

# 社区发现在复杂网络划分中的应用

谭莹,张然,朱东生

(长沙理工大学 计算机与通信工程学院,湖南 长沙 410000)

**摘要:**在无标度网络中,社区结构是普遍存在的一种网络结构特性,社区结构是网络中间层的描述,是对网络的自然压缩。文中基于这一事实,将社区结构发现方法加入到多层网络划分框架中,提出了基于社区结构的多层网络划分改进策略。该方法首先对无标度网络进行社区发现;然后以发现的社区结构为单位,对原网络进行压缩;之后对压缩后的网络进行初始划分;最后将划分结果还原为对原网络的划分。在进行初始划分时,为获得较好的划分效果,引入了0-1规划方法,并使用K-L算法进行优化。通过对比实验,结果表明把社区结构引入多层网络划分方法中,可以获得更好的划分。

**关键词:**网络划分;无标度网络;社区结构;多层网络划分

中图分类号:TP39

文献标识码:A

文章编号:1673-629X(2014)11-0234-04

doi:10.3969/j.issn.1673-629X.2014.11.059

## Application of Community Discovery in Complex Network Division

TAN Ying, ZHANG Ran, ZHU Dong-sheng

(School of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha 410000, China)

**Abstract:** In the scale-free network, the community structure is ubiquitous structural properties of a network, community structure is the description of network middle layer, which is a natural compression for network. Based on this fact, the community structure discovery methods are added into multi-layer network framework, propose an improved multi-layer network division strategy based on community structure. This method first carries out the community discovery for scale-free networks, then with the discovered community structure as a unit, conduct the original network compression, later divide the network compressed initially, finally the result will be reverted to the original network division. During initial division, in order to get a better division results, introduce the 0-1 programming methods and algorithms and optimized by the use of K-L. By comparing the experiment, the results show that introduction of community structures into multi-layer network division method, can get a better division.

**Key words:** network segmentation; scale-free network; community structure; multi-layer network division

## 0 引言

随着通信技术的不断发展,不同的对象所形成的网络的规模不断<sup>[1]</sup>扩大;随着信息采集技术的不断发展,人们获取各种信息的能力不断增强;随着存储器价格的不断下降<sup>[2]</sup>,可以存储的信息越来越多。这就使得人们拥有了大量的网络数据,如何对这些海量的数据进行分析,从中提取出感兴趣的知识,对以后的决策等进行支持。面对如此庞大的数据量,普通的单个计算机远远不能满足要求,因此使用并行计算或者分布式计算来解决这类问题。

并行计算或者分布式计算面临的一个首要问题就是,如何把所掌握的信息分配到不同的计算节点上,以

达到最优的性能。这可以转化为网络划分的问题,把同一个区域的数据放在同一个计算节点上,划分后保持各个区域的规模近似相等。这样就可以在保证并行性的同时,减少节点间的信息传输。这也说明了网络划分在处理这一问题时的重要性,划分得好,可以大大加快数据分析处理的效率;反之,则会大大降低。

文中针对多层划分技术在处理无标度网络方面的局限性,提出了基于社区结构的网络划分方法。使用社区结构对原网络进行压缩,有以下优点:

(1)社区结构是网络的中间层表示,是对网络的一个自然压缩;

(2)社区结构保留了原网络的拓扑特性,对划分

收稿日期:2013-10-22

修回日期:2014-03-03

网络出版时间:2014-09-11

基金项目:国家“863”高技术发展计划项目(2013AA01A212)

作者简介:谭莹(1986-),女,硕士研究生,研究方向为数据处理。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20140911.0941.002.html>

的结果可以直接还原为对原网络的划分;

(3) 无标度网络中几乎都含有社区结构,因此使用社区结构对网络进行压缩具有通用性;

