

考虑地理距离的复杂网络社区挖掘算法

陈 娱^{1,2}, 许 珺^{1*}

(1. 中国科学院地理科学与资源研究所 资源与环境信息系统国家重点实验室, 北京 100101;

2. 中国科学院大学, 北京 100049)

摘要: 复杂网络具有社区结构的性质, 即社区内节点的连接比社区间的连接更为紧密。目前, 具有复杂网络拓扑结构的社区挖掘算法已有很多, 但在很多地理空间的复杂网络中节点间的紧密度, 不仅与其连接关系有关, 同时与它们之间的距离有关。因此, 本文提出将节点间的地理距离考虑到社区挖掘的过程中, 修改基于模块度增量矩阵的Newman快速算法(简称CNM算法), 将 $1/d_{ij}^n$ (d 为节点 i 与节点 j 之间的距离)作为边权, 对加权网络进行社区挖掘, 从而发现既相互联系紧密又在地理空间上相互接近的社区。最后, 本文用国内航线网络作为实例, 将算法用于挖掘航线网络中城市的社区结构, 得到10个在航线网络中联系紧密且在空间分布上具有一定地域性的城市社区, 与我国的主要经济区域分布比较一致。本算法考虑地理相关性和连接紧密性, 较好地识别出空间网络的社区结构。

关键词: 复杂网络; 位置信息; 模块度; 空间距离; 航线网络

DOI: 10.3724/SP.J.1047.2013.00338

1 引言

近年来, 复杂网络的研究成为一个热门的课题。例如, 对城市路网拓扑结构的分析^[1]、机场航线网络的分析^[2-3]、人口迁移网络流的研究分析^[4]等。网络实际上都具有一个共同的性质——社区结构, 即整个网络是由若干个“群”或者“团”构成的, 社区内部节点之间的连接相对紧密, 而社区之间的连接相对比较稀疏^[5]。目前, 复杂网络的社区结构挖掘算法, 主要有图形分割算法(如Laplacian谱平分法^[6]、Kernighan-Lin算法^[7]等)和分级聚类算法(如GN算法^[8]、Newman快速算法^[9]等)。

但是, 这些算法都是以网络的拓扑结构进行社区挖掘。很多地理现象, 如Tobler第一定律所述, 具有空间相关性, 即在空间上越接近则相关性越强。可见, 地理对象间的关系不仅仅与它们之间的连接有关, 还受到空间距离的影响。因此, 对受到距离影响的复杂网络进行社区分割时, 不仅需要考虑网络的结构, 还需要考虑网络节点之间的空间距离。例如, Paul Expert等人发现手机之间的通量与

机主之间的距离有关, 故在对手机通话流网络进行社区挖掘的研究中, 考虑了地理距离, 从而挖掘出更多原本被隐藏的小社区^[10]。

本文在以网络拓扑结构的社区挖掘算法基础上, 提出考虑距离的复杂网络社区挖掘算法。本文采用Newman及Clauset等人提出的基于模块度增量矩阵的Newman快速算法(简称CNM算法)^[11], 修改该算法, 把距离加入其中, 对于有边直接相连的节点对, 将它们距离 n 次幂的倒数作为边权, 对加权网络进行社区挖掘。在此基础上, 本文用国内航线网络试验了算法, 只以网络的拓扑结构挖掘到4个社区, 而考虑网络中节点的位置信息, 挖掘到10个社区。

2 包含位置信息的复杂网络社区挖掘

在很多复杂网络中, 节点是具有空间位置信息的地理对象。假设图1是一个具有位置信息的小型社交网络, 每一个节点代表一个人, 每条边表示两个人之间有联系。如果只考虑节点间的连接关系, 得到图2所示的红色和绿色两个明显社区, 社区内

收稿日期: 2013-02-27; 修回日期: 2013-03-04。

基金项目: 国家自然科学基金项目(41171296); 国家“863”计划课题(2012AA12A211)。

作者简介: 陈 娱(1989-), 女, 江苏镇江人, 硕士生, 主要从事地理信息系统和复杂网络的研究。E-mail: chenyu@lreis.ac.cn

*通讯作者: 许 珺(1972-), 女, 博士, 副研究员, 主要从事地理信息系统、地理空间认知和知识表达的研究。E-mail: xujun@lreis.ac.cn

