软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2014,25(3):547−559 [doi: 10.13328/j.cnki.jos.004462]

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

基于信息流动分析的动态社区发现方法*

索 勃, 李战怀, 陈 群, 王 忠

(西北工业大学 计算机学院,陕西 西安 710072)

通讯作者: 索勃, E-mail: caitou@mail.nwpu.edu.cn

摘 要: 随着社交网络和微博等互联网应用的逐渐流行,其用户规模在迅速膨胀.在这些大规模网络中,社区发现可以为个性化服务推荐和产品推广提供重要依据.不同于传统的网络,这些新型网络的节点之间除了拓扑结构外,还进行频繁的信息交互.信息流动使得这些网络具有方向性和动态性等特征.传统的社区发现方法由于没有考虑到这些新的特征,并不适用于这些新型网络.在传染病动力学理论的基础上,从节点间信息流动的角度,提出一种动态社区发现方法.该方法通过对信息流动的分析来发现联系紧密、兴趣相近的节点集合,以实现动态的社区发现.在真实数据集上的实验结果表明:相对于传统的社区发现方法,所提出的方法能够更准确地发现社区,并且更能体现网络中社区的动态变化.

关键词: 社交网络;社区发现;信息流动分析;传染病动力学模型

中图法分类号: TP311 **文献标识码**: A

中文引用格式: 索勃,李战怀,陈群,王忠.基于信息流动分析的动态社区发现方法.软件学报,2014,25(3):547−559 http://www.jos.org.cn/1000-9825/4462.htm

英文引用格式: Suo B, Li ZH, Chen Q, Wang Z. Dynamic community detection based on information flow analysis. Ruan Jian Xue Bao/Journal of Software, 2014,25(3):547−559 (in Chinese). http://www.jos.org.cn/1000-9825/4462.htm

Dynamic Community Detection Based on Information Flow Analysis

SUO Bo, LI Zhan-Huai, CHEN Qun, WANG Zhong

(College of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China)

Corresponding author: SUO Bo, E-mail: caitou@mail.nwpu.edu.cn

Abstract: As the Internet applications, such as social networks and micro-blogs, become popular, their scale of users has been increasing rapidly. Community detection in these large-scale networks could provide important insights into customer behavior for service recommendation and product marketing. The difference of these networks from traditional ones is that besides topology, they have frequent information interaction between nodes. Information flow makes these networks directed and dynamic. Traditional community detection approaches fall short in these networks because they do not consider these new characteristics. Inspired by the dynamics of infectious disease theory, this paper proposes a novel community detection approach based on information flow analysis. This approach effectively groups the nodes with frequent information interaction in the same community. Between communities, there would be little information flow. This paper experiments on real-world networks demonstrate that compared with previous community detection methods, the proposed approach is more effective at identifying the dynamics in the networks.

Key words: social network; community detection; information flow analysis; epidemic model

复杂网络分析在社会学、传染病学和生物学等领域有着广泛的应用.网络由节点的集合构成,节点之间的联系通过边来表示.网络中节点之间的联系可以通过不同的方式建立,例如:超链接文档构成的网络中,节点间

* 基金项目: 国家重点基础研究发展计划(973)(2012CB316203); 国家自然科学基金(61033007); 国家高技术研究发展计划(863)(2012AA011004); 西北工业大学研究生创业种子基金(Z2013125, Z2013126)

收稿时间: 2012-02-21; 修改时间: 2012-07-23; 定稿时间: 2013-07-30

的联系通过链接建立;引文作者网络中,作者间的联系则通过文章引用建立.

在复杂网络中,社区发现是抽取网络特征、理解网络内在结构的重要手段.社区发现通过对所有节点所构成的集合进行划分,使得同一子集内的节点关系“密切”,而不同子集间关系“松散”.合理地定义“密切”和“松散”的具体含义,可以使得社区内部节点呈现出某些共同属性.例如:在超链接文档构成的互联网中,同一社区内的文档通常具有相似的话题或观点;在引文作者网络中,每个社区内的作者都有着相近的研究方向;在生物网络中,相同社区的蛋白质形成一个功能团,共同完成某项生物功能;在社交网络中,相同社区内的用户有着相近的兴趣爱好.现有的社区发现方法都是围绕着如何判断节点间的关系“密切”或者“松散”而展开.简单的方法将社区定义为完全图,或者条件较为宽松的类完全图;稍复杂的方法通过定义“质量函数”对网络进行划分,其中,Clauset,Newman 和 Moore 提出的 CNM 算法[1]已在许多种网络中得到成功应用.

传统的社区发现方法中,网络作为静态拓扑图处理,不用考虑节点间的信息交互因素,这在微博等逐渐兴起的社交网络中不再适用.在微博及其应用所构成的社交网络中,不同节点间的信息交互非常频繁;拓扑结构仅代表用户之间交互的可能性,而实际交互的程度则由节点之间的信息流动情况决定.单纯依靠拓扑结构的社区划分方法因为忽略了社交网络中的信息流动,具有明显的局限性,不能适用现代社交网络的新的特征,划分结果的准确性无法得到保证.文献[2]的研究指出:到 2009 年 6 月,Twitter 的用户平均拥有 126 个关注者;简单地用拓扑结构来衡量节点交互,会导致社区划分结果不准确,社区的整体结构也会趋于静态.

针对传统的社区发现方法在解决动态网络(如社交网络)社区划分时所面临的问题,本文以传染病动力学理论为基础,提出了一种基于信息流动分析的动态社区发现方法.我们首先利用平均接触率来量化节点间的交互程度,然后再运用含有信息流动特征的模块度函数 Q_t 来实现社区划分(T-CNM 算法).我们的动态社区发现方法可以取得这样的一个结果:社区内部信息流动频繁,而社区之间则只存在偶尔的信息流动.为了验证方法的有效性,我们将该方法应用于真实的动态网络中,包括微博网络以及论文作者相互引用所构成的网络,并与 CNM 算法[7]和基于着色的动态社区发现算法[5]进行了比较.实验结果表明,基于信息流动分析的社区发现方法能更够准确地反映出动态网络中社区的结构及其变化.

本文的主要贡献如下:

1) 通过对比传染病传播过程和信息流动过程的异同点,对社交网络中信息流动的过程进行分析,阐明了将传染病动力学模型应用于社交网络的合理性(见第 2.1 节);
2) 结合传染病动力学模型,提出了利用平均接触率来反映社交网络中信息流动过程的方法,对节点交互环境下节点间关系的紧密程度进行量化(见第 2.2 节);
3) 在社区发现质量函数 Q 的基础上,加入节点之间的信息流动特征,提出了社区模块度函数 Q_t 和 T-CNM 算法,提高了社区发现的层次性和合理性,并能够从社区的角度反映出各时刻社交网络结构的变迁过程(见第 2.3 节).

本文第 2 节对比了传染病传播过程和信息流动过程,并描述如何应用信息流动分析来计算节点间平均接触率,以及利用 T-CNM 算法来实现动态社区发现.第 3 节是实验验证.第 4 节介绍相关工作.第 5 节总结我们的工作,并展望未来的研究工作.

