ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.20, Supplement, December 2009, pp.184−192

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

IC:动态社会关系网络社区结构的增量识别算法*

单 波+, 姜守旭, 张 硕, 高 宏, 李建中

(哈尔滨工业大学 计算机科学与技术学院,黑龙江 哈尔滨 150001)

IC: Incremental Algorithm for Community Identification in Dynamic Social Networks

SHAN Bo+, JIANG Shou-Xu, ZHANG Shuo, GAO Hong, LI Jian-Zhong

(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

+ Corresponding author: E-mail: bobactive@163.com

Hou MB, Xu QL, Guo SQ. IC: Incremental algorithm for community identification in dynamic social networks. *Journal of Software*, 2009,20(Suppl.):184−192. http://www.jos.org.cn/1000-9825/09022.htm

Abstract: Community structure in social networks (SNS) could provide interesting information, such as the pattern of social activities between individuals and the trend of social development. Traditional methods to identify communities on static social networks will miss interesting laws how SNS change. The few methods on modeling and analyzing of community structures in dynamic social networks, which are obtaining more and more attention recently, fail to identify large networks in acceptable time. This paper proposes an incremental new method to identify community structure in dynamic social networks. Utilizing the time locality that there's little change in adjacent network snapshots, the paper incrementally analyzes social networks to avoid repeatedly partitioning the whole networks. Experiments demonstrate that this approach offers orders-of-magnitude performance improvement over state-of-the-art approaches on large scale networks (105 nodes) and can produce nice community structures which reflect the essence of SNS.

Key words: dynamic social network; community identification; incremental algorithm

摘 要: 社会关系网络(SNS)中社区结构的识别有助于得出有意义的个体间活动模式和社会发展规律,传统的静态SNS社区结构识别的方法不能发现SNS的变化规律,而最近受到广泛关注的动态SNS社区识别方法普遍存在可扩展性差的缺点.描述了动态SNS的数学模型,并在此基础上提出了动态SNS中发现社区结构的增量式新方法.提出方法利用动态网络时间局部性即相邻采样时刻网络变化不大的特点,通过增量分析避免对整个网络中的个体全部重新划分,达到较高的算法效率.分析和实验结果表明,效率高于现有方法,在大规模网络上(10^5结点量级)效率提升在一个数量级以上,发现的社区结构很好地反映出社会关系网络的本质结构.

关键词: 动态社会关系网络;社区识别;增量算法

* Supported by the National Basic Research Program of China under Grant No.2006CB303000 (国家重点基础研究发展计划(973)); the National Natural Science Foundation of China under Grant Nos.60703012, 60533110, 60773063, 60903017 (国家自然科学基金); the Heilongjiang Province Fund for Young Scholars of China under Grant No.QC06C033 (黑龙江省青年科技专项资金); the NSFC-RGC of China under Grant No.60831160525 (NSFC/RGC联合科研基金)

Received 2009-05-01; Accepted 2009-07-20

1 引 言

1.1 应用背景

对社会关系网络(SNS)的研究最初源自社会学领域,随着计算机科学和网络技术的发展尤其是Web 2.0 的流行和普及,这项研究在计算机科学领域也逐渐流行起来[1−10].在SNS的相关研究中,图是非常重要的建模工具,把被研究个体抽象成顶点、把个体之间的联系抽象成边就构成了图结构.通过对这种特定的图的研究可以分析和挖掘SNS内部所隐含的模式与信息,其应用范围遍及社会学、人类行为学、信息传递和疾病传播学以及互联网在线社区.

SNS 研究中,一个基本而又非常重要的问题是社区结构识别.SNS 的社区结构可以直观地理解为:将 SNS 中的所有个体划分为若干互不相交的子集合,集合内部的个体联系比较紧密,而不同集合之间的个体联系比较松散, 每个这样的集合就是一个社区.一个社区内部的成员往往具有相似的兴趣、爱好,甚至具有相似的行为特点,但并非仅仅如此,因为社区成员之间相互影响、相互合作,在很多应用背景中以一个具有特定功能的整体而存在.社区结构刻画了 SNS 的组织性和团体性,它的准确识别,对于整个 SNS 的宏观或微观的行为模式的进一步研究,具有重大的意义.

在绝大多数社会关系网络的实际应用中,SNS 中个体的兴趣和社会角色会随时间发生改变,个体之间的联系也可能会发生变化.此类应用环境中,这样的问题非常常见:某种社会模式的前因和后果是什么?流行性疾病的传播速度有多快?某个社区兴趣主题在发生怎样的转移?社区的成员数目在怎么变化,以后会怎样?诸如此类问题,静态 SNS 的模型是回答不了的.为了解释这样的问题和现象,SNS 分析中必须考虑时间因素,建立起 SNS 的动态数学模型,并进行动态 SNS 社区结构的识别和分析.