(4) 社区结构发现的计算复杂度并不高,可以达到线性时间复杂度。

因此可以使用社区结构对原网络进行压缩,从而得到较高质量的网络划分。

## 1 相关研究

### 1.1 二分法

二分划分方法<sup>[3]</sup>是指通过调整局部节点的区域归属,使得划分时的割边数达到局部最优的一种方法。Kernighan 和 Lin 所提出的 K-L 算法, Fiduccia 和 Mattheyses 所提出的 F-M 算法都是二分划分方法,其中 K-L 方法是最为经典的算法之一。

K-L 算法的基本思想是通过不断地置换两个区域中的节点,找到使得划分的割边数最小的两个区域,该方法得到的划分两个区域的大小一定是差距最小的。具体过程如下:

(1) 输入图的平衡二分划分,  $G = (V, E) = \{G_1, G_2 \mid G_1 = (V_1, E_1), G_2 = (V_2, E_2), V = V_1 \cup V_2, \emptyset = V_1 \cap V_2\}$ ;

(2) 选择两个未交换过的节点进行交换,并对这两个节点进行标记;

(3) 如果交换后所得的划分割边数较小,则确定要交换这两个节点,否则将这两个节点还原;

(4) 重复步骤(2)和步骤(3)直到所有的节点都被标记。

### 1.2 分层网络划分

随着多层划分技术<sup>[4]</sup>的兴起<sup>[5]</sup>,多层网络划分方法成为当前流行的划分方法。多层划分方法主要分为图压缩、压缩图划分和图还原三个过程<sup>[6]</sup>,见图 1。

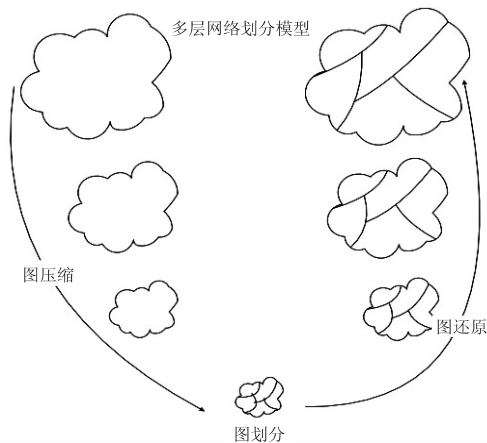


图 1 多层网络划分模型

图压缩阶段是通过一定的规则把图中某些节点结合在一起形成一个节点<sup>[7]</sup>,不断地进行节点的合并直

到图的规模达到足够小。当网络规模足够小时,可以在较短的时间内得到  $k$ -路划分的结果。图还原过程是通过节点合并规则的逆反过程<sup>[8]</sup>,把对压缩图<sup>[9]</sup>的划分结果逐步还原为对原图的划分<sup>[10]</sup>。

## 2 大规模复杂网络划分

### 2.1 问题引入

多层网络划分技术<sup>[11]</sup>是目前最为流行的图划分技术,它一共包含三个阶段:网络规模压缩、初始划分和划分结果还原<sup>[12]</sup>。Karypis 等人对这个问题进行了深入的研究,并通过实验证实了,在处理无标度网络方面<sup>[13]</sup>,目前的压缩方法存在着局限性<sup>[14]</sup>。因为在压缩阶段,算法收敛速度过快,从而使得压缩后的网络密度变大,在进行划分的过程中导致割边的条数过多。这个问题虽然可以通过划分结果还原的阶段进行一定的优化,但是这样做就大大增加了计算的复杂度。

### 2.2 相关定义

定义 1:中间图  $G = (V, E)$  中包含有社区结构  $C_1, C_2, \dots, C_m$ , 其中社区结构满足:  $C_j \cap C_k = \emptyset, j \neq k; G(V, E) = \bigcup_j C_j$ 。以社区结构为顶点,以社区间的联系为边的图称为中间图。

定义 2:加权中间图。在中间图中,如果以  $|C_j|$  作为对应节点的权值,  $C_j$  和  $C_k$  之间的连边数或者连边的权值之和作为对应节点间的连边权重,则称这样的中间图为加权中间图。

定义 3:区域中心。在加权中间图中,称未进行标记的具有最大权值的节点为区域中心。

定义 4:均衡节点数。网络  $G$  中含有  $n$  个节点,把这  $n$  个节点划分成  $k$  个区域,称  $n$  与  $k$  的比值为均衡节点数。