1 传染病动力学理论

传染病动力学理论主要采用“仓室”数学模型.其中,SI 仓室模型具备传染病动力学模型的所有基本特征,也能够满足动态网络中信息流动分析的需求.在以下讨论中,我们将以 SI 仓室模型为基础介绍信息流动分析方法.SI 仓室模型针对某类传染病将相关地区的人群分成两类,或者说两个仓室:易感者(susceptibles)类和染病者(infectives)类.易感者类表示 t 时刻未染病但有可能被该类疾病传染的人群,我们把它的数量记为 $S(t)$;染病者类表示 t 时刻已被感染成病人而且具有传染力的人群,我们把它的数量记为 $I(t)$.假设总人口数为 $N(t)$,考虑在一个封闭的环境中,传染病的感染比人口的出生、死亡、流动等种群动力因素显著得多,可以认为此环境中总人口始终保持为一个常数,即 $N(t)=k$,则有 $S(t)+I(t)=k$.若将传染病自 S 仓室到 I 仓室的传染过程划分为许多相等的时

间片,并假定传染仅发生在 t 到 t+1 的过程中.将 t 时已经被感染了 θ 个时间片的染病者数量定义为 $I_\theta(t)$,则:

$$I(t)=\sum_{\theta=0}^{t} I_\theta(t) \tag{1}$$

这样,从 t 到 t+1 的传染过程中,易感者变为染病者的数量为 $I(t+1)-I(t)$.

传染病的全部传染过程如图 1 所示,其中,灰色节点表示 t 时之前已存在的染病者集合,白色节点表示 t 时新增加的染病者集合.虚线标记集合随时间的延续,实线表示传染过程.可以看出:t+1 时,染病者的数量由 t 时已存在的染病者(灰色节点)和这些染病者新传染的易感者(白色节点)两部分构成.

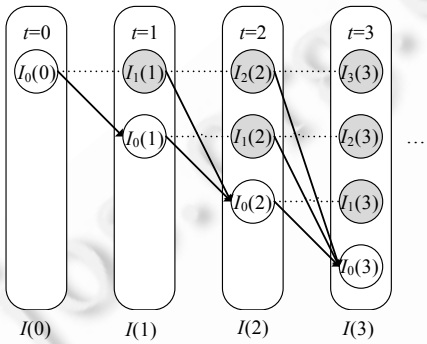


Fig.1 Spread of epidemic

图 1 传染病传染过程

假定传染病传染的"强度"随染病者已被传染的时间而变化,记为 β_θ.它表示 θ 时,单位时间内每个染病者对每个易感者的传染率.这样,t 时到 t+1 时,图 1 中所示的传染病传染过程可表示为

$$I(t+1)=I(t)+S(t)\cdot\sum_{\theta=0}^{t}\beta_\theta I_\theta(t) \tag{2}$$

假定各时间片大小趋于 0,则可认为 t 连续增加,由公式(2)可得,当 $\Delta t\to 1$ 时:

$$\lim_{\Delta t\to 1} I(t+\Delta t)-I(t)=S(t)\cdot\int_0^t \beta_\theta I_\theta(t)\mathrm{d}\theta\cdot\Delta t \tag{3}$$

通常认为,在传染病从 S 仓室到 I 仓室的传染过程中,传染率 β_θ 保持为常数 β,则由等式(3)可以得出染病者数量的增长速度为

$$\lim_{\Delta t\to 1}\frac{I(t+\Delta t)-I(t)}{\Delta t}=\beta\cdot S(t)\cdot\int_0^t I_\theta(t)\mathrm{d}\theta=\beta\cdot S(t)\cdot I(t) \tag{4}$$

2 基于信息流动分析的社区发现方法

2.1 基于传染病动力学模型的信息流动分析

在社交网络中的节点针对某一信息也可以划分为两个仓室:传播者(senders)类和接收者(receivers)类,这与传染病动力学中的 SI 仓室模型完全相同.传播者类表示 t 时刻已传播某一信息的节点集合,与传染病动力学中染病者类似,将传播者数量标记为 $I(t)$;接收者类与易感者相似,表示 t 时刻已接收到某一信息,可能会传播这条信息的节点集合,将其数量标记为 $S(t)$.这样,社交网络中信息传播的过程就变为接收者类节点逐渐变为传播者的过程,如图 2 所示.同样,虚线标记集合随时间的延续,灰色节点表示 t 时之前存在的信息传播者集合,白色节点表示 t 时新增的传播者集合,阴影节点表示 t 时所有非传播者节点.

可以看出,信息流动过程与传染病传播过程都是将人群总体划分为两个仓室,并在两个仓室的动态交互过程中保持人群总数的恒定,交互过程方向始终保持不变,且过程不可逆.这些都是传染病传播和信息流动的前提,二者完全相同.在 SI 仓室模型中,假设所有个体"充分接触",即通过在总人口中随机选择来决定每个易感者

所接触的个体.并且假定 t 时每个个体接触的人群数量均为 $S(t)$,每个个体的接触率都等于β_θ.显然,将这些假设应用于社交网络不再准确.综上,传染病传播过程和信息流动过程的异同点见表 1.

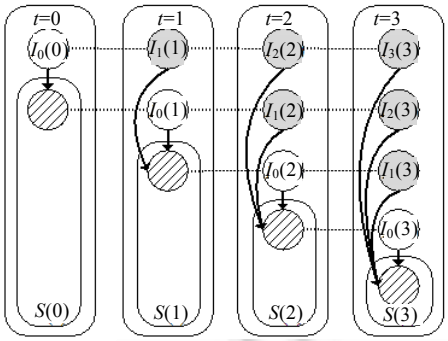


Fig.2 Flow of information

图 2 信息流动过程

Table 1 Similarities and differences of the transmissions of epidemic and information

表 1 传染病传播和信息流动过程异同

	传染病传播过程	信息流动过程
仓室数量	两个:染病者、易感者	两个:传播者、接收者
人群总量	恒定	恒定
仓室交互方向	不变:由染病者向易感者	不变:由传播者向接收者
过程可逆	不可逆	不可逆
个体接触	充分接触,随机传播	由拓扑结构决定
接触程度	与染病者和易感者无关	与传播者和接收者相关

针对信息流动过程的特殊性,由于社交网络中每个节点只能随机接触其拓扑相邻的节点,接触的节点数量由相邻的节点数量决定.因此,传播的过程如图 2 所示,信息仅能从传播者集合(灰色、白色)向他们相邻的节点(阴影)传播.我们用$\Theta(v,t)$表示 t 时社交网络中节点 v 与相邻节点的接触在 $S(t)$中的比例,$I_\theta(t)$仍表示已将信息传播了θ个时间片的节点所构成的集合.则在社交网络中,等式(2)变为

$$I(t+1)=I(t)+S(t)\sum_{v\in\bigcup_{\theta=0}^{t}I_\theta}\Theta(v,t)\cdot\beta_{v,t} \tag{5}$$

其中,$\beta_{v,t}$表示 t 时节点 v 与相邻节点的接触率.