1.2 相关工作

传统的SNS研究,通常把社会关系网络建模为一个静态的图,这个图一般是通过对所有时间的社会关系网络聚集或者抽取某一个时刻的社会关系网络快照得到的.Newman等人提出了基于betweeness[4,6]的分裂式层次方法,同时给出了衡量社区结构质量好坏的测度modularity Q[4,5],为这一问题的研究指明了方向.在此基础上,有许多研究者提出了新的方法改进了SNS社区分析的运行效率或者准确度,比如Guimera等人[7,8],Clauset等人[9].

然而,上述研究的一个主要局限是它们都认为SNS是固定、一成不变的,这一点与常见的社会关系网络真实情况并不一致.比如在图 1 中,两个网络在 3 个不同采样时刻发生网络拓扑结构的变化,如果通过聚合的方法,就都会得到G_1这个静态的图,每条边的权重都为 1,表示 3 个顶点之间的关系同等重要.可是分析图 1(a)会发现,在网络变化过程中,顶点A和B具有相同的行为,它们要么互相连接,要么都与C顶点连接.A与B的这一共性在图 1(b)中却不存在.从这个例子可以看到,实

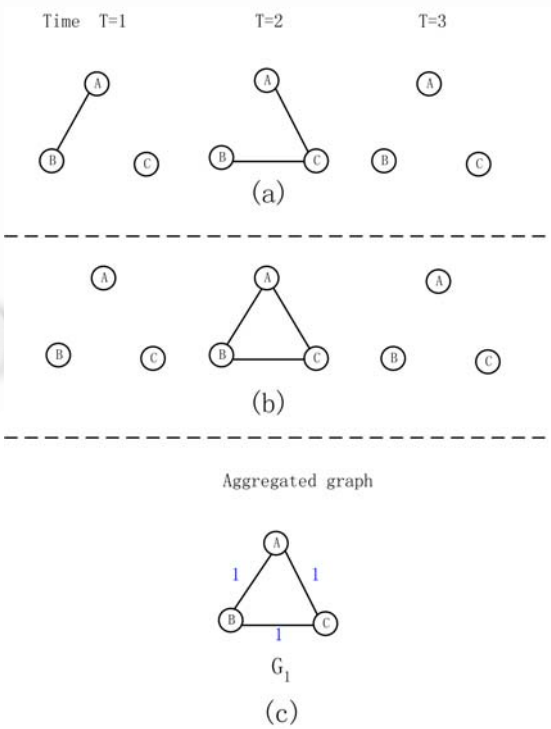


Fig.1 Two dynamic networks and the aggregated network
图 1 两个动态网络和它们的聚集网络

际相差很大的两个网络,通过聚合得到的结果图可能是一样的,如果仅仅用聚合得到的图来分析网络,就会丢失一些在网络变化过程中所包含的重要信息.

把动态变化的SNS在不同的时刻抽样,在每个时刻得到网络的一个拓扑图,按照时间顺序把这些拓扑图排列起来,构成一个图序列,这是动态SNS社区研究常用的建模方法[10−14].

动态 SNS 的社区结构就可以直观的理解为:

(1) 在某时刻,将所有个体划分成若干互不相交的子集合,集合内部个体联系比较紧密,而不同集合之间的个体联系比较松散.这个划分就是一个社区结构,一个子集合是一个社区.

(2) 不同时刻的社区结构变化不大.

基于动态变化观点研究SNS、分析社区结构的工作在最近三四年才开始,模型构建和算法效率都不是很理想,还有很多的工作要做.Berger-Wolf和Saia[10]给出了分析动态社会关系网络的基本框架,第 1 次提出了meta-group概念用以描述动态SNS的社区结构.Tantipathananandh等人[11]在文献[10]研究基础上,提出了用动态规划的方法优化目标代价函数,在小组和个体两个层次上,分两阶段实现社区结构划分.Tantipathananandh等人方法的主要特点是目标代价函数包含了所有时间内个体的社区归属,而不着重考虑相邻时刻的社区结构的局部性,其形式化的问题是一个NP完全问题.Zhou等人[12]提出了带文档的动态社区识别方法,第 1 次明确了基于网络拓扑结构和历史社区先验信息的动态SNS社区结构识别框架,对宏观层次上的方法选择具有重大意义.Lin等人[13]不再采用分两层次两阶段的社区识别策略,而是利用图上的混合模型[14],将两个阶段一次性的结合了起来,对于稀疏图其算法效率最高,复杂度为$\boldsymbol{O}(n^2k)$.

