

互联网宏观拓扑的社团发现

张君 赵海林 川刘 晓

(东北大学计算机科学与工程学院 沈阳 110819)

摘要 自然界中存在的大量复杂系统都可以通过复杂网络加以描述,社团结构是继小世界特性和无标度特性之后发现的最为重要的复杂网络特性。社团发现对理解互联网的宏观拓扑结构至关重要。针对互联网宏观拓扑的结构特性,基于边聚簇算法思想,设计了一个基于路由特征的社团发现算法,以互联网宏观拓扑中的探测边频为影响因子定义边相似性,改造边聚簇算法中的关键聚簇过程,以发现互联网宏观拓扑中的社团结构。实验结果表明,所提算法与原算法相比,具有更高的分割密度。进一步以边介数替代探测边频,将该算法应用在其它类型网络中,同样取得了较好的效果。

关键词 复杂网络,社团发现,路由特征,互联网宏观拓扑

中图分类号 TP301.5, TP393.01

文献标识码 A

DOI 10.11896/j.issn.1002-137X.2016.11.028

Community Detecting of Internet Macroscopic Topology

ZHANG Jun ZHAO Hai LIN Chuan LIU Xiao

(School of Computer Science and Engineering, Northeastern University, Shenyang 110819, China)

Abstract A large number of complex systems in nature can be described by complex networks. Community structure is the most important feature of complex networks following the small-world and scale-free features. Community detecting is very important for understanding the macroscopic topology structure of Internet. Aimed at the structure features of the macroscopic topology of Internet, based on link clustering method, we proposed a community detecting algorithm which redefines link similarity with routing features to transform the link clustering process. It gives a better community structure in Internet macroscopic topology. It is further applied into other networks of different types by using link betweenness instead of link frequency, and better community structure can be gotten.

Keywords Complex network, Community detecting, Routing features, Internet macroscopic topology

1 概述

社团是由网络中具有相似属性的节点组成的集合,社团的特点是内部节点联系紧密而社团间节点联系松散。社团结构在一定程度上反映了网络的拓扑关系,揭示了网络中的功能实体,有助于理解网络的拓扑结构及功能结构,对于分析网络结构和特性至关重要。

社团结构的研究已有很长的历史,它最早与计算机科学中的图形分割以及社会学中的层次聚类密切相关^[1,2]。1986年 McMillan 等人在 Hillary 提出的“社团”^[3]概念的基础上,完善了社团的定义和相关理论^[4]。2002年, Girvan 和 Newman 提出复杂网络除了小世界和无标度特性外,还具有社团结构特性^[5],引起了学术界的广泛关注。

面对网络中大量的目标对象,人们常常需要将这些对象分类以进行研究或处理,比如生物学中控制病毒传播、物理学、计算机图形学中图形分割、社会学中发现犯罪团体、寻找潜在客户等,这都需要有效的社团发现工具,传统的社团发现算法^[6-9]大多数基于社团无重叠的假设,而且不能发现社团的

层次结构。现有研究发现,许多真实网络存在社团重叠现象和社团层次结构^[10,11],因此传统社团发现算法并不能很好地反映真实复杂网络的社团特征。2010年, Ahn 等人提出了基于边聚簇的社团发现算法^[12],解决了传统算法不能发现重叠和层次组织结构的问题,从而带来了突破性的进展,引起了国内外学者的广泛关注。目前,关于复杂网络社团发现的研究正在积极展开,众多专家学者潜心研究,并且提出了各自独到的见解^[13-20]。

Internet 作为当今人类社会传播信息的主要载体,在政治、经济、文化、教育乃至日常生活等方面的影响力与日俱增,越来越多的科研人员在发展网络建设、提升网络性能方面进行着不懈的努力。伴随着需求的增加与技术的进步, Internet 的规模日益庞大,拓扑结构也日趋复杂,社团发现对理解互联网的宏观拓扑结构意义重大。但是由于互联网宏观拓扑结构的复杂性,现有的算法在互联网的社团发现中存在一定的局限,不能得到令人满意的结果。