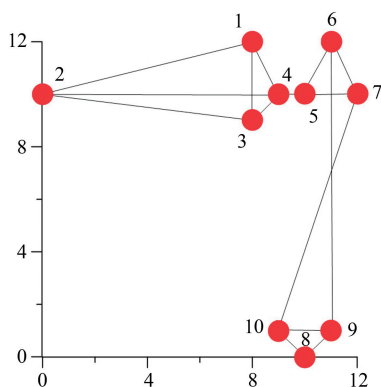


图1 10个具有位置信息的节点网络图

Fig.1 A network contains 10 nodes with location information

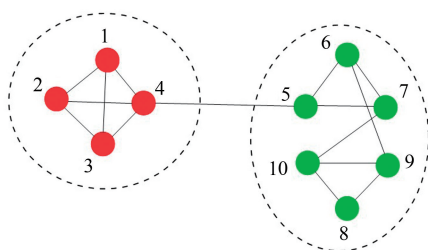


图2 拓扑空间中网络的社区结构

Fig.2 The communities in the topological space

部相互联系比较紧密,而社区之间只通过4和5两个人联系起来。

在弹性社交网络中,随着个体的移动,为了随时发现距离比较接近的好友,需要考虑节点之间的位置关系。

如果考虑图1中各个节点的位置,认为越接近的节点属于一个社区的可能性越大,而某一节点位置的移动会造成它与其他节点之间的距离拉大或缩小,那么得到的社区就会是不同的结果,如图3所示,节点5、6、7被划分到红色社区中。

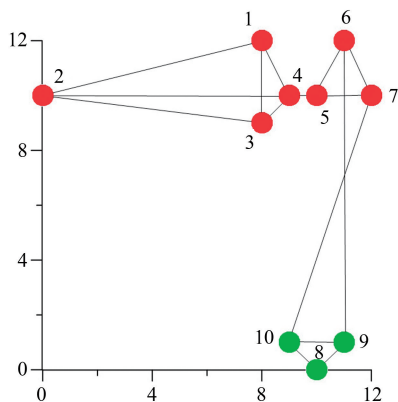


图3 考虑距离的社区挖掘结果

Fig.3 Distance/topological based communities

因此,随着复杂网络应用的扩展,只考虑网络拓扑关系的社区挖掘算法是不足的,在很多应用中都需考虑网络节点的位置信息。鉴此,本文提出考虑距离的社区挖掘算法,旨在考虑节点间连接的紧密与空间的接近。

3 基于距离的复杂网络社区挖掘算法

本文采用模块度增量矩阵的Newman快速算法(CNM算法),即由Newman、Clauset和Moore等人于2004年提出的算法。初始时,它将网络看成是点集,每个节点“各自为家”,而后将连接性强的节点对合并为一个社区,重复寻找连接性强的社区进行合并,整个流程形成了一个自下而上的树状图,如图4所示。底部的原点代表网络中的节点,随着虚线的不断上移,节点逐步合并为更大的社区,如果凝聚过程在虚线处停止就对应着一种社区划分结构。对此,Newman等人提出模块度(modularity)的概念^[12]。

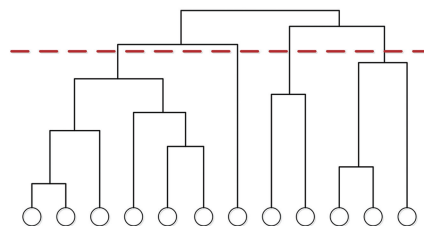


图4 凝聚算法流程树状图

Fig.4 The dendrogram generated by the aggregate algorithm

3.1 模块度

模块度(Q)是Newman等人提出的衡量网络社区划分质量的标准,是网络的一个固有属性值。

模块度定义为:

$$Q = \frac{1}{2m} \sum_{c \in P} \sum_{v, w \in c} \left(A_{vw} - \frac{k_v k_w}{2m} \right) \quad (1)$$