2.2 节点间信息流动的平均接触率

在社交网络信息流动分析中,我们希望通过信息的传播过程,准确反映出节点间的关系,而且节点交互的历史信息记录了每个时刻信息传播的情况.这样,t 时信息传播者的数量 $I(t)$、接收者的数量 $S(t)$均为已知.在此,$\beta_{v,t}$便反映了 t 时节点 v 传播的某一信息对它相邻节点的影响程度,影响大,则接触率高;反之,接触率低.定义Δt_{vu}表示从节点 v 将信息传播至节点 u 到节点 u 变为传播者的时间间隔.在社交网络分析中,信息传播的方向、时间已经确定,因此,对于每个Δt_{vu}得出的是信息传播时的平均接触率,表示为 $\overline{\beta}_{v,t}$,且 $\overline{\beta}_{v,t}=\overline{\beta}_{v,t+\Delta t_{vu}}$.这样,等式(5)变为

$$I(t+\Delta t_{vu})-I(t)=S(t)\Theta(v,t)\cdot\overline{\beta}_{v,t}\cdot\Delta t_{vu}+\sum_{\Delta t_{mn}\in(0,\Delta t_{vu})}\sum_{w\in I_\delta}S(t+\Delta t_{mn})\Theta(k,t+\Delta t_{mn})\cdot\overline{\beta}_{w,t}\cdot\Delta t_{mn} \tag{6}$$

等式右侧第 1 项表示节点 v 相邻节点由信息接收者变为传播者的数量;第 2 项表示Δt_{vu}时间间隔内,除节点 v 外其他节点所引起的传播者数量变化.对于等式(6)左侧,在Δt_{vu} 的时间间隔内,每当有信息接收者变为信息传播者,右侧则有 $S(t+\Delta t_{mn})\Theta(k,t+\Delta t_{mn})\cdot\overline{\beta}_{w,t}\cdot\Delta t_{mn}$ 与它对应.这样,t 时,节点 u 由信息接收者变为传播者,节点 v 与其相邻节点的平均接触率 $\overline{\beta}_{v,t}$ 为

$$\bar{\beta}_{v,t}=\frac{1}{S(t)\Theta(v,t)\cdot\Delta t_{vu}} \tag{7}$$

在具体计算时,$\Theta(v,t)$可根据具体问题不同定义.我们接下来的讨论和实验中,将$\Theta(v,t)$定义为 t 时节点 v 相邻的节点中接收者占所有接收者的比例,因此,$S(t)\cdot\Theta(v,t)$为 t 时节点 v 相邻节点中的接收者数量.

对于社交网络 $G(V,E)$,其中,V 和 E 分别是节点和边的集合,设 T 是时序构成的集合.假定节点间所传播的信息 r 构成集合 R,则 $r\in R$.信息 r 的传播者构成集合 $V_r=\{v_r\}\subseteq V$.$\Gamma:(V,R)\to T$ 为信息 r 和传播者与时序的映射.则具体计算节点 v 传播信息 r 时的平均接触率 $\bar{\beta}_{v,\Gamma(v,r)}$ 的方法见算法 1.

算法 1. $\bar{\beta}_{v,\Gamma(v,r)}$ Calculation Algorithm.

Input: $G(V,E,T)$, R, $\Gamma:(V,R)\to T$;

Output: $B=\{\bar{\beta}_{v,\Gamma(v,r)}\mid v\in V,r\in R\}$.

1. begin
2. for $r\in R$ do
3. $V_r\leftarrow\{v_r|$all the infectors of $r\}$;
4. for $v\in V_r$ do
5. $V_r'\leftarrow\{v_r\mid v_r\in V_r\wedge(v,v_r)\in E\}\cup\{v\}$;
6. sort V_r' in ascending order according to T using Γ;
7. for $u\in V_r'\wedge u\neq v$ do
8. $\bar{\beta}_{v,\Gamma(v,r)}\leftarrow 1/((\Gamma(u,r)-\Gamma(v,r))\cdot|V_r'|)$;
9. $V_r\leftarrow V_r-\{u\}$;
10. end

算法 1 对于每一条 R 中的信息,首先取出所有这条信息的传播者构成集合 V_r(第 2 行、第 3 行);然后,针对每一个染病者,取出传播相同信息的临界点子集,并对这些具有邻接关系的传播者按传播时间进行排序(第 5 行、第 6 行);最后,根据信息传播的方向以及信息在两节点间的传播时间,利用公式(7)计算出对应节点对间的平均接触率(第 8 行).由于信息只能从传播者向接收者传播,因此在每次平均接触率计算结束后,需要将对应的传播者从接收者集合移出(第 9 行).

以上为传播者 v 相邻节点 u 由信息 r 的接收者变为传播者时,v 的平均接触 $\bar{\beta}_{v,\Gamma(v,r)}$ 的计算方法.由于在社区划分过程中需要考虑的是各个节点之间的信息交流,因此需要得到$\Gamma(u,r)$时节点 v 与相邻节点 u 的接触率.又因为节点 v 与相邻节点的平均接触率近似等于 v 向相邻节点 u 传递信息 r 的过程中,$\Gamma(u,r)$时的平均接触率 $\bar{\beta}_{v,\Gamma(u,r)}$,本文为了减少计算的复杂程度,将 $\bar{\beta}_{v,\Gamma(u,r)}$ 近似当作 v 与相邻节点 u 在$\Gamma(u)$时的平均接触率 $\bar{\beta}_{v,u,\Gamma(u,r)}$ 进行后续计算.即: $\bar{\beta}_{v,u,\Gamma(u,r)}=\bar{\beta}_{v,\Gamma(u,r)}$.令$\Delta t_{vu}=\Gamma(u,r)-\Gamma(v,r)$,则 $\bar{\beta}_{v,u,\Gamma(u,r)}=\bar{\beta}_{v,\Gamma(v,r)+\Delta t_{vu}}=\bar{\beta}_{v,\Gamma(v,r)}$.由此,根据算法 1 即可求出节点间在$\Gamma(u,r)$时刻的平均接触率 $\bar{\beta}_{v,u,\Gamma(u,r)}$.