通过对互联网宏观拓扑的社团结构的研究发现,网络中连接不同社团的边往往是探测边频很大的边。本文基于边聚

到稿日期:2015-10-09 返修日期:2016-03-09 本文受国家自然科学基金资助项目(60973022)资助。

张君(1967—),女,博士,讲师,主要研究方向为复杂网络, E-mail: zhangjun1@cse.neu.edu.cn; 赵海(1959—),男,教授,博士生导师,主要研究方向为复杂网络、普适计算、嵌入式系统及数据融合等; 林川男,博士生,主要研究方向为无线传感器网络、复杂网络; 刘晓女,博士生,主要研究方向为复杂网络。

簇的算法思想,设计了一个基于路由特征的社团发现算法。该算法在使用局部拓扑信息进行边相似性计算的基础上,增加了边的路由特征进行“边加权”,以探测边频为影响因子定义边相似性,改造边聚簇算法中的关键聚簇过程,在互联网宏观拓扑结构的社团发现中效果显著。该算法可以发现重叠的社团结构和层次组织社团结构,而且可以根据边聚簇的阈值调节社团的粗细粒度。边介数有类似于探测边频的性质,以边介数替代探测边频可以使该算法移植到其它类型的网络中。通过对6个不同类型和规模的真实网络进行测试,验证了改进的算法的普适性。

2 边聚簇社团发现算法简介

边聚簇社团发现算法^[12]限制使用的边对为相邻的边对(有一个共同节点),对于相连的边对 e_{ik} 和 e_{jk} ,称其共同节点 k 为基节点,称节点 i, j 为拱节点。在定义两条边的相似性时需要忽略基节点的影响。利用 Jacard^[21]指数给出边相似性的定义如下:

$$S(e_{ik}, e_{jk}) = \frac{|n_+(i) \cap n_+(j)|}{|n_+(i) \cup n_+(j)|} \quad (1)$$

其中, $n_+(i)$ 为节点 i 的邻居节点的超集,其定义为:

$$n_+(i) = \{x | d(i, x) \leq 1\} \quad (2)$$

其中, $d(i, x)$ 表示节点 i 和 x 之间的最短路径长度。

边聚簇社团发现算法对所有的相连边对都计算出它们的相似性值,然后进行排序,得到一个有序的边相似性序列表。边聚簇过程分为两个阶段:首先将每一条边初始化为一个社团;然后按照从大到小的顺序对边相似性列表中的边进行聚合,如果边已经并入过其它社团,那么也对其社团进行合并,重复这一处理过程直到所有的边都合并到同一个社团内。这一处理过程的历史记录都被保存到一个树图里,其中树图的每一个叶子都对应一条边,因此社团的层次结构信息也都被保存下来,两个社团合并时所对应的相似性值定义为社团的聚簇强度。

层次聚类方法不断将小社团合并形成一个大社团并构造出一个系统树图,这个树图可以表示为一个社团。对于树图的不同层次,可以对树图进行分割从而得到不同聚簇强度的多个社团。

3 互联网宏观拓扑的社团发现算法

3.1 设计思想

在抽取的互联网拓扑中,探测边频是指一条边在一个探测周期内出现的频度,它代表了一条边在其它节点的路由链路中的影响力。在互联网中,信息的传播往往选择代价较小的路径,如果连接两个社团的边的探测边频较大,则两个社团之间的信息传递的代价就较小;反之亦然。因此探测边频较大的边即是影响社团之间通讯的关键边。为更合理地发现互联网的社团结构,选择探测边频作为边的权重,在基于局域的邻居节点信息的基础上将边权重作为影响因子加入原边相似性定义式(1)中,并设计两条聚簇原则:1)边权重越小的边优先划分入同一个社团;2)边权重更相近的两条边优先划分入同一个社团。由此,构造新的边相似性定义如下:

$$S(e_{ik}, e_{jk}) = \frac{|n_+(i) \cap n_+(j)|}{|n_+(i) \cup n_+(j)|} \cdot (1 + \frac{1}{(w_{ik} + w_{jk})(|w_{ik} - w_{jk}| + 1)}) \quad (3)$$

