网络社团的演化是一个渐进的过程,很少出现一些显著的变化[10,12],即整个网络的社团演化应该具有较小的 *CFlu* 值(或较大的 *CCor* 值).从表 4 最后一列可以看出,CBCD 的 *CCor* 要明显高于其他两种算法,这说明其划分的社团结构更加合理.

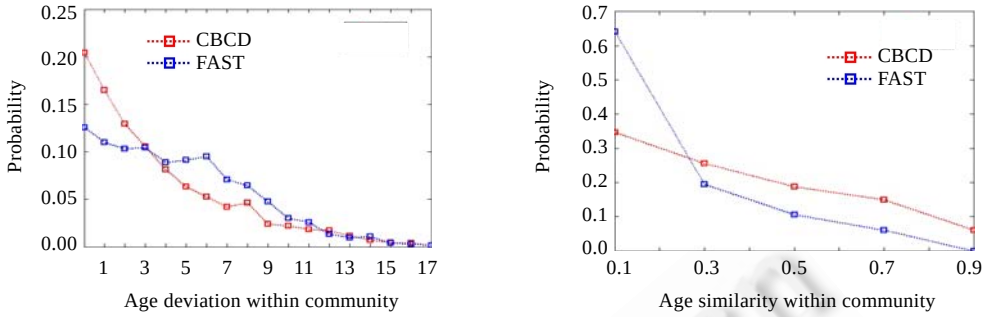
Table 4 Results of community detection algorithms on call graphs

表 4 几种社团发现结果的比较

Name	$\lvert\{T^{(i)}\}\rvert$	Alg.	$\lvert\{C^{(i)}\}\rvert$	$\bar{V}(C^{(i)})$	Avg.Time (s)*	Avg.*CCor*
Calls *A*	18	GN[5]	13	13	4.3	0.019
		FAST[24]	11	16	0.10	0.015
		CBCD	10	13	0.46	0.037
Calls *B*	18	GN	14	11	1.5	0.016
		FAST	23	8	0.13	0.015
		CBCD	7	10	0.35	0.017
Calls *C*	16	GN	356	10	59.3	0.031
		FAST	221	12	1.8	0.031
		CBCD	106	9	1.9	0.037

此外,从其社团的实际意义来讲,社会网络中的社团应该更能反映人们的某种行为趋向.虽然对较大网络来看(如 Calls *C*)很难验证发现的社团的合理性,但从社会学角度来看,社团成员应具有某种同质性(homogeneity,如相似的年龄、社会地位、爱好等)[7].为了验证本节提出的社团发现算法的实效,这里对 Calls *C* 数据中基于 FAST 和 CBCD 两种算法所获得的社团结构进行了评价.图 10(a)描述了社团内成员年龄的均方差,值越大,表明社团成员的年龄差别越大.可以看出,基于 CBCD 算法所发现的社团结构成员年龄差别要小于应用 FAST 算法所得到的结果.除了年龄的相似特征以外,社团成员还具有相似的地域活动范围[17].图 10(b)描述了社团成员活动地域的相似程度(通过其通话的基站信息获得,其值越大,说明社团成员的活动地域越相似),可以看出,使用 CBCD 算法获得的社团成员的活动地域更加相似,甚至相似度能达到 0.9(FAST 算法获得的结果没有能达到相

似度为 0.9 的社团).



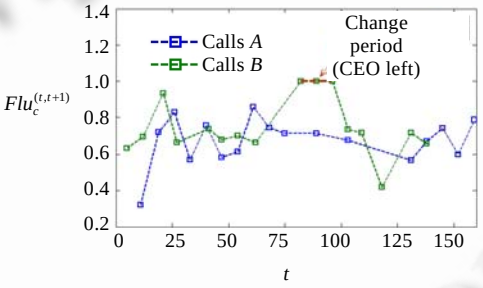
(a) Age deviation within community
(a) 社团内个体年龄的平均标准差