以上方法所求得的节点 u,v 间的接触率在 T 上并不连续,而是由节点转发的信息 r 的时间$\Gamma(u,r)$决定的.为了体现出信息接收者对于传播者所传播信息接受程度随时间的连续变化,我们对 $\bar{\beta}_{v,u,\Gamma(u,r)}$ 进行插值.假设节点 v 传播的信息的时序构成集合$\{\Gamma(u,r)|r\in R\wedge(u,v)\in E\}$,对于其元素按升序排序得到节点 v 的相邻节点传播时序 $T_v=\{\tau_1,\tau_2,\ldots,\tau_i,\ldots\}$,则任意时刻 $t(\tau_i<t<\tau_{i+1})$时节点 v 与节点 u 之间的平均接触率为 $\bar{\beta}_{v,u,\tau_i}$ 和 $\bar{\beta}_{v,u,\tau_{i+1}}$ 的加权平均:

$$\bar{\beta}_{v,u,t}=\frac{\omega_{t,\tau_{i+1}}}{\omega_{\tau_i,t}+\omega_{t,\tau_{i+1}}}\bar{\beta}_{v,u,\tau_i}+\frac{\omega_{\tau_i,t}}{\omega_{\tau_i,t}+\omega_{t,\tau_{i+1}}}\bar{\beta}_{v,u,\tau_{i+1}} \tag{8}$$

其中,$\omega_{\tau_i,t}$ 和 $\omega_{t,\tau_{i+1}}$ 为时间 τ_i 和 τ_{i+1} 两时刻接触率的相应权值,权值可采用任意合理的表示方式.在本文中,为了体现节点信息传播时间短、平均接触率高的特点,取 $\omega_{\tau_i,t}=(t-\tau_i)^2,\omega_{t,\tau_{i+1}}=(\tau_{i+1}-t)^2$.

$$\bar{\beta}_{v,u,t}=\frac{(\tau_{i+1}-t)^2}{(\tau_{i+1}-t)^2+(t-\tau_i)^2}\bar{\beta}_{v,u,\tau_i}+\frac{(t-\tau_i)^2}{(\tau_{i+1}-t)^2+(t-\tau_i)^2}\bar{\beta}_{v,u,\tau_{i+1}} \tag{8′}$$

由上述公式得到 T 中任意时刻 t 节点 v 到相邻节点 u 平均接触率 $\bar{\beta}_{v,u,t}$,所反映的是节点 v 的相邻节点 u 对于 v 所传播信息的接受程度,平均接触率的高低与接受程度一致.随着节点 u 兴趣的不断变化,节点 u 根据自身对于某话题的关注程度对 v 所传播的信息选择传播或者忽略.通过对比中不同时刻 t 的社区划分的快照,则体现了整个社交网络结构随时间的变化.针对带有各时刻平均接触率的社交网络,我们想要得到的是这样的一种社区划分结果:社区内部信息流动频繁,社区间偶尔存在信息流动.

2.3 T-CNM算法

Newman 和 Girvan 最先提出了模块度 Q 的概念[3],表示在一个有向带权重的网络中,目前的社区结构与具有相同节点数、相同边数、相同权重分布的随机图之间的差别.由于社区内部节点联系“紧密”,因此,属于同一社区的节点的模块度一定高于它们所生成的随机图.通过不断地进行社区合并,寻找网络中模块度的最大值,从而完成社区划分.在实际的应用中,基于模块度的社区划分方法可以利用贪心算法进行优化,Clauset,Newman 和 Moore 提出了一种有效的优化方法,即 CNM 算法[1].

假定社交网络 $G(V,E)$中,W_{vu} 表示节点 v 到 u 的边的权重,W 表示社交网络中所有边的权重之和, s_v^{out} 为从节点 v 流出的权重值和, s_u^{in} 为流入节点 u 的权重之和.G 的社区发现结果Π由 V 中若干节点的集合π_i构成,其中,

$$\Pi=\{\pi_1,\pi_2,...\},\pi_i\cap\pi_j=\varnothing(i\neq j),\bigcup_{\pi_i\in\Pi}\pi_i=V.$$

Q 的定义如下:

$$Q(\Pi)=\sum_{\pi_i\in\Pi}\left(\sum_{v,u\in\pi_i}\frac{W_{vu}}{W}-\frac{s_v^{out}s_u^{in}}{W^2}\right) \tag{9}$$

CNM 算法初始状态下,每个节点单独形成一个社区,随后计算合并任意两个社区π_i,π_j时 Q 的变化 $\Delta Q_{\pi_i,\pi_j}^{\Pi}$.

算法不断地寻找最大的 $\Delta Q_{\pi_i,\pi_j}^{\Pi}$,并对相应的社区$\pi_i$和$\pi_j$进行合并,直至 $\Delta Q_{\pi_i,\pi_j}^{\Pi}$ 降低为止.

CNM 算法采用了两种数据结构来寻找最大的 $\Delta Q_{\pi_i,\pi_j}^{\Pi}$:

(1) 任意两社区对(π_i,π_j)构成的平衡二叉树;

(2) 按照 $\Delta Q_{\pi_i,\pi_j}^{\Pi}$ 排序的任意两社区对(π_i,π_j)构成的最大堆.

我们在CNM算法的基础上,结合节点间信息流动,利用平均接触率进行社区发现,即T-CNM算法.该算法既保留了原有 CNM 算法在社区发现时的优点,又引入了信息流动的分析.这样,节点间的静态拓扑结构不再是影响社区发现结果的唯一要素,社区发现的结果将更具有动态性、时效性.同时,通过 T-CNM 算法可以得到具体某个时刻的社区发现结果.不同于以往从整个网络或者单个节点的粒度对社交网络的变迁过程的分析,T-CNM 算法可以从社区的角度对不同时刻社区发现的结果进行比较,实现社交网络的演进过程的分析.

在考虑到结合节点间平均接触率这一要素后,t 时刻,社区的模块度函数表示为 Q_t,网络中所有节点间的总接触率表示为 W_t, $s_{v,t}^{out}$ 为节点 v 向其他节点流出的总接触率, $s_{u,t}^{in}$ 为流入节点 u 的接触率之和,结合公式(8′),可以得出:

$$W_t=\sum_{v,u\in V}\bar{\beta}_{v,u,t}=\sum_{v,u\in V}\frac{\omega_{t,\tau_{i+1}}}{\omega_{\tau_i,t}+\omega_{t,\tau_{i+1}}}\bar{\beta}_{v,u,\tau_i}+\frac{\omega_{\tau_i,t}}{\omega_{\tau_i,t}+\omega_{t,\tau_{i+1}}}\bar{\beta}_{v,u,\tau_{i+1}},$$

$$s_{v,t}^{out}=\sum_{(v,m)\in E}\bar{\beta}_{v,m,t}=\sum_{(v,m)\in E}\frac{\omega_{t,\tau_{i+1}}}{\omega_{\tau_i,t}+\omega_{t,\tau_{i+1}}}\bar{\beta}_{v,m,\tau_i}+\frac{\omega_{\tau_i,t}}{\omega_{\tau_i,t}+\omega_{t,\tau_{i+1}}}\bar{\beta}_{v,m,\tau_{i+1}},$$

$$s_{u,t}^{in}=\sum_{(n,u)\in E}\bar{\beta}_{n,u,t}=\sum_{(n,u)\in E}\frac{\omega_{t,\tau_{i+1}}}{\omega_{\tau_i,t}+\omega_{t,\tau_{i+1}}}\bar{\beta}_{n,u,\tau_i}+\frac{\omega_{\tau_i,t}}{\omega_{\tau_i,t}+\omega_{t,\tau_{i+1}}}\bar{\beta}_{n,u,\tau_{i+1}},$$

则模块度函数 Q_t 即为公式(9′)所示,具体算法见算法 2.

$$Q_t(\Pi)=\sum_{\pi_i\in\Pi}\left(\sum_{v,u\in\pi_i}\frac{\bar{\beta}_{v,u,t}}{W_t}-\frac{s_{v,t}^{out}\cdot s_{u,t}^{in}}{(W_t)^2}\right) \tag{9′}$$

算法 2. T-CNM.