其中, w_{ik} 和 w_{jk} 表示边 e_{ik} 和 e_{jk} 的边权重。

式(3)体现了所设计的两条聚簇原则:式中边的权重越小,则右式的分母 $(w_{ik} + w_{jk})(|w_{ik} - w_{jk}| + 1)$ 越小,从而整个分式的值变大,则边的相似性系数就越高,因此两条边就会被优先划分入同一个社团;如果 w_{ik} 和 w_{jk} 接近,则 $w_{ik} - w_{jk}$ 接近于0,从而分母 $(w_{ik} + w_{jk})(|w_{ik} - w_{jk}| + 1)$ 变小,则整个分式的值变大,从而边的相似性系数变高,因此两条边就会被优先划分入同一个社团。同时,为避免分母为0,将 $|w_{ik} - w_{jk}|$ 加1。

3.2 算法描述

互联网宏观拓扑的社团发现算法描述如下:

- (1)读取网络拓扑及各边的权重;
- (2)设置边相似性阈值 *threshold* 的初值;
- (3)对于所有相连的边对,计算它们的相似性值;
- (4)将相似性值按由大到小的顺序进行排序,形成一个列表;
- (5)每条边都初始化为一个社团;
- (6)从列表头部取出一个相似性值 $S(e_{ik}, e_{jk})$,如果 $S(e_{ik}, e_{jk}) < \text{threshold}$,则算法结束;否则继续执行步骤(7);
- (7)如果边 e_{ik} 和 e_{jk} 已经在同一个社团里,则转步骤(6);否则将边 e_{ik} 和 e_{jk} 所在的社团合并,并设置新社团的聚簇强度为 $S(e_{ik}, e_{jk})$;
- (8)重复步骤(6)、(7)。

下面以图1为例来说明该算法的操作过程。依据式(3)计算出各边的相似性系数,按照由大到小的顺序进行排列,结果如表1所列。设 $\text{threshold} = 0.2$,根据上述算法,首先将边 e_{13}, e_{23} 合并成一个社团,记为 C_1 ;然后将边 e_{12}, e_{26} 依次加入到该社团,将边 e_{48}, e_{49} 合并成一个社团 C_2 ;再将边 e_{25}, e_{37} 依次加入到社团 C_1 ,将边 e_{14} 加入到社团 C_2 。此时,继续访问列表取得相似性值 $0.1786 < 0.2$,则算法结束。网络的社团划分结果如图2所示,共3个社团: $C_1(e_{13}, e_{23}, e_{12}, e_{26}, e_{25}, e_{37})$, $C_2(e_{48}, e_{49}, e_{14})$ 和 $C_3(e_{68})$ 。

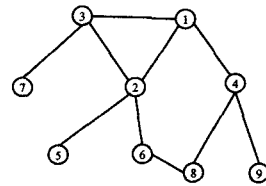


图1 一个简单的网络示例

表1 图1所示网络中任意两邻边的相似性系数

边1	边2	相似性值	边1	边2	相似性值
e_{13}	e_{23}	0.5313	e_{14}	e_{49}	0.2016
e_{12}	e_{13}	0.5167	e_{68}	e_{48}	0.1786
e_{12}	e_{23}	0.5167	e_{23}	e_{26}	0.1716
e_{12}	e_{26}	0.3403	e_{23}	e_{37}	0.1716
e_{48}	e_{49}	0.2583	e_{14}	e_{48}	0.1679
e_{25}	e_{26}	0.2574	e_{26}	e_{68}	0.1458
e_{23}	e_{25}	0.2125	e_{13}	e_{14}	0.1440
e_{12}	e_{25}	0.2067	e_{12}	e_{14}	0.1259
e_{13}	e_{37}	0.2059			

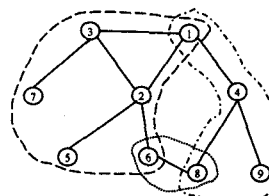


图2 图1所示网络的社团划分结果

在上述算法中,要得到合理的社团结构,关键在于边相似性阈值 *threshold* 的初值设置。目前,大多数节点层次聚类算法中普遍使用模块度作为衡量社团结构划分是否合理为标准,模块度优化方法已经成为复杂网络社团发现的一种基准方法,得到了广泛的应用。但是对于有重叠的社团结构,模块度不容易定义也不方便在算法中实现。为此引入基于社团内部边密度的分割密度,对边簇质量进行计算。分割密度表示一个网络的综合边密度,不同于模块度,分割密度不会有分辨率极限问题。其定义如下^[12]:

对于一个有 N 个节点和 M 条边的网络 $G(N, M)$, 设 $P = \{p_1, p_2, \dots, p_c\}$ 为分割后的边社团集合, 社团 p_c 中的边数为 $m_c = |p_c|$, 节点数为 $n_c = |\bigcup_{e_{ij} \in p_c} \{i, j\}|$, 那么社团 c 的分割密度为:

$$D_c = \frac{m_c - (n_c - 1)}{n_c(n_c - 1)/2 - (n_c - 1)} \quad (4)$$

式中使用社团内最多和最少(保持联通)可能的边数对边密度的计算进行了调节,假如当社团内只有一条边($n_c = 2$)时,则 $D_c = 0$ 。对于整个网络,其分割密度为:

$$D = \frac{1}{M} \sum_c m_c \cdot D_c = \frac{2}{M} \sum_c m_c \frac{m_c - (n_c - 1)}{(n_c - 2)(n_c - 1)} \quad (5)$$

文献[12]的研究表明,分割密度最大时对应的社团结构为最佳划分,因此,对树图的每一个层次都计算其分割密度,然后选择分割密度最大的层次对应的边相似性值作为边相似性阈值 *threshold* 的初值。

3.3 实验结果

本文的实验网络为互联网 IPv6 IP 级网络拓扑,来源于 CAIDA(Cooperative Association for Internet Data Analysis)数据。CAIDA 是一个在全球范围内对 Internet 结构及数据进行分析的国际合作机构,是互联网宏观拓扑分析领域中具有代表性和较大影响力的大型科研项目之一,其主要研究目标是实现对 Internet 数据的获取、分析、研究以及共享。CAIDA 对于 Internet 拓扑数据探测的架构主要是 Ark(Archipelago, 始于 2007 年 9 月,之前为 Skitter 项目)架构,是一种分布式的探测方式,利用 Scamper 技术对随机选取的目的地址发送报文,来实现探测源点之间的协作测量。在 CAIDA 探测结果的基础上使用 gephi 分析工具获取网络拓扑以及探测边频作为本文的实验基础,使用的网络由 3306726 个节点和 5291325 条边构成。

将上述算法应用于 IPv6 IP 级网络拓扑,并与原边聚簇社团发现算法^[12]进行比较。边相似性值与分割密度的曲线图如图 3 所示,其中图 3(a)对应原算法,图 3(b)对应本文算法。可见,基于探测边频的边聚簇社团发现算法能够得到比原算法更高的分割密度,因此相比原算法能更好地发现互联网拓扑中的社团结构。

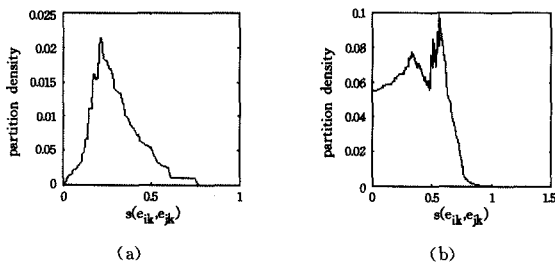


图 3 IPv6 IP 级网络上的分割密度变化曲线

4 在其它类型复杂网络上的应用

互联网中的探测边频与网络理论中的边介数相似,只是探测边频是实际探测数据上的统计结果,而边介数则是根据一个静态网络拓扑计算得来的,它们反映的都是网络拓扑的路由特征。作为网络中节点的中心地位和影响力的衡量尺度,节点介数被定义为通过它的节点对之间的最短路径的数目,表征的是测量节点对其它节点之间的信息流的影响力,特别是对那些信息流主要通过最短的可能路径进行传播的网络。GN 算法^[5]将这一定义延伸到边的情况下,定义边的“边介数”为通过它的节点对之间的最短路径的数目。如果一个网络划分成的社团之间仅由少数几条边松散地连接,那么不同社团之间的所有最短路径必经过这几条边之一。因此,连接不同社团的边将会有比较高的边介数。基于此,在其它类型的复杂网络中,可以使用边介数作为边权重参与计算。为验证算法的普适性,选择了空手道俱乐部网、美国政治书网、海豚社会网、词语关系网络和蛋白质交互网络等几个真实网络进行实验,实验结果如图 4—图 8 所示,其中图(a)对应原算法,图(b)对应本文算法。