其中,如果节点 v, w 相连,则 $A_{vw}=1$,其他情况 $A_{vw}=0$; C 表示某一社区, P 表示网络的社区划分结构, k 为节点的度, m 为网络的总边数。

模块度是社区内部边的数目占网络总边数的比例减去社区内部边数的期望值占网络总边数的比例。 Q 值越大表示社区内部连接的稠密程度大于随机分布下的期望值。 Q 的上限为 $Q=1$, Q 越接近1,就说明网络的社区结构越明显。

定义 e_{ij} 表示网络中连接2个不同社区的节点

的边占有所有边的比例,这两个节点分别属于社区*i*和社区*j*。

$$e_{ij} = \frac{1}{2m} \sum_{v \in i, w \in j} A_{vw} \quad (2)$$

定义一个辅助变量 a_i

$$a_i = \sum_j e_{ij} \quad (3)$$

它表示与第*i*个社区中的节点相连的边在所有边的数目中所占的比例。

可由公式(2)、(3)将模块度公式(1)推导为:

$$Q = \sum_i (e_{ii} - a_i^2) \quad (4)$$

3.2 CNM算法及其流程

在CNM算法思想中,社区的凝聚是通过计算合并两个社区后模块度增量值来判断的,社区的合并总是向着增量值最大的方向进行。合并两个社区后的模块度增量可由公式(4)得到:

$$\Delta Q_{ij} = e_{ij} + e_{ji} - 2a_i a_j = 2(e_{ij} - a_i a_j) \quad (5)$$

为了提高速度,该算法直接构建 ΔQ 矩阵,从而快速地找到 $\max \Delta Q$ 。CNM算法的流程为:

(1)初始化:每个节点看做一个社区,网络初始的模块度值为0。构建对称矩阵*E*,其中元素 e_{ij} 为:

$$e_{ij} = \begin{cases} \frac{1}{2m}, & \text{节点}i\text{与节点}j\text{相连} \\ 0, & \text{其他} \end{cases} \quad (6)$$

其中,*m*为网络中的总边数。

$$a_i = k_i / 2m \quad (7)$$

其中, k_i 为节点*i*的度。

合并任意两个节点*i*和*j*都对应一个模块度增量值,根据公式(4)构建对称的模块度增量矩阵,初始的模块度增量矩阵中的元素满足:

$$\Delta Q_{ij} = \begin{cases} \frac{1}{2m} - k_i k_j / (2m)^2, & \text{节点}i\text{和}j\text{相连} \\ 0, & \text{其他} \end{cases} \quad (8)$$

(2) 构建一个最大堆*H*,将模块度增量矩阵每一行的最大值存放在最大堆*H*中;

(3) 从最大堆*H*中选择最大的 ΔQ_{ij} ,合并对应的社区*i*和*j*,并将合并后的社区标记为*j*,然后更新模块度增量矩阵,将第*i*行元素和第*i*列元素删除,更新第*j*行和第*j*列,更新分为以下3种情况:

$$\Delta Q'_{jk} = \begin{cases} \Delta Q_{ik} + \Delta Q_{jk}, & \text{社区}k\text{与社区}i\text{社区}j\text{相连} \\ \Delta Q_{ik} - 2a_i a_k, & \text{社区}k\text{只与社区}i\text{相连} \\ \Delta Q_{jk} - 2a_j a_k, & \text{社区}k\text{只与社区}j\text{相连} \end{cases} \quad (9)$$

(4) 辅助变量 a_i 的更新:

$$a'_j = a_i + a_j; a'_i = 0 \quad (10)$$

(5) 重复步骤(3)直到模块度增量矩阵中的最大元素都小于0后,停止凝聚过程,此时得到网络的最佳社区结构。因为在整个算法过程中,模块度*Q*只有一个最大值,当模块度增量矩阵中最大的元素小于0以后,*Q*值只会不断变小,故此认为此时得到的社区结构为最优的结果。