然而,在保持高准确度的同时,算法效率是社区识别最关心的要素之一.Danon等人[14]指出,当SNS中成员个数达到 10^4~10^5规模的时候,Newman等人[4]的算法会运行数小时甚至几天的时间,这是用户无法接受的.文献[10−13]等方法在处理大规模数据的时候,也遇到了可扩展性差的问题.

为了解决以上问题,本文提出了增量式地识别动态SNS中社区结构的方法.真实数据统计发现,现实中的SNS变化都是很缓慢的,合适的相邻采样时刻之间网络变化很小.对于非常相似的图反复的进行顶点划分,会有很多的重复操作.如何合理有效地利用过去已经得到的社区结构来指导当前的网络社区划分,是本文主要考虑的问题.本文采用增量分析的方法,重点关注两个相邻时刻之间图的变化差量,避免了对整个网络中的个体全部重新进行划分.分析得出,提出算法的时间复杂度比已有最新方法[13]降低了一个数量级.实验结果表明,本文所采用的方法效率很高,找到的社区结果很好的反映出SNS的真实结构.

2 问题描述

2.1 静态SNS的数学模型

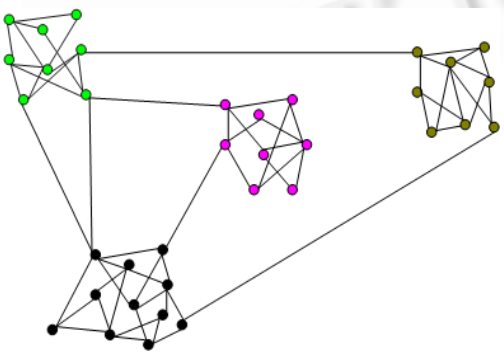


Fig.2 Graph mode of a social network
图 2 一个社会关系网络的图模型

用图 G 来描述 SNS,图的顶点对应 SNS 中的人或其他个体,比如在线社区 Facebook 中的注册会员.如果 SNS 中两个人之间有联系,那么在图中对应的两个顶点之间就存在一条边.利用高度发展的图的理论,再加上数据挖掘的手段,可以识别 SNS 社区.图 2 给出了一个 SNS 的例子,顶点代表人,边代表两个人之间的联系.这个网络被划分成 4 个社区,分别用 4 种颜色来表示.

我们用无向图 $G=(V,E)$来描述静态的 SNS.其中,

(1) V 是顶点集合.顶点 $v\in V$,表示 SNS 中有人与 v 相对应.

(2) E是边集合.边$e=(v_1,v_2)\in E$, 表示分别与v_1和v_2相对应的人有相互的联系.

定义 1. 静态SNS的社区结构CS是顶点集合V的一个分划$P=(C_1,C_2,C_3,\ldots,C_k)$,而且对任意一个集合$C_i$,要求满足:

(1) C_i内部的顶点之间连接紧密.即集合内部边比较稠密.

(2) C_i与C_j两个集合($i\neq j$)的顶点之间连接松散.即集合间的边比较稀疏.

其中,一个集合C_i就被称作一个社区,k是社区的个数,通过先验知识得到.

需要指明的是,在静态SNS社区结构的定义中,顶点之间联系的紧密与松散是一个相对的概念.在文献[1,6,7,11,12]等研究工作中,都没有给出一个统一的数学描述.紧密、松散程度可以有多种衡量测度,常用的测度有凝聚度[16] $\sum_i\left|\{(u,v)|u,v\in C_i\}\right|$和分离度$\sum_{i,j}\left|\{(u,v)|u\in C_i,v\in C_j\}\right|$,这也是本文所用度量,用凝聚度与分离度比值判断分划P是否满足社区结构定义.

在文献[4]中,Newmam等人提出了判断社区结果好坏的测度Modularity Q,定义一个$n\times n$矩阵$\mathbf{e}$,元素e_{ij}表示两个顶点分别在社区C_i和社区C_j的边占全部边的比例,得到:

$$Q=\sum_i(e_{ii}-a_i^2) \tag{1}$$

Newman 等人指出,如果社区结构的划分是随机、任意的,那么 $e_{ii}=a_i^2$,Q=0,表明这个社区结构很差.而一个比较大的 Q 值代表非常好的社区结构.