Input: $G(V,E)$, $B=\{\bar{\beta}_{v,\Gamma(v,r)}\mid v\in V,r\in R\}$, t;

Output: $\Pi=\{\pi_i\}$.

1. begin
2. $\Pi\leftarrow\{v\in V\mid\{v\}\}$;
3. for $v,u\in V$ do
4. $W_t\leftarrow Beta\_Calculation(v,u,t)+W_t$;
5. while (true)
6. for $\pi_i,\pi_j\in\Pi$ do
7. for $v,u\in\pi_i\cup\pi_j$ do
8. $\bar{\beta}_{v,u,t}\leftarrow Beta\_Calculation(v,u,t)$;
9. for $m\in\{m\mid m\in V\wedge(v,m)\in E\}$ do
10. $s_{v,t}^{out}\leftarrow Beta\_Calculation(v,m,t)+s_{v,t}^{out}$;
11. for $u\in\{u\mid u\in V\wedge(n,u)\in E\}$ do
12. $s_{u,t}^{in}\leftarrow Beta\_Calculation(n,u,t)+s_{u,t}^{in}$;
/*Q_t is calculated according to equation (9′)*/
13. $\Delta Q_{\pi_i,\pi_j}^{\Pi}\leftarrow Q_t(\Pi-\pi_i-\pi_j+(\pi_i\cup\pi_j))-Q_t(\Pi)$;
14. $(\pi_i^*,\pi_j^*)\leftarrow\arg\max_{(\pi_i,\pi_j)\in\Pi^2}\Delta Q_{\pi_i,\pi_j}^{\Pi}$;
15. if $(\Delta Q_{\pi_i^*,\pi_j^*}^{\Pi}<0)$ break;
16. $\Pi=\Pi-\pi_i^*-\pi_j^*+(\pi_i^*\cup\pi_j^*)$;
17. end
18. Function float *Beta_Calculation*(v,u,t)
19. begin
20. $B_{v,u}\leftarrow\{\bar{\beta}_{v,\Gamma(u,r)}\mid r\in R\wedge(v,u)\in E\}$;
21. sort $B_{v,u}$ according to $\Gamma(u,r)$ in ascendant order $\{\tau_1,\tau_2,\ldots,\tau_n\}$;
22. if ($t<\tau_1$), $\tau_t\leftarrow t$, $\tau_{i+1}\leftarrow\tau_1$;
23. else if ($t>\tau_n$), $\tau_t\leftarrow\tau_n$, $\tau_{i+1}\leftarrow t$;
24. else find τ_i,τ_{i+1} where $\tau_i<t<\tau_{i+1}$;
25. return $\bar{\beta}_{v,u,t}$; /* $\bar{\beta}_{v,u,t}$ is calculated according to equation (8′)*/
26. end

算法 2 首先令各节点自身形成一个社区(第 2 行),然后调用函数 *Beta_Calculation* 计算接触率的总和 W_t(第 4 行).接着,尝试将任意两个社区进行合并,在合并过程中调用函数 *Beta_Calculation*,计算模块度所需的 $s_{v,t}^{out}$,$s_{u,t}^{in}$ 和 $\bar{\beta}_{v,u,t}$,并计算社区合并后社区整体的模块度 $Q_t(\Pi-\pi_i-\pi_j+(\pi_i\cup\pi_j))$,比较合并前的变化(第 6 行~第 13 行).在计算完所有社区对合并后模块度的变化 $\Delta Q_{\pi_i,\pi_j}^{\Pi}$ 后,取出得到最大的 $\Delta Q_{\pi_i,\pi_j}^{\Pi}$ 时两社区 π_i^* 和 π_j^*(第 14 行).若 $\Delta Q_{\pi_i^*,\pi_j^*}^{\Pi}<0$,则社区发现过程结束(第 15 行);否则,将 π_i^* 和 π_j^* 合并(第 16 行),继续该过程.在计算模块度 Q_t 时,需要用到 t 时刻节点间的平均接触率 $\bar{\beta}_{v,u,t}$,即函数 *Beta_Calculation*.函数 *Beta_Calculation* 首先取出传播者 v 和接

收者 u 在 B 中的所有平均信息接触率 $\overline{\beta}_{v,\Gamma(u,r)}$ (第 20 行),然后将其按时间排序(第 21 行).接着,根据时间 t 找出距离 t 最近的信息传播时刻 τ_i,τ_{i+1}(第 22 行~第 24 行).最后,根据公式(8′)返回求得的节点 v,u 在 t 时刻的平均接触率 $\overline{\beta}_{v,u,t}$ (第 25 行).

3 实 验

为了验证基于信息流动分析社区发现方法的有效性,我们选取了以下几个数据集(见表 2):

(1) 通过新浪微博提供的 API 抓取的 2010 年 12 月部分节点的信息流动记录;

(2) 2010 年 DBLP 引文文献数据集[4].

Table 2 Statistics of the datasets

表 2 数据集统计特征

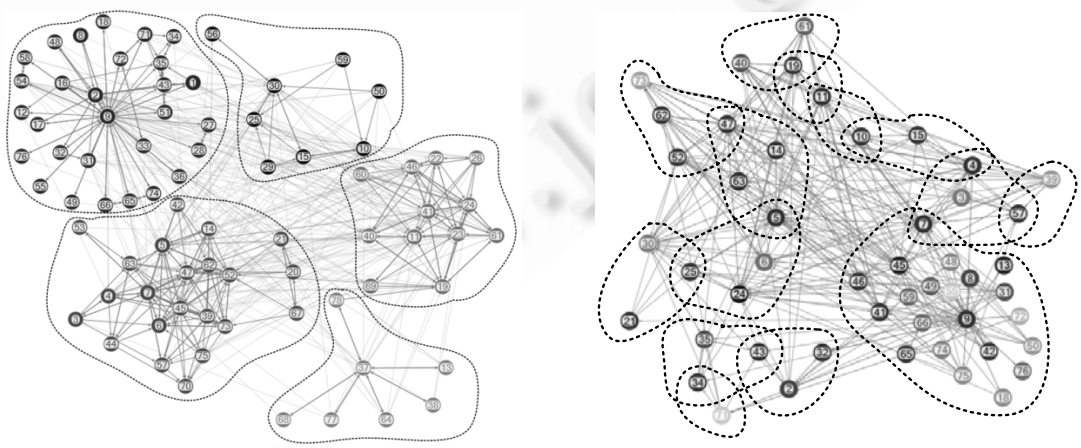
数据集	微博数据	DBLP
节点	78	179 035
边	1062	1 535 204
平均度数	13.61	8.57
最大度数	53	3 165
方向性	有方向	有方向

其中,

- 微博数据中信息传播者和接收者对应发布微博的节点和转发/评论微博的节点;
- DBLP 数据中,我们将引文文献解释为信息的流动,被引用的作者是信息的传播者,引用者是信息的接收者.