3.3 考虑距离的CNM算法

对于有边直接相连的节点*i*和*j*,计算2个节点间的距离,将距离*n*次幂的倒数作为边权,即 $w_{ij} = 1/d_{ij}^n$,将网络看作为一个加权网络进行社区挖掘。

模块度*Q*值修改为:

$$Q^w = \sum_i [e_{ii}^w - (a_i^w)^2] \quad (11)$$

其中,定义一个对称矩阵 E^w , e_{ij}^w 就是它的元素,表示网络中连接社区*i*和社区*j*的节点的边权占网络总边权的比例。初始时:

$$e_{ij}^w = \begin{cases} w_{ij}/2w, & \text{如果节点}i\text{和}j\text{相连} \\ 0, & \text{其他} \end{cases} \quad (12)$$

其中,*w*为网络的边权之和,即:

$$w = \sum_{i,j} w_{ij} = \sum_{i,j} 1/d_{ij}^n \quad (13)$$

辅助变量 a_i^w 为矩阵 E^w 中对角线上的各元素之和,因此,

$$a_i^w = \sum_j e_{ij}^w \quad (14)$$

$$\text{初始时, } a_i^w = \sum_j e_{ij}^w = \frac{\sum_j w_{ij}}{2w}$$

在构建的模块度增量矩阵中 ΔQ_{ij}^w 的初始值为:

$$\Delta Q_{ij}^w = \begin{cases} \frac{1}{2w} - \frac{\sum_j w_{ij} \sum_i w_{ij}}{(2w)^2}, & \text{节点}i\text{和}j\text{相连} \\ 0, & \text{其他} \end{cases} \quad (15)$$

在初始化之后,构建最大堆 H^w 存放 ΔQ^w 矩阵每行的最大值,合并最大堆中最大的 ΔQ_{ij}^w 对应的社区*i*和社区*j*,并更新 ΔQ^w 矩阵,重复该过程直到模块度增量矩阵中所有的元素都小于0时停止,从而得到考虑节点间地理距离后网络的最佳社区结构。

4 网络社区挖掘的应用实例分析

本文选取国内航线网络作为研究对象。航线

网络是将城市看作节点,有航班往来的城市之间用边相连。Guimerá等人研究发现世界航线网络是一个无标度小世界网络^[13-14];Bagler等人研究了印度航线网络的最短路径长度、度分布,以及聚类系数、中心性等复杂网络特征度量^[2],Guida等人研究了意大利航线网络的拓扑性质^[15]。目前,很多有关国内航线网络的研究^[16-17]表明,国内航线网络显示出小世界模型的特征属性:具有较高的聚类系数,同时又具有较小的最短路径长度^[18]。

城市之间的航线连接在一定程度上反映了城市之间的经济联系。本文将考虑距离的CNM算法用于国内航线网络的社区划分,其包括目前通航的国内城市169个(其中,将同一座城市的机场归并为一个,例如,北京首都机场与北京南苑机场合并为北京,上海浦东与上海虹桥合并为上海,重庆江北机场和万州机场合并为重庆等),共包含航线1413条(数据来源于中国民航网、去哪儿网及携程网等,获取时间为2012年3月)。该网络的平均度大小为16.72,度分布符合幂律分布,网络的平均最短路径长度为2.09108,聚类系数为0.718,该网络具有无标度网络和小世界网络的特性。

4.1 国内航线网络社区挖掘结果

本文首先采用模块度增量矩阵的Newman快速算法(CNM算法),对国内航线网络进行社区挖掘,得到4个社区。该算法只考虑网络的拓扑关系,4个社区分别用红色、黄色、绿色和蓝色表示。其中,红色社区共含有87个节点,包含几个度很大节点,即航线很多的机场,如北京,度为130;上海,度为112;广州,度为104,等等,这些城市间的航线连接关系很紧密;黄色社区共含有58个节点,其中,包含几个度较大的节点:长沙、大连、厦门、天津等;蓝色社区含有17个节点,度最大的节点为西安,其他节点为金昌、固原等与西安通航的小城市;绿色社区含有7个节点。社区的空间分布如图5所示,可见国内航线网络的社区结构地域性不强。