同一个社区中的顶点往往具有一些共同的属性,而且并不仅仅如此,很多的应用背景中顶点间相互作用构成一个功能模块整体.比如,在线博客社区会有一些潜在的主题;在生物信息学中,一个社区会构成基本的功能单位.社区结构的识别,对于主题或模块功能层次上的研究,具有很现实的意义.

在真实的应用环境中,SNS总是会随时间推移而改变的[10−13],网络的社区结构也有可能跟着发生改变.如何为动态SNS建立模型并分析其社区结构,成为此时必须解决的一个问题.

2.2 动态SNS的数学模型

由于动态SNS总是在变化的,有些时候某两人之间有联系,而到了另外的一些时候,这两个人可能就停止联系了.为了方便刻画动态SNS的数学模型,需要先对动态SNS进行采样,在不同时刻得到一个静态SNS的所对应的无向图,我们把这个无向图称作动态SNS在这个时刻的网络快照.比如,在第 1 个时刻得到快照G_1,在第 2 个时刻得到快照G_2,依次类推,得到G_{T-1},G_T等这些快照.

用一个图序列$G_1,G_2,\ldots,G_{T-1},G_T$表示从 1 时刻到第$T$个时刻的动态SNS,其中$G_t=(V_t,E_t)$是动态SNS在第$t$时刻的快照,$G_{t-1}$是动态SNS第$t-1$ 时刻的快照,t=1,2,…,T−1,T.图序列是常用的描述动态SNS的方法[12,13].

定义 2. 动态SNS在t时刻社区结构$\boldsymbol{CS_t}$是顶点集合V的一个分划$P_t=(C_{t,1},C_{t,2},C_{t,3},\ldots,C_{t,k})$,如果:

(1) CS_t满足定义 1,对任意i>=1.

(2) CS_t与CS_{t-1}变化不大.对于给定常数∂,满足:

$$\frac{|\Delta C_t|}{|C_{t-1,i}|}<\partial\ ,\ \forall\, t\text{>=}2 \tag{2}$$

其中,$\Delta C_{t,i}=\left(C_{t,i}\setminus C_{t-1,i}\right)\cup\left(C_{t-1,i}\setminus C_{t,i}\right)$,表示集合$C_{t-1,i}$与$C_{t,i}$之间的差量.

当t=1 时CS_t,即CS_1,只要求满足定义 1 就可以了.CS_1是通过静态SNS的社区划分得到的.

Lin等人[13]指出,SNS的变化实际上是非常缓慢的.通过对Enron[17]等大量的真实数据分析统计我们也发现了发现这一特点.即,在相邻的$t-1$ 时刻和t时刻,SNS的拓扑结构变化相对于整个图来讲非常小,即:G_t与G_{t-1}差别不大.

定义 3. 社会一个动态关系网络从G_{t-1}到G_t变化是缓慢的,如果满足:

$$\frac{|\Delta E_t|}{|E_t|}<\alpha \tag{3}$$

和

$$\frac{|\Delta V_t|}{|V_t|} < \beta \tag{4}$$

其中,$\triangle E_t=(E_t\backslash E_{t-1})\cup(E_{t-1}\backslash E_t)$和$\triangle V_t=(V_t\backslash V_{t-1})\cup(V_{t-1}\backslash V_t)$是$G_t$与$G_{t-1}$的差量,$\alpha$ 与 β 是两个先验参数,与应用背景有关.

2.3 问题的形式化定义

利用静态 SNS 和动态 SNS 中社区结构的定义,问题的形式化表示:

已知: t–1 时刻的社区结构CS_{t-1}和当前t时刻的网络拓扑图G_t

求: t时刻的社区结构CS_t,使得CS_t满足定义 2.

3 IC 算法的提出

3.1 算法基本思想

通过定义 2 可看出,在求 t 时刻 SNS 社区结构的时候,主要考虑两个方面:一,相邻时刻之间的拓扑结构变化不会太大,即:要参考 t–1 时刻的历史社区结构信息;二,要符合当前 t 时刻 SNS 拓扑图.简而言之,就是历史加拓扑,决定当前的社区结构.

令C_u和C_v分别表示在t–1 时刻顶点u和v所在的社区集合.令$e^+=(u,v)\in(E_t\backslash E_{t-1})$,表示$t$–1 时刻不存在,而在$t$时刻产生的边.令$e^-=(u,v)\in(E_{t-1}\backslash E_t)$,表示$t$–1 时刻存在,而$t$时刻消失了的边.$e^0$表示在$t$–1 和$t$时刻都存在的边,即$e^0\in E_t$且$e^0\in E_{t-1}$.