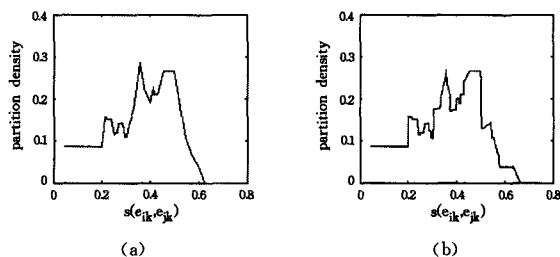


图 4 空手道俱乐部网的分割密度变化曲线

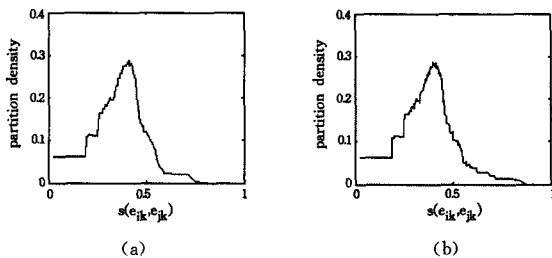


图 5 美国政治书网的分割密度变化曲线

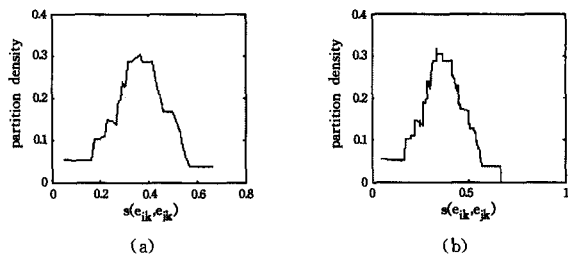


图 6 海豚社会网的分割密度变化曲线

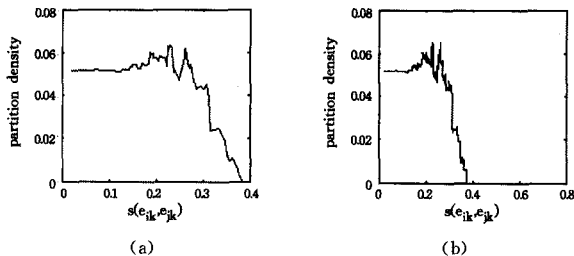


图 7 词语关系网络的分割密度变化曲线

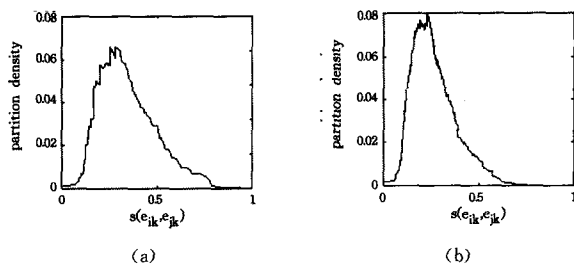


图8 蛋白质交互网络的分割密度变化曲线

表2列出了上述网络对应的最大分割密度,可见本文算法在互联网IPv6 IP级网络拓扑的社团发现中效果显著,在其他类型复杂网络的社团发现中也有较好的应用。

表2 不同网络的最大分割密度

网络	原算法	本文算法
IPv6 IP级网络	0.0213	0.0982
空手道俱乐部网络(karate)	0.2679	0.2680
美国政治书网(polbooks)	0.2866	0.2864
海豚社会网	0.3038	0.3156
词语关系网	0.0632	0.0658
蛋白质交互网络 PPI(yeast)	0.0686	0.0798

结束语 本文分析了互联网中边的路由特征即探测边频与社团结构的关系,发现探测边频较大的边是联系不同社团的关键边,根据这一关系设计改造了边聚簇社团发现算法中的边相似性定义式,对于其它类型网络则使用类似性质的边介数作为影响因子改造边相似性定义式,并改进了边聚簇算法中的关键聚簇过程,从而设计了一种基于路由特征的边聚簇社团发现算法。分别对互联网拓扑和其它几种类型的复杂网络进行测试,结果表明该算法在互联网宏观拓扑的社团发现中效果突出,在其它类型的复杂网络中也有较好的应用。