4.2 考虑距离的国内航线网络社区挖掘分析

若将城市间的紧密度与航线网络和城市之间的距离联系起来,这些有机场的城市之间的紧密度不仅与航线相连有关,也遵循Tobler第一定律,地理位置越接近越紧密。

对国内航线网络用改进的CNM算法进行社区

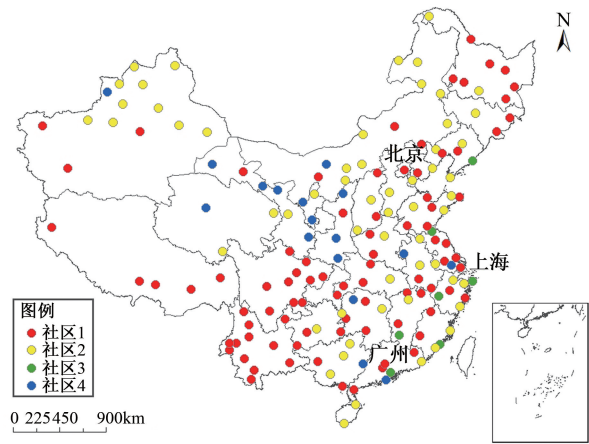


图5 国内航线网络社区划分结果

Fig.5 The communities of China flight network in the topological space

挖掘,将城市间的距离考虑到社区挖掘的过程中,将距离 n 次幂的倒数作为边权。 N 的取值采用试验法,分别试验了 n 取值为1、2、3的情况,其中,以 n 取值为2结果最佳。 n 取值为1时社区的地域性不明显, n 取值为2社区的空间分布具的地域性就已显示出来,而 n 取值为3时则社区划分太破碎,划分出很多小社区。因此,这里只讨论 n 取值为2的结果。记录算法过程中每一次合并后网络的模块度值,得到模块度函数曲线如上图6所示。这是一个典型的凝聚算法的模块度函数曲线,初始时,169个城市各自为一个社区,模块度值为0.003,合并对应的两个城市,重复该过程,当社区凝聚为10个社区时,模块度 Q 达到峰值0.54121,这时得到的就是网络的最优社区结构,之后模块度值呈现快速下降的趋势。

社区挖掘的结果在地理空间上的分布如图7所示。对比两种结果可以发现,由于距离的引入,社

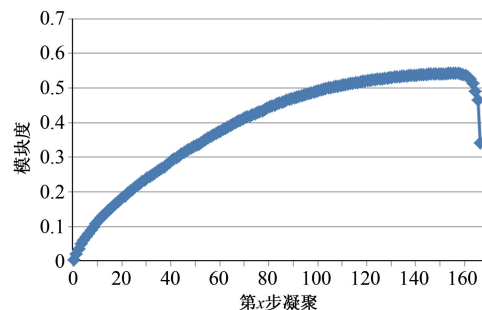


图6 算法过程中的模块度变化曲线

Fig.6 The modularity Q over the course of the algorithm (the x axis shows the number of joins)

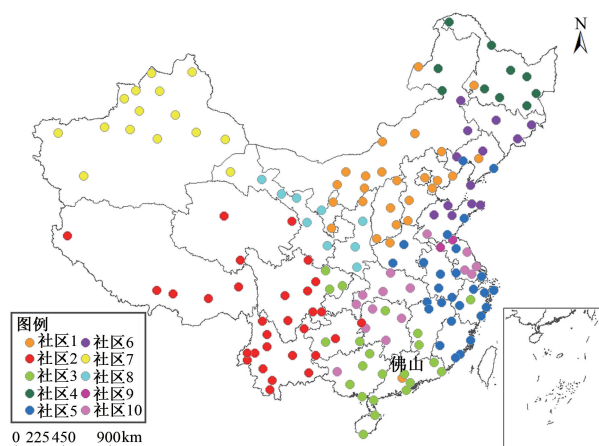


图7 考虑距离的国内航线网络社区划分图

Fig.7 The communities of China flight network generated by the distance based algorithm