定义 4. ***IVt*** 表示 t 时刻增量相关顶点集合,即:

IVt={$v|e^+=(u,v)$且C_u和C_v是不同集合

或$e^-=(u,v)$且C_u和C_v是同一集合

或v是v^+,即新加入的顶点

}

用 d_v 表示顶点 v 的度,即与顶点 v 相连的边的个数.

定义 5. 顶点v对C_i的社区依存度 $aff_{v,i}$:

$$aff_{v,i}=\frac{\left|\left\{u\middle|u\in C_{t-1,i}\ \text{ and }\ (u,v)\in E_t\right\}\right|}{d_v} \tag{5}$$

社区依存度 $aff_{v,i}$ 表示顶点对社区依赖程度,即顶点 v 与社区 i 内顶点之间的边占所有与 v 相连的边的比例.

定义 6. 称一个顶点v在t时刻改变社区归属,如果在t–i时刻$v\in C_{t-1,i}$,且在t时刻$v\in C_{t,j}$,但$i\neq j$.

顶点 v 改变社区归属的条件是:

$$\frac{aff_{v,j}-aff_{v,i}}{aff_{v,i}}>\varepsilon \tag{6}$$

其中,ε是$aff_{v,i}$变化比例阈值,取值由用户给出.ε是衡量CS_{t-1}在演变成CS_t过程中所起作用的参数.ε值越大,社区结构越不易改变,表明历史社区结构信息在新的社区结构获得过程中起较大作用;ε值小,表明G_t即当前拓扑结构所起作用比较大.

在t–1 时刻的社区结构CS_{t-1}的基础上,新加入的边的两端顶点如果都处于同一社区C_i内,会使凝聚度增大,反之,如果新加入的边处于两个社区之间,即两端顶点分别处于不同的社区,就会使分离度增大,破坏t–1 时刻的社区结构CS_{t-1}.动态SNS随时间变化的局部性特征,是增量分析社区结构的立足点.

命题 1(社区增量定理). 在动态 SNS 社区结构中,如果顶点 v 在 t 时刻改变了社区归属,那么 v 是增量相关顶点, 即 $v\in IVt$.

证明:假设∃顶点 v 改变了社区归属,但 $v\notin$ IVt.那么与顶点 v 相连的边(u,v)只可能是以下 3 种之一:

(1) $(u,v)=e^0$:

e^0在$t-1$ 和t时刻都存在,边的存在对 $\frac{aff_{v,j}-aff_{v,i}}{aff_{v,i}}$ 没有任何贡献;

(2) $(u,v)=e^+$且C_u和C_v是相同集合:

t 时刻加入(u,v)边会使得 $aff_{v,i}$ 增大;在 $aff_{v,j}$ 不变的情况下,$\frac{aff_{v,j}-aff_{v,i}}{aff_{v,i}}$ 会减小;

(3) $(u,v)=e^-$且C_u和C_v是不同集合:

t时刻减少边(u,v)使 $aff_{v,j}$ 减小,其中$u\in C_{t-1,j}$, 那么在 $aff_{v,i}$ 不变的情况下 $\frac{aff_{v,j}-aff_{v,i}}{aff_{v,i}}$ 值会减小;

因为 $t-1$ 时刻 v 属于社区 i,故 $t-1$ 时刻 $\frac{aff_{v,j}-aff_{v,i}}{aff_{v,i}}<\varepsilon$ 对任意的 j 成立,而(u,v)的3种情况下 $\frac{aff_{v,j}-aff_{v,i}}{aff_{v,i}}$ 都会不会增大;因此,在 t 时刻,$\frac{aff_{v,j}-aff_{v,i}}{aff_{v,i}}<\varepsilon$.这不满足顶点 v 改变社区归属的条件.而在假设中,顶点 v 改变了社区归属.因此,假设不成立. □

3.2 IC算法

算法思路:

(1) t 时刻大部分顶点社群归属都与 $t-1$ 时刻相同,只是有很少的部分顶点需要改变.

(2) 在CS_{t-1}的基础上,找到增量相关顶点集IVt中那些满足改变社区归属条件的顶点,划归到新社区中去.

(3) 其余的顶点维持其原来的社区归属.

输入:$t-1$ 时刻社区结构$CS_{t-1}=\{C_{t-1,i}\},i=1,2,\ldots,k$

t时刻SNS拓扑图G_t;

输出: t时刻社区结构$CS_t=\{C_{t,i}\},i=1,2,\ldots,k$

算法 1(IC 算法).