(b) Mobility similarity within community
(b) 社团内个体移动位置的相似度

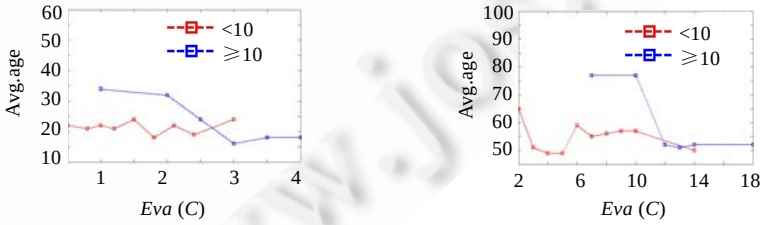
Fig.10 Community evaluation

图 10 社团评价

对于大型社会网络,timeline 可能是比较平稳的,即从整个网络的角度很难发现某些局部事件.这时,timeline 仅反映一些社会性事件,如人们在周末、节日期间和在平时工作时间通话模式的不同.为了发现网络的局部事件,可以使用社团波动分析来跟踪社团的演化.图 11(a)展示了同一公司在两段不同时期社团的波动变化情况,其中:较深色代表这个公司第 1 段时期的社团演化(Calls *A*,2005/10~2006/03),可以看出其波动一直是一个比较均匀的值;较浅色代表这个公司在第 2 段时期的社团演化(Calls *B*,2007/12~2008/04).可以明显地看到一段较高的波动区域(80~100).实际上,公司在这段时间发生了一系列中、高层领导的人事变动.



(a) Fluctuation of communities in Calls *A* and Calls *B*
(a) Calls *A* 和 Calls *B* 中社团演化的波动情况



(b) Community evolution of Cond-mat
(b) 对 Cond-mat 中社团的演化分析

(c) Community evolution of Calls *C*
(c) 对 Calls *C* 中社团的演化分析

Fig.11 Analysis on community evolution

图 11 社团演化分析

除了关键事件对社团演化的影响以外,人们还关注具有何种结构特征的社团能够“活”得更久.为了分析社

团结构与其演化寿命(即社团最长演化路径的长度)的关系,这里应用第 3.5 节提到的社团评价方法,针对 Cond-mat 和 Calls *C* 两个数据集进行分析.图 11(b)和图 11(c)描述了社团的平均寿命和 $D(C,std(C))$之间的关系,其中,较深色的线代表大社团,较浅色的线代表小社团.可以看出:在 $D(C,std(C))$值较小的情况下,大社团趋向于具有更长的寿命.同时,随着 $D(C,std(C))$的增加,当达到某一值后,其寿命将显著下降,并最终趋于平稳;而对于小社团,$D(C,std(C))$对其寿命的影响则比较小.通过对社团寿命的考察发现:对于大社团,如果具有一个紧密的内部关系,并且这种内部关系的权重比较均匀(即社团内部个体的角色相似、地位相近)时,其演化更加稳定,活得更长;而对于不具有这种特性的大社团和规模较小的社团来讲,这种现象则不明显.

5 结论及展望

社会网络研究的兴起,为人们研究现实社会人类的交往行为提供了有效的工具.本文从网络演化出发,提出了一种全新的社会网络分析框架,用来分析动态网络的时间特征和演化模式.不同于传统的直接基于时间片的演化分析,本文首先利用 timeline 来发现网络的演化趋势,并基于事件进行网络分段,通过对抽取出的平稳演化段的近似,得到描述其结构的近似图.这样,对网络的演化分析就转换为对这些近似图的追踪分析.这样做不仅能够简化分析,而且能够有效地降低噪声带来的影响.本文还提出了基于近似图的社团划分及追踪方法,通过对多个真实社会网络的分析,验证了本文提出的分析方法的有效性和实用性.此外,从社团演化的角度可以看到,社团结构与其寿命有着紧密的联系.实际上,人们对其自身复杂的行为模式还知之甚少,本文提出的方法和得到的结论只是从社会网络演化的角度进行了一个方向性的探索.下一步工作将涉及网络或社团中特定模式的研究,如人们通信行为和活动范围(基站信息)的周期性等.此外,社团的演化被认为是一个复杂的时间过程,还有许多其他因素会影响社团的演化模式,对这些因素的分析也将是下一步工作的重点.

致谢 感谢匿名审稿人给予本文的珍贵建议以及本课题组其他成员对本文数据和实验提供的有力支持和帮助.