区更具有地域性,即将地理空间上接近且连接关系紧密的城市划分到一个社区。

例如,图7,北京、上海、广州3市,在拓扑空间中,它们连接紧密,属于同一个社区,而考虑地理距离之后,由于3市距离相距很远,得到的结果是3市属于不同的社区;而有一些节点仍然归属到地理空间上距离较远的社区,例如,佛山(图7中标注的橘红色点),由于它只与北京南苑机场有航班,因此,它与北京同属于一个社区。

除此之外,从图7的社区空间分布来看,它与中国经济圈分布有很大程度上的吻合,表1为我国7大经济区域分布表^[9]。比较发现,这种既考虑网络的连接关系又考虑城市间距离的社区挖掘算法得到的结果,在空间上形成了一种与经济联系紧密度密切相关的分布形态。

表1 我国7大经济区域及其对应的主要省份

Tab.1 Seven main economic regions and the corresponding provinces of China

经济区域	主要省(区)
东北地区	黑龙江、吉林、辽宁
环渤海地区	京、津、河北、山东、内蒙古7盟市
长三角地区	上海、江苏、浙江、安徽
东南沿海地区	广东、广西、福建、海南
中部地区	湖南、湖北、江西
西南华南部分地区	四川、重庆、云南、贵州、西藏
西北地区	陕西、甘肃、宁夏、青海、新疆

总之,考虑地理距离之后的社区发现结果具有一定地域性,同时没有丧失掉复杂网络的拓扑关

系。相比原算法得到更多的社区,这些小社区的节点综合考虑了地理空间上接近与连接关系紧密。对于不同的研究目的和应用需求,我们需要探讨是否考虑地理距离。例如,本文实例中的航线网络,若只关注网络的拓扑结构特征,那么无需加以空间上的研究(如探讨航线网络中哪些城市之间通航较为方便,就可以用原算法得到的结果加以分析);若研究与空间分布密切相关,则需要在网络结构的基础上,加以地理距离的探讨,本文修改后的算法为此提供了一个很好的研究方法。

5 结论和展望

本研究修改了模块度增量矩阵的Newman快速算法,对于有边直接相连的节点对,将它们距离 n 次幂的倒数作为边权,对加权网络进行社区挖掘。结果表明,相比原算法,考虑地理距离的复杂网络社区挖掘算法得到的社区显示出地域性的特征,原算法得到4个社区,修改后的算法得到10个社区,使得有航线紧密相连且地理距离相近的城市划分到一个社区。

Expert等人认为距离越近,通话的概率越大,如果根据拓扑关系对网络进行分割,倾向于将近距离的节点划分为同一社区。他们通过比较一定距离下实际的连接边数量与同一距离下连接边数量的期望值,剔除了距离的影响,目的在于挖掘紧密联系且不受空间约束的社区^[10]。本文考虑到地理相关性,旨在挖掘受空间距离约束的社区,因而加强了距离在网络分割中的作用。航线网络一般受空间的约束不大,往往距离非常近的城市之间航线较少或没有,而距离较远的城市之间航线反而多。根据地理相关性,城市之间的关系是受空间约束的,但是难以量化表达。本文考虑距离的社区挖掘算法,从航空网络中挖掘出受空间约束的城市社区。此种方法同样适用于位置的社交网络、移动网络等复杂网络的社区发现。

本研究的创新点在于将节点的地理位置信息考虑到社区挖掘的过程中,认为网络中节点间的紧密性不仅仅与它们的连接关系相关,同时也符合距离越接近越紧密,从而得到一种地域上相近且连接紧密的社区挖掘结果。但是,权重如何确定,距离对节点间紧密性的影响如何衡量等问题,有待今后深入研究。

参考文献:

- [1] 李树彬,吴建军,高自友,等.基于复杂网络的交通拥堵与传播力学分析[J].物理学报,2011,60(5):050701(1-9).
- [2] Bagler G. Analysis of the airport network of India as a complex weighted network[J]. Elsevier, 2008,387(12): 2972-2980.
- [3] Li W, Cai X. Statistical analysis of airport network of China [J]. Physical Review E, 2004,69(2):046106(1-6).
- [4] Guo D. Flow mapping and multivariate visualization of large spatial interaction data[J]. IEEE, 2009,15(6): 1041-1048.
- [5] 汪小帆,李翔,陈关荣.复杂网络理论及其应用[M].北京:清华大学出版社,2006.
- [6] Pothén A, Simon H, Liou K P. Partitioning sparse matrices with eigenvectors of graphs[J]. SIAM. J. Matrix Anal. & Appl., 1990,11(3):430-452.
- [7] Kernighan B W, Lin S. A efficient heuristic procedure for partition graphs[J]. Bell System Technical Journal, 1970 (49):291-307.
- [8] Girvan M, Newman M E J. Community structure in social and biological networks[J]. Proceedings of the National Academy of Science, 2001(99):7821-7826.
- [9] Newman M E J. Fast algorithm for detecting community structure in networks[J]. Physical Review, 2004,69(6): 066133(1-5).
- [10] Expert P, Evans T S, Blondel V D, *et al.* Uncovering space-independent communities in spatial networks[J]. PNAS, 2011,108(19):7663-7668.
- [11] Clauset A, Newman M E J, Moore C. Finding community structure in very large networks[J]. Physical Review E, 2004,70(6):066111(1-6).
- [12] Newman M E J, Girvan M. Finding and evaluating community structure in networks[J]. Physical Review E, 2004, 69(2):026113(1-15).
- [13] Guimerà R, Mossa S, Turtzschl A, *et al.* The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles[J]. PNAS, 2005,102(22): 7794-7799.
- [14] Guimerà R, Amaral L A N. Modeling the world-wide airport network[J]. The European Physical Journal. B, 2004 (38):381-385.
- [15] Guida M, Maria F. Topology of the Italian airport network: A scale-free small-world network with a fractal structure?[J] Chaos, Solitons and Fractals, 2007,31(3): 527-536.
- [16] Wang J E, Mo H H, Wang F H, *et al.* Exploring the network structure and nodal centrality of China's air transport network: A complex network approach[J]. Journal of Transport Geography, 2011(19):712-721.
- [17] 明朝辉,韩松臣,张明.基于复杂网络理论的中国民航机场航线网络静态特征挖掘和应用[J].江苏教育学院学报(自然科学),2011,27(3):22-27.
- [18] Watts D J, Strogatz S H. Collective dynamics of 'small-world' networks[J]. Nature, 1998,393(6684):440-442.
- [19] 宋岭,魏秀丽.中国经济区域划分综述[J].新疆财经,2000 (2):48-49.

A Distance-based Method of Community Detection in Complex Networks

CHEN Yu^{1,2} and XU Jun^{*}

(1. Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing 100101, China; 2. University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: One feature discovered in the study of complex networks is community structure, it means that vertices are gathered into several groups that the edges within groups are more than those between them. Detecting the community structure hidden in the networks has extensive application prospects. Recently, many approaches have been developed for finding communities, such as Gieven-Newman algorithm, Newman fast algorithm and so on. Also, there are some approaches consider both the network topology and the attributes of vertices, such as SA-cluster algorithm. However, a few studies focused on considering geographic distance between vertices. Tobler's first law says that near things are more related than distant things. Based on this proposition, we think that the strength of interaction between vertices is concerned with geographic distance. By defining the weights of

the edges in the network as the function of distance between two directly connected nodes, we modified the fast modularity maximization algorithm (CNM algorithm). The weight is defined in such a way that the closer the distance, the greater the weight. The algorithm is tested on the flight network of China. We consider the strength of interaction between two cities is related with both the connection of flights and distance. We find 10 communities in the air transport network, and the distribution of communities displays regionally.

Key words: complex network; community structure; modularity; spatial distance; flight network

***Corresponding author:** XU Jun, E-mail: xujun@lreis.ac.cn