① {对 $\forall i\in\{1,2,3,\ldots,k\}$

② {对 $\forall v\in C_{t-1,i}$

③ if($v\in IVt$)

④ {求 j,使得 $aff_{v,j}$ 最大,$j=1,2,3,\ldots,k$

⑤ if($i\neq j$ 且 $\frac{aff_{v,j}-aff_{v,i}}{aff_{v,i}}>\varepsilon$)

⑥ 把 v 划归社区 j;

else

⑦ v 所属社区不变;

}

}

}

3.3 算法时间复杂度分析

算法的第 1 和第 2 行虽然对所有的社区和所有的顶点都进行扫描,但第 4~7 行的具体操作,是以第 3 行为判断条件的;如果条件不满足,就不会有第 4~7 行的操作.因此,运算中真正进行操作的遍数是 IVt 的集合大小 d,也就是所有与图差量有关的顶点的数目.根据定义 3,实际的 SNS 的变化是缓慢的.d 的大小与数据的采样周期有关,通过统计发现,真实社会关系网络的 d 的规模为 $O(n)$.

在每次循环中,第 4 行有 $O(k)$个单元操作,第 5、6 行有 $O(1)$个单元操作.因此每一遍需要 $O(k)$时间.

综合起来,总的运行时间 $O(d\times k)$.由于 d 的规模为 $O(n)$,IC 算法的总的时间复杂度为 $O(n\times k)$.

4 实 验

4.1 实验设计

实验环境为 Pentium(R)4 2.93GHz CPU, 512MB DRAM, 80GB SCSI HD, WinXP.

采用公开的安然公司邮件数据作为实验数据.安然公司邮件是 SNS 研究中重要的真实数据,它可在 CMU 计算机学院网站(http://www-2.cs.cmu.edu/~enron/)下载.用顶点表示邮件联系人,如果两个联系人之间有信件联系,就在对应顶点之间建立一条无向边.从 2000 年 4 月 1 号开始直到 2002 年 4 月 1 号,采样周期为三个月,每周期得到一个网络快照图,总共有 8 个图,由此构成一个跨越 8 个采样周期的动态 SNS.

定义 7. 动态 SNS 的社区结构稳定性测度 ***S*** 是衡量社区结构变化大小的度量.即:

$$S=1-\frac{\sum_{t=2}^{T}\sum_{i=1}^{K}\triangle C_{t,i}}{(T-1)n},$$

其中,$\sum_{t=2}^{T}\sum_{i=1}^{K}\triangle C_{t,i}\Big/(T-1)$表示“改变社区号的顶点个数”的平均值.

首先进行参数选择的实验.用户参数ε是顶点改变社区归属条件的参数,对算法结果具有重要的影响.通过观察ε对平均 modularity Q 和稳定性测度 S 的影响,来进行ε取值的选择.

然后进行对比实验.在相同的实验环境下,把IC算法与Tantipathananandh等人[11]和Lin等人[13]的算法做对比实验,为简便起见,他们的算法分别简称为Framework和FacetNet.

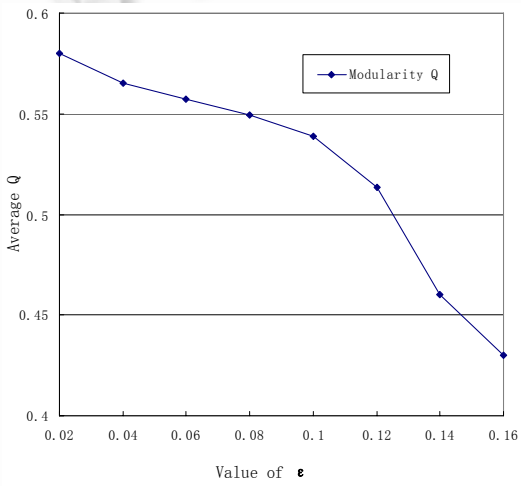


Fig.3 Average Q with the change of ε

图 3 Q 的平均值随ε的变化情况

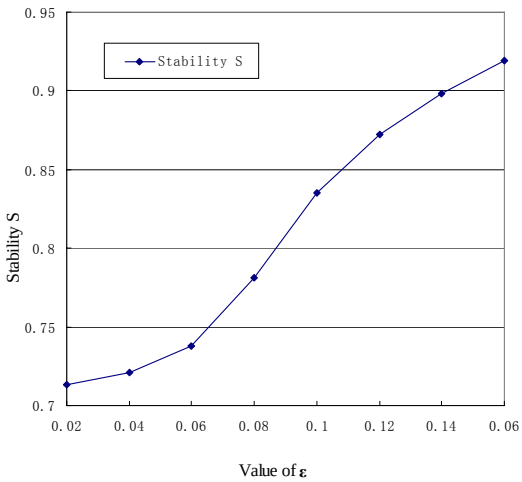


Fig.4 Value of S with the change of ε

图 4 S 的平均值随ε的变化情况

4.2 参数选择

通过分析不同的ε值对算法结果的影响.