我们将从有效性和动态性两个方面对基于信息流动的社区发现结果进行分析,通过与文献[5,7]中的社区发现结果的对比,来验证本文的社区发现方法在信息流动频繁的社交网络中的有效性,动态性通过分析各个社区在某时间段内的变化进行说明.

首先,我们来检验基于信息流动速率的社区划分方法的有效性.在微博数据上进行实验,图 3(a)所示的是基于模块化的社区划分结果,即仅从节点间的拓扑结构的角度,不考虑信息流动.可以看出,仅依靠拓扑结构划分社区后,社区内部节点拓扑关系比较紧密,社区间拓扑关系比较松散.同样,对于微博数据,利用文献[5]中的算法,依靠节点间的交互信息进行社区划分,划分结果如图 3(b)所示(孤立点已省去).



(a) 基于模块度的社区发现[7] (b) 基于着色的动态社区发现[5]

Fig.3

图 3

通过对比图 3(a)和图 3(b)可以看出:考虑节点间信息交互后,划分的社区规模减小,节点间的信息流动影响到了社区的划分结果.例如,节点 24 和节点 25 在图 3(a)中分属不同社区,图 3(b)中则处于同一社区.

图 3(b)所示的社区划分方法在微博等带有信息流动社交网络分析中,能够在一定程度上根据节点间信息流动的记录产生合适的社区划分结果.但也存在不足,主要原因在于节点间频繁的信息流动增加了文献[5]中节点着色的复杂度;同时,每条微博数据记录的是两个节点间的交互情况,这也导致该方法划分社区的结果过于松散,所得到的社区规模普遍偏小.并且,该方法在划分社区时完全忽略了节点间的拓扑情况,而微博应用中某一特定时刻信息交互常发生于少量节点间,这样,大多数节点在该时刻都作为不参与交互活动的节点,这些都导致划分结果中孤立点较多.

基于信息流动速率的社区划分,则在利用节点信息流动记录的基础上兼顾了社区的拓扑结构.图 4 所示在 2010 年 12 月 5 日该算法的社区划分结果,其中,节点间信息流动速率大小通过节点间边的宽度表示.可以看出:信息流动频繁的节点都处于同一个社区,并且图 4 的社区划分结果相比图 3(a)中的社区划分结果,社区内的节点关系更为紧密,即在依靠拓扑结构所得到的社区发现的层级结构基础之上,划分的层次更为细致.例如,对于“时尚”相关信息,“娱乐综艺明星”(24,25,61)与“时尚服饰设计师”(40,41,60,46,45)分属不同社区;同时,又避免了产生如图 3(b)中社区规模偏小、孤立点较多的情况.

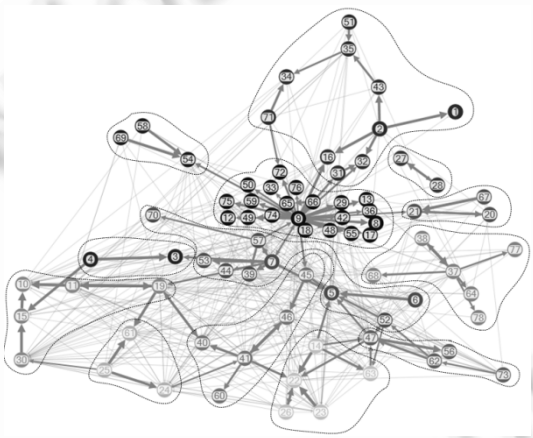


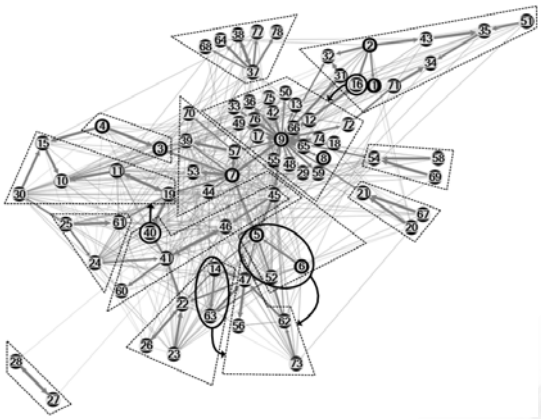
Fig.4 Dynamic community detection based on information flow

图 4 基于信息流动分析的动态社区发现

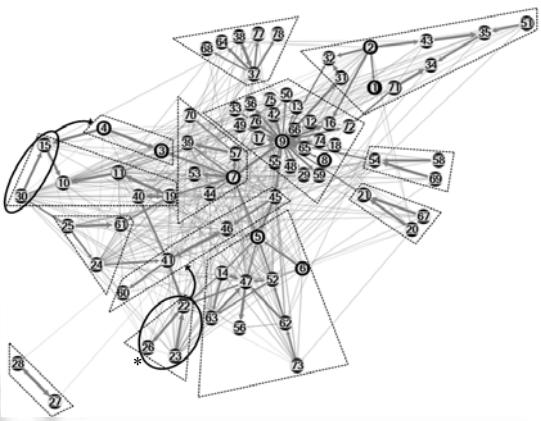
在社区发现的动态性方面,通过将各不同时间的社区划分结果进行比较,基于信息流动速率的社区划分方法能够反映出整个社交网络结构的变化过程.

图 5 所示的是微博数据中节点每隔 5 日的社区变化情况,其中,节点间信息流动速率通过边的宽度表示.我们将整个网络社区变化趋势也做了标记.

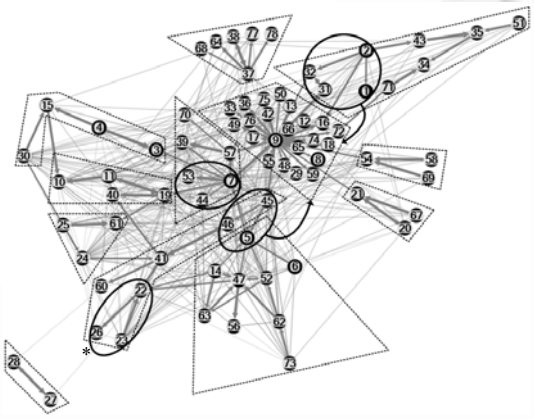
对图 5(a)~图 5(f)进行分析可以看出:社区的主体节点在一定时期内相对保持静止,整个社交网络结构变化包含了社区的分拆、合并等过程.例如:由节点(22,23,26)所构成的社区(用*号标记)在 10 日~15 日过程中并入社区(41,45,46,60),15 日~20 日过程中逐渐从该社区分拆出来,至 25 日又逐渐并入另一社区(14,47,52,56,62,63,73),25~30 日又与该社区分离,并入社区(5,6,41,45,46,60)中.又如:社区(7,39,44,53,57,70)在 15 日~20 日的信息流动过程中逐渐拆分为两个社区(39,57,70)和(7,44,53),20 日后,这两个社区又合并为原来的社区.通过分析还可以看出:部分社区虽然在拓扑结构上与其他社区相连,但是社区间却没有信息流动,信息仅在本社区内流动,因此这些社区的构成节点也保持稳定.例如:社区(37,38,64,68,77,78)、社区(54,58,69)和社区(20,21,67)等都是内部信息流动远大于社区间的信息流动,在 30 天内,这些社区的构成也保持不变.