参考文献

- [1] Michael R G, David S J. Computers and Intractability: A Guide to the Theory of NP-Completeness[M]. New York: W. H. Freeman & Co Ltd, 1979
- [2] Scott J. Social network analysis[J]. Sociology, 1988, 22(1): 109-127
- [3] Hillary G. Definitions of Community: Areas of Agreement [J]. Rural Sociology, 1955, 20(2): 111-123
- [4] McMillan D W, Chavis D M. Sense of community: A definition and theory[J]. Journal of Community Psychology, 1986, 14(1): 6-23
- [5] Girvan M, Newman M E J. Community structure in social and biological networks[J]. PNAS, 2002, 99(12): 7821-7826
- [6] Newman M E J. Fast algorithm for detecting community structure in networks[J]. Physical Review E, 2004, 69(6): 066133
- [7] Clauset A, Newman M E J, Moore C. Finding community structure in very large networks[J]. Physical Review E, 2004, 70(6): 066111
- [8] Du H, Feldman M W, Li S, et al. An algorithm for detecting community structure of social networks based on prior know-

ledge and modularity[J]. Complexity, 2007, 12(3): 53-60

- [9] Duch J, Arenas A. Community detection in complex networks using extremal optimization[J]. Physical Review E, 2005, 72(2): 027104
- [10] Li D, Leyva I, Almendral J A, et al. Synchronization interfaces and overlapping communities in complex networks[J]. Physical Review Letters, 2008, 101(16): 3958-3964
- [11] Lancichinetti A, Fortunato S, Kertész J. Detecting the overlapping and hierarchical community structure in complex networks [J]. New Journal of Physics, 2009, 11(3): 19-44
- [12] Ahn Y Y, Bagrow J P, Lehmann S. Link communities reveal multiscale complexity in networks[J]. Nature, 2010, 466: 761-764
- [13] Zhao Shu, Ke Wang, Chen Jie, et al. Community detection algorithm based on clustering granulation[J]. Journal of Computer Applications, 2014, 34(10): 2812-2815 (in Chinese)
- 赵姝, 柯望, 陈洁, 等. 基于聚类粒化的社团发现算法[J]. 计算机应用, 2014, 34(10): 2812-2815
- [14] Cheng Ze-kai, Zhang Jia-yu. New community detection algorithm based on node similarity[J]. Computer Engineering and Design, 2014, 35(5): 1688-1693 (in Chinese)
- 程泽凯, 张佳玉. 基于节点相似度的社团发现算法[J]. 计算机工程与设计, 2014, 35(5): 1688-1693
- [15] Liu Ke, Huang Jian-bin, Sun He-li, et al. Label propagation based evolutionary clustering for detecting overlapping and non-overlapping communities in dynamic networks[J]. Knowledge-Based Systems, 2015, 89(c): 487-496
- [16] Hua Ye, Hu Fang-yu, Cao Jing-hua. A Local Community Detection Algorithm Based on Relative Intimacy[J]. Computer Simulation, 2014, 31(11): 278-281 (in Chinese)
- 华烨, 胡芳宇, 曹菁华. 一种基于相对关系亲密度的局部社团发现算法[J]. 计算机仿真, 2014, 31(11): 278-281
- [17] Alvarez A J, Sanz-Rodríguez C E, Cabrera Juan L. Weighting dissimilarities to detect communities in networks[J]. Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 2015, 373(2056): 1-10
- [18] Yuan Hui-hui, Cao Yu-lin, Wang Xiao-ming. Community discovery in multi-layered social network based on edge clustering [J]. Application Research of Computers, 2014, 31(2): 351-353, 377 (in Chinese)
- 袁辉辉, 曹玉林, 王小明. 基于边聚类的多层社会网络社团发现算法[J]. 计算机应用研究, 2014, 31(2): 351-353, 377
- [19] Wang Tao, Wang Hong-jue, Wang Xiao-xia. A novel cosine distance for detecting communities in complex networks[J]. Physica A: Statistical Mechanics and its Applications, 2015, 437: 21-35
- [20] Chen Feng-jiao, Li Kan. Detecting hierarchical structure of community members in social networks[J]. Knowledge-Based Systems, 2015, 87(c): 3-15
- [21] Jaccard P. Etude comparative de la distribution florale dans une portion des Alpes et du Jura[J]. Bull Soc vaudoise Sci nat, 1901, 37(142): 547-579