References:

[1] Sun JM, Faloutsos C, Papadimitriou S, Yu PS. GraphScope: Parameter-Free mining of large time-evolving graphs. In: Proc. of the KDD 2007. 2007. 687−696. http://www.sigkdd.org/kdd2007/

[2] Asur S, Parthasarathy S, Ucar D. An event-based framework for characterizing the evolutionary behavior of interaction graphs. In: Proc. of the KDD 2007. 2007. 913−921. http://www.sigkdd.org/kdd2007/ [doi: 10.1145/1281192.1281290]

[3] Girvan M, Newman MEJ. Community structure in social and biological networks. Proc. of the National Academy of Sciences of the United States of America, 2002, 99(12):7821−7826. [doi: 10.1073/pnas.122653799]

[4] Hopcroft J, Khan O, Kulis B, Selman B. Tracking evolving communities in large linked networks. Proc. of the National Academy of Sciences, 2004,101(Suppl. 1):5249−5253. [doi: 10.1073/pnas.0307750100]

[5] Palla G, Barabási AL, Vicsek T. Quantifying social group evolution. Nature, 2007,446(7136):664−667. [doi:10.1038/nature05670]

[6] Berger-Wolf TY, Saia J. A framework for analysis of dynamic social networks. In: Proc. of the KDD 2006. 2006. 523−528. http://www.sigkdd.org/kdd2006/ [10.1145/1150402.1150462]

[7] Jure L, Jon MK, Christos F. Graphs over time: Densification laws, shrinking diameters and possible explanations. In: Proc. of the KDD 2005. 2005. 177−187. http://www.sigkdd.org/kdd2005/ [doi: 10.1145/1081870.1081893]

[8] Lin YR, Chi Y, Zhu SH, Sundaram H, Tseng BL. Facetnet: A framework for analyzing communities and their evolutions in dynamic networks. In: Huai JP, Chen R, eds. Proc. of the WWW 2008. New York: ACM Press, 2008. 685−694. [doi: 10.1145/1367497.1367590]

[9] Long B, Xu XY, Zhang ZF, Yu PS. Community learning by graph approximation. In: Proc. of the ICDM 2007. 2007. 232−241. http://www.ist.unomaha.edu/icdm2007/ [doi: 10.1109/ICDM.2007.42]

[10] Backstrom L, Huttenlocher D, Kleinberg J, Lan XY. Group formation in large social networks: Membership, growth, and evolution. In: Proc. of the KDD 2006. 2006. 44−54. http://www.sigkdd.org/kdd2006/ [doi: 10.1145/1150402.1150412]

[11] Jin EM, Girvan M, Newman MEJ. Structure of growing social networks. Physical Review E, 2001,64(4):046132. [doi: 10.1103/PhysRevE.64.046132]

[12] Tantipathananandh C, Berger-Wolf TY, Kempe D. A framework for community identification in dynamic social networks. In: Proc. of the KDD 2007. 2007. 717−726. http://www.sigkdd.org/kdd2007/ [doi: 10.1145/1281192.1281269]

[13] Tong HH, Papadimitriou S, Sun JM, Yu PS, Christos F. Colibri: Fast mining of large static and dynamic graphs. In: Proc. of the KDD 2008. 2008. 686−694. http://www.sigkdd.org/kdd2008/ [doi: 10.1145/1401890.1401973]

[14] Newman MEJ. Fast algorithm for detecting community structure in networks. Physical Review E, 2004,69(6):066133. [doi: 10.1103/PhysRevE.69.066133]

[15] Dérenyi I, Palla G, Vicsek T. Clique percolation in random networks. Physical Review Letters, 2005,94(16):160202. [doi: 10.1103/PhysRevLett.94.160202]

[16] Tomita E, Tanaka A, Takahashi H. The worst-case time complexity for generating all maximal cliques and computational experiments. Theoretical Computer Science, 2006,363(1):28−42. [doi: 10.1016/j.tcs.2006.06.015]

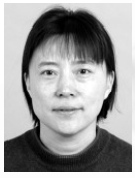
[17] González MC, Hidalgo CA, Barabási AL. Understanding individual human mobility patterns. Nature, 2008,453(7196):779−782. [doi:10.1038/nature06958]

[18] Bollinger JA. Bollinger on Bollinger Bands. New York: McGraw-Hill, 2001.

[19] Rossman B. On the constant-depth complexity of *k*-clique. In: Dwork C, ed. Proc. of the 40th Annual ACM Symp. on Theory of Computing. Victoria: ACM Press, 2008. 721−730.

吴斌(1969－),男,湖南长沙人,博士,副教授,CCF 高级会员,主要研究领域为数据库,数据挖掘,复杂网络与复杂系统.

杨胜琦(1984－),男,硕士生,主要研究领域为数据挖掘,社会网络分析.



王柏(1962－),女,博士,教授,博士生导师,CCF 高级会员,主要研究领域为分布式计算技术,数据挖掘.