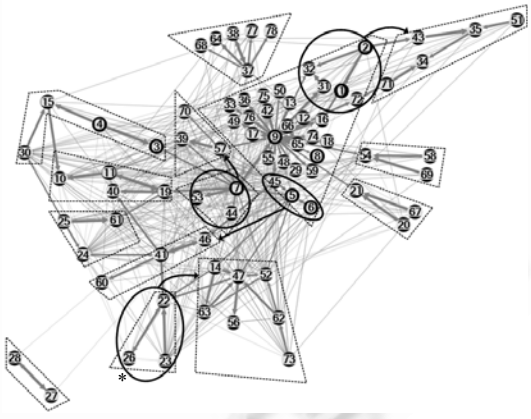
(a) 12 月 5 日~12 月 10 日社区变化情况



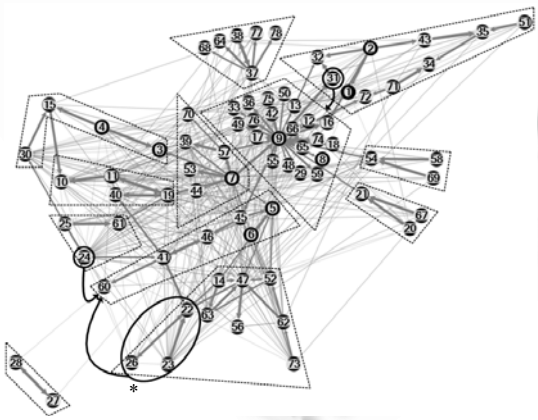
(b) 12 月 10 日~12 月 15 日社区变化情况



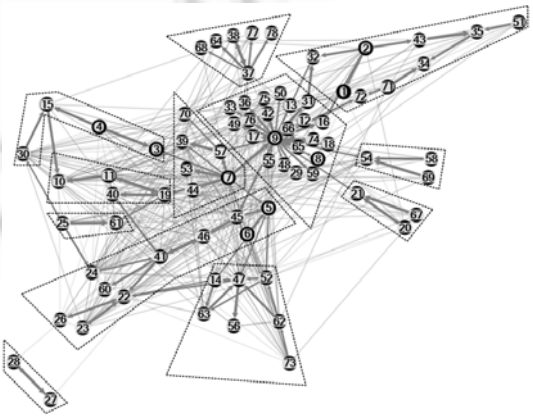
(c) 12 月 15 日~12 月 20 日社区变化情况



(d) 12 月 20 日~12 月 25 日社区变化情况



(e) 12 月 25 日~12 月 30 日社区变化情况



(f) 12 月 30 日社区划分结果

Fig.5 Analysis of community evolution on the micro-blog data

图 5 微博数据社区变化分析

在 DBLP 数据上利用本文的算法进行社区发现,对 2000 年~2005 年社区发现结果中规模大于 3 000 个节点

的社区进行分析,其中,社区构成发生较大变化的社区共有 7 个,具体社区描述见表 3.社区动态变化过程如图 6 所示,其中,每个节点表示一个社区,节点大小表示社区规模大小,边表示社区变化的方向,阴影节点为新出现社区.从图中可以看出,各个社区相对稳定,尽管在社区间存在节点的转移,但是这种转移主要发生在相关的社区之间.例如 2004 年~2005 年,社区 c4,c5 中部分节点转移至社区 c3,从 c4 社区转移至 c3 社区的节点主要是研究方向为编程语言的节点,节点自 c5 社区向 c3 社区的转移则反映了部分节点从数据库、数据挖掘向人工智能、软件工程的研究方向的转变.

Table 3 Communities description of DBLP

表 3 DBLP 数据社区描述

社区编号	社区描述
c1	Simulation; Electrical & electronic engineering; Manufacturing
c2	Algorithms & theory; Networks & communications
c3	AI; Software engineering; Programming languages; Mathematics
c4	Hardware & architecture; Operating systems; Programming languages; Databases
c5	Databases; DataMining
c6	Machine learning & pattern recognition; AI; DataMining
c7	Natural language & speech; Multimedia; Information retrieval; Machine learning & pattern recognition

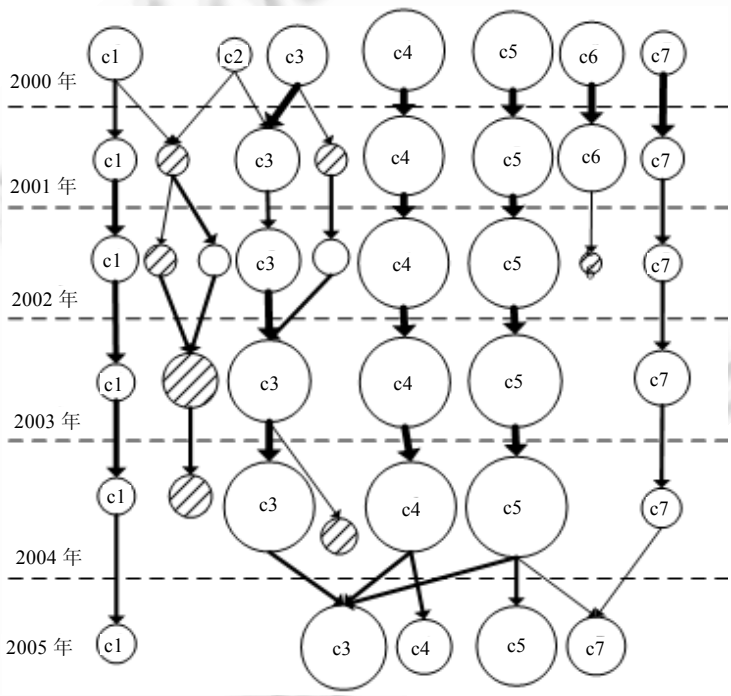


Fig.6 Analysis of community evolution on DBLP

图 6 DBLP 社区变化分析

4 相关工作

传统的社区发现方法通常利用网络的拓扑结构,采用基于介数[3]、图谱[7]、随机漫步[8]、或信息论[9]等方法实现.Newman 和 Girvan 最先引入模块性[3]的概念,起初只是用其来作为社区划分的终止条件,目前已经成为社区发现中最常用的质量函数,对社区发现影响深远,在社区发现界开启了新纪元[19].基于模块性的社区发现方法,就是寻找最优的 Q 值,使得这种差别最大化.由于寻找最优 Q 值的算法复杂度太高,很多方法都对 Q 值的计算过程进行了优化,转而寻求次优的 Q 值,常见的如贪心算法[1]、模拟退火[10]、极值优化[11]等.这些方法都去除

了节点交互的时间信息,将网络看作一个静态图处理.