首先要观察的是ε的变化对modularity Q[11]测度的影响.对每一个ε的值,分别求出 8 个采样周期的社区结构的modularity Q的均值.

从图 3 可以看出,随着ε值的增大,Q 均值下降,表明阈值ε越高社区结构就越不符合网络当前的拓扑结构.反言之,ε的值越小,Q 均值越高,当前社区结构与网络拓扑就越一致.

其次要分析ε的变化对社区结构稳定性的影响.

从图 4 可以看出,ε值的增大会使得社区结构的稳定性提高,这符合顶点改变社区归属的条件的定义.

综合起来,为了使得到的社区结构既不过分的违背当前网络的拓扑又不过分的缺乏稳定性,ε的值应该取在

一个比较折衷的范围内.

后面的实验中,设定阈值ε=0.1,以使社区结构具有较好的社区特征(见定义 1)和较高的稳定性(见定义 7).

4.3 对比分析

首先对比测试的是算法的时间性能随网络规模变化的情况.从上面的动态SNS中抽取不同规模的部分顶点构成不同大小的网络,顶点数目分别为 12,118,925,10233,91479,分别对应 10^1,10^2,10^3,10^4,10^5五个数量级.

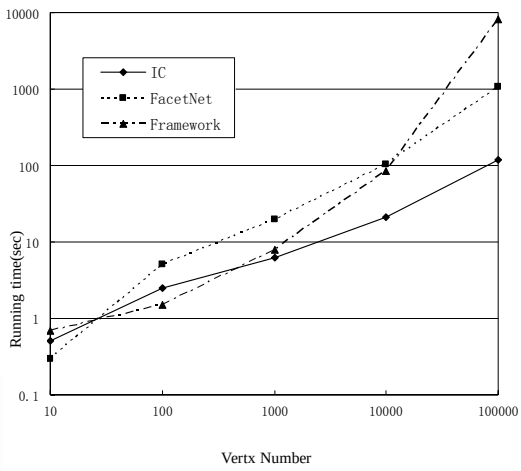


Fig.5 Running time with networks size

图 5 不同规模网络的算法运行时间

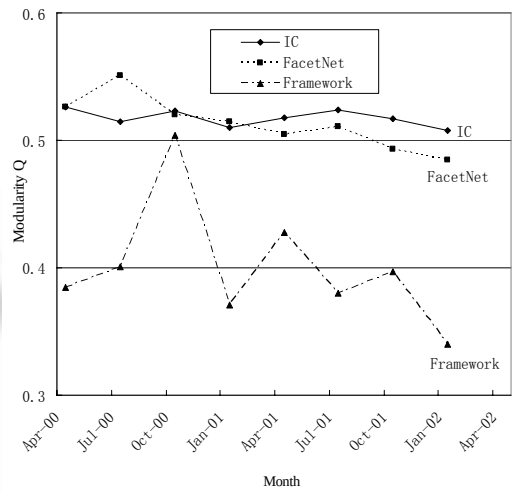


Fig.6 Q of three algorithms' results

图 6 3 种算法结果的 modularity Q

从图 5 中可以看出,当网络规模n<=10^3的时候,3 种算法的运行时间差不多,FacetNet算法的运行时间相对长一些;但当网络规模接近 10^5的时候,Framework算法的运行时间呈几何倍数增长.比如,在顶点数为 91479 的时候,Framework的运行时间已经达到了 8137s,也就是两个多小时,FacetNet的运行时间达到 17m,而此时IC算法才2m.可见,IC算法的运行时间远低于Framework和FacetNet的运行时间,具有很好的可扩展性,适合处理大规模数据.

其次,要评价实验得到的社区结构.Modularity Q[4]这一测度一直都被SNS的研究者所采用,本文也使用Q来衡量社区结构质量的高低.