近来,许多研究开始关注社区的动态变化.White[12]最早提出了在节点关系随时间变化的网络中进行社区发现.Leskovec[13]基于图的拓扑特性,例如大规模网络中的度分布、“小世界”特性,对网络的动态变化进行了研究,从理论上指出:这些网络的密集程度随时间呈指数性增长;同时,网络的有效直径逐渐减少.并且,据此提出了一种图生成模型.Backstrom[14]研究了社区的形成以及变化过程,提出通过决策树来估计节点加入两个社区的可能性,该方法还可以发现有增长趋势的社区.Kumar[15]通过分析两个真实网络的结构和变化过程,量化了社交网络所具有的特性.并基于这些特性提出了一种社交网络动态分析的模型.以上方法都是从图的特性方面对于网络的变化过程进行分析,并没有通过分析网络中社区的变化来解释网络的变化过程.

关于社区随时间的变化,Berger-Wolf 和 Saia[16]提出了一种追溯社区变化过程的框架.在他们随后的工作中,Tantipathananandh 等人[5]将动态社区发现的过程形式化为图着色问题,并利用动态规划和启发式算法优化了社区发现的结果.此后,他又将之前的模型扩展到任意的动态网络中去,并采用半定松弛度和循环启发式思维,近似去求解优化问题[20].Falkowski[17]利用不同时刻的社区划分情况,通过统计方法分析它们的重叠程度,找出稳定和变化的社区.虽然这些社区发现方法包含了对网络中节点交互信息的分析,但关注更多的是节点的变化过程,对于社区所反映的网络结构关注较少.我们所提出的动态社区发现方法将传染病动力学理论应用于节点关系的分析中,量化了节点间关系的动态变化;同时,在社区划分的过程中利用传统的社区发现方法,使得社区发现的结果包含了更多的社交网络结构性信息.此外,Zhou 等人[21]提出了考虑了社区信息内容的模型 COCOMP,使社区发现更加高效.

5 结论和下一步工作

本文针对传统社区发现方法忽略社交网络动态特性的不足,利用传染病动力学理论提出基于节点间信息流动的动态社区分析方法.通过分析信息在社交网络中的流动过程,将节点间密切程度量化为平均接触率,有效地反映节点间关系随时间的动态变化,为从社区的角度解释社交网络结构的变化提供了一种有效的途径.通过实验验证了新的动态社区发现方法相比传统的社区发现方法的合理性、有效性.此外,对于已得到的动态社区变化,如何解释这一变化的原因,以及社区动态变化对应的未来趋势预测、分析等算法,还有待进一步研究.

References:

[1] Caluset A, Newman MEJ, Moore C. Finding community structure in very large networks. Physical Review E, 2004,70(6):066111. [doi: 10.1103/PhysRevE.70.066111]

[2] The guardian. http://www.guardian.co.uk/technology/blog/2009/jun/29/twitter-users-average-api-traffic

[3] Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Physical Review E, 2004,69(2):026113. [doi: 10.1103/PhysRevE.69.026113]

[4] Tang J, Sun J, Wang C, Yang Z. Social influence analysis in large-scale networks. In: Proc. of the 15th ACM SIGKDD Int’l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2009. [doi: 10.1145/1557019.1557108]

[5] Tantipathananandh C, Berger-Wolf T, Kempe D. A framework for community identification in dynamic social networks. In: Proc. of the 13th ACM SIGKDD Int’l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2007.

[6] Freeman LC. Centrality in social networks conceptual clarification. Social Networks, 1979,1(3):215−239.

[7] Newman MEJ. Finding community structure in networks using the eigenvectors of matrices. Physical Review E, 2006,74(3): 036104. [doi: 10.1103/PhysRevE.74.036104]

[8] Pons P, Latapy M. Computing communities in large networks using random walks. In: Proc. of the Computer and Information Sciences (ISCIS 2005). Berlin, Heidelberg: Springer-Verlag, 2005. 284−293.

[9] Fortunato S, Latora V, Marchiori M. Method to find community structures based on information centrality. Physical Review E, 2004, 70(5):056104. [doi: 10.1103/PhysRevE.70.056104]

[10] Reichardt J, Bornholdt S. Statistical mechanics of community detection. Physical Review E, 2006,74(1):016110. [doi: 10.1103/PhysRevE.74.016110]

[11] Duch J, Arenas A. Community detection in complex networks using extremal optimization. Physical Review E, 2005,72(2): 027104. [doi: 10.1103/PhysRevE.72.027104]

[12] White HC, Boorman SA, Breiger RL. Social structure from multiple networks. Part I: Blockmodels of roles and positions. American Journal of Sociology, 1976,730−780.

[13] Leskovec J, Kleinberg J, Faloutsos C. Graphs over time: Densification laws, shrinking diameters and possible explanations. In: Proc. of the 11th ACM SIGKDD Int'l Conf. on Knowledge Discovery in Data Mining. ACM Press, 2005.

[14] Backstrom L, Huttenlocher D, Kleinberg J, Lan XY. Group formation in large social networks: Membership, growth, and evolution. In: Proc. of the 12th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2006. [doi: 10.1145/1150402.1150412]

[15] Ravi K, Novak J, Tomkins A. Structure and evolution of online social networks. In: Proc. of the Link Mining: Models, Algorithms, and Applications. New York: Springer-Verlag, 2010. 337−357.

[16] Berger-Wolf TY, Saia J. A framework for analysis of dynamic social networks. In: Proc. of the 12th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2006. [doi: 10.1145/1150402.1150462]

[17] Tanja F, Bartelheimer J, Spiliopoulou M. Mining and visualizing the evolution of subgroups in social networks. In: Proc. of the 2006 IEEE/WIC/ACM Int'l Conf. on Web Intelligence. IEEE Computer Society, 2006. [doi: 10.1109/WI.2006.118]

[18] 马知恩,周义仓,王稳地,靳祯.传染病动力学的数学建模与研究.北京:科学出版社,2004.

[19] Santo F. Community detection in graphs. Physics Reports, 2010,486(3):75−174.

[20] Chayant T, Berger-Wolf TY. Finding communities in dynamic social networks. In: Proc. of the 2011 IEEE 11th Int'l Conf. on Data Mining (ICDM). IEEE, 2011. [doi: 10.1109/ICDM.2011.67]

[21] Zhou WJ, Jin HX, Liu Y. Community discovery and profiling with social messages. In: Proc. of the 18th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2012. [doi: 10.1145/2339530.2339593]

索勃(1987－),男,辽宁锦州人,博士生,主要研究领域为图数据管理,数据挖掘.
E-mail: caitou@mail.nwpu.edu.cn

陈群(1976－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为云计算,RFID 数据管理, XML 数据管理.
E-mail: chenbenben@nwpu.edu.cn

李战怀(1961－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为数据库理论与技术.
E-mail: lizhh@nwpu.edu.cn

王忠(1989－),男,硕士,主要研究领域为海量数据处理,图数据挖掘.
E-mail: Zhongwang.cs.npu@gmail.com