图 6 的横坐标是 8 个采样周期的起始日期,纵坐标是 3 种算法各自得到的社区结构的 Q 测度值.可以看出.本文的 IC 算法与 FacetNet 算法的 Q 值相差不多,IC 的稍微好一点,但二者都比 Framwork 的要好很多.另外图 6 中 Framework 算法的曲线在第 2~第 5 个采样周期波动很大,在实际中,安然公司长期积累的信誉与经济矛盾在这个时期达到顶峰,因而危机爆发前公司邮件表现出骚动和不稳定,在 Framework 的曲线中表现为受噪声干扰而产生的强烈波动.IC 和 FacetNet 曲线波动小,具有较好的抗噪声干扰能力,表现出不错的稳定性.

综合可知,IC 算法在保证得到比较不错的社区结构的同时,运行时间远远小于其他两个算法.

5 总结与未来工作

本文提出一种动态社会关系网络上社区识别的增量式新方法.在给出静态和动态 SNS 的数学模型之后,分别形式化的定义了这两种模型之中的社区结构;基于连续两个采样时刻的网络图之间变化很小的特点,通过分析变化差量在社区识别中的重要作用,提出了动态 SNS 中社区识别的增量式算法 IC 算法.分析结果指出 IC 算法的时间复杂度比现有最新方法降低一个数量级;在真实数据上的对比实验表明,IC 算法发现的社区结构是有意义的.

识别出社区结构之后,就可以把社区作为基本单位来更进一步的研究和分析 SNS 的动态模式和演变规律.

另外,具有动态 SNS 特征的 Web 空间网络规模巨大、数据资源丰富,本文提出的社区结构识别方法将有助于发现其上新的有意义的知识,我们将重点关注这一领域.

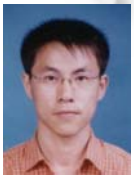
致谢 在此,我们向对本文的工作给予支持和建议的同行,尤其是李建中教授领导的哈工大数据与知识工程研究中心的同学和老师们表示最诚挚的感谢.

References:

[1] Girvan M, Newman MEJ. Community structure in social and biological networks. Proc. Natl. Acad. Sci, 2002,99(12):7821–7826.
[2] Newman MEJ. Fast algorithm for detecting community structure in networks. Physical Review E, 2004,69(6):66–133.
[3] Newman MEJ. Detecting community structure in networks. The European Physical Journal B, 2004,38:321–330.
[4] Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Phys, Rev. E, 2004,69:26–113.
[5] Newman MEJ. Modularity and community structure in networks. Proc. Natl. Acad. Sci, 2006. 8577–8582.
[6] Newman MEJ. A measure of betweenness centrality based on random walks. Social Networks, 2005,27:39–54.
[7] Guimera R, Sales-Pardo M, Amaral LAN. Modularity from fluctuations in random graphs and complex networks. Physical Review, 2004,E 70:25–101.
[8] Guimera R, Amaral LAN. Cartography of complex networks: Modules and universal roles. JSTAT, 2005,02001:1–13.
[9] Clauset A, Newman MEJ, Moore C. Finding community structure in very large networks. Physical Review, 2004,E 70:66–111.
[10] Berger-Wolf TY, Saia J. A framework for analysis of dynamic social networks. In: Proc. of the KDD. 2006. 523–528.
[11] Tantipathananandh C, Berger-Wolf TY. A framework for community identification of dynamic social networks. In: Proc. of the KDD. 2007. 717–726.
[12] Zhou D, Councill I, Zha HY, Giles CL. Discovering temporal communities from social network documents. In: Proc. of the ICDM. 2007. 745–750.
[13] Lin YR, Chi Y, Zhu SH, Sundaram H, Tseng BL. FacetNet: A framework for analyzing communities and their evolutions in dynamic networks. In: Proc. of the WWW 2008. 2008. 685–694.
[14] Yu K, Yu SP, Tresp V. Soft clustering on graphs. In: Proc. of the NIPS. 2005.
[15] Danon L, Duch J, Arenas A, Diaz-Guilera A. Comparing community structure identification. J. Stat. Mech, 2005.
[16] Tan PN, Steinbach M, Kumar V. Support envelopes: A technique for exploring the structure of association patterns. In: Proc. of the KDD 2004. 2004.
[17] http://www-2.cs.cmu.edu/~enron

单波(1985－),男,山东滕州人,硕士生,主要研究领域为图数据挖掘,社会关系网络.

姜守旭(1968－),男,博士,教授,博士生导师,主要研究领域为对等计算,传感器网络.



张硕(1982－),男,博士生,主要研究领域为图数据管理,图数据挖掘.

高宏(1966－),女,博士,教授,博士生导师,主要研究领域为并行数据库,图数据管理挖掘.

李建中(1950－),男,博士,教授,博士生导师,主要研究领域为并型数据库,传感器网络.