定义 5:边缘节点区域  $V_i$  中满足下列条件的节点称为区域  $V_i$  中的边缘节点,简称边缘节点。

(1) 和区域外部的节点有边直接相连;

(2) 不是区域的中心。

定义 6:均衡度。均衡度是评价网络划分后,各个区域均衡程度的一种度量,它是最大区域与最小区域规模的差和均衡节点数的比值。

在实际进行区域划分时,使所有的划分区域的绝对均衡是不现实的,也是没有必要的。因为:网络中的节点不一定是划分区域的整数倍,在这种情况下要使各个划分区域绝对均衡是不可能的;可能存在这样的一些节点,把它划入一个区域之后虽然使得区域的规模略大,但是可以大大地降低划分的割边数或者划分的割边的权值和。在进行区域划分时,只要保证区域的均衡度不大于某个阈值,就可以认为该划分是均衡的。

### 3 基于社区的网络划分方法

基于社区结构的网络划分方法,其基本思路如下:

首先,发现网络中隐含的社区结构,并将发现社区结构的网络进行压缩,这样就得到了加权的中间图。

其次,为了使划分后各个区域的规模是均衡的,使用均衡节点数来控制各个区域的规模。这样做之后就类似于一个 0-1 背包问题,如何把这些具有权值的节点放入背包中,使得背包尽可能被装满。

之后,对 0-1 规划所得到的初始划分,只能保证划分是均衡的,但是不一定是最优的,因为割边数可能会比较大。此时可以通过局部调整中间图中节点的归属问题,以使割边数达到尽可能得小。

最后,将经过优化的对加权中间图的划分,还原为对原始网络的划分。

#### 3.1 算法 1

输入:加权网络  $G$ ,划分区域数  $k$ ,伸缩度阈值  $\rho$ ;

输出:网络  $G$  的  $k$ -路划分。

- (1) 计算均衡节点数  $p$ ;
- (2) 计算绝对伸缩度  $q$ ;
- (3) 对加权网络  $G$  进行社区发现;
- (4) 压缩网络  $G$  得到中间图  $g$ ;
- (5) 对  $g$  使用 0-1 规划的方式,进行初始划分;
- (6) 利用伸缩度,使用 K-L 方法,对初始划分进行

优化;

- (7) 重复(3)和(4),直到划分出  $k$  个区域;

- (8) 把划分还原为对原网络  $G$  的划分。

算法 1 对基于社区的网络划分方法的流程进行了概括,算法 2 和算法 3 分别对算法 1 中的第五步和第 6 步进行详细的介绍。

#### 3.2 算法 2

输入:加权中间图  $g$ ,均衡节点数  $p$ ;

输出:加权  $G$  中的一个划分  $V_1$  和  $G-V_1$ 。

- 1 find  $\max(w(v) \text{ in } V)$
- 2 find  $\sigma(w) = \{w_1, w_2, \dots, w_m\}$
- 3 以变量  $y_j$  代表节点  $w_j$  是否属于区域中心  $v$  所在区域
- 4 设定目标函数  $njo \mid \sum y_j w_{jj} - p + x(w) \mid$
- 5 约束条件:  $y_j = 0 \text{ or } 1$
- 6 求解 0-1 规划
- 7 if  $\text{sum}(x_i) = 1$  then
- 8 all  $\sigma(w)$  with  $v$
- 9 find  $\sigma(V_1)$  goto 3
- 10 else  $\text{sum}(x_i) = 0$  then
- 11 exit
- 12 else

13 some  $\sigma(w)$  with  $v$

14 exit

15 end if

#### 3.3 算法 3

输入:初始划分  $V_1, G-V_1$ ,绝对伸缩度  $q$ ;

输出:优化后的划分  $V_1, G-V_1$ 。

- 1 find  $\theta(V_1)$  and  $\sigma(V_1)$
- 2 find pairs  $(u, v)$  with  $v \in (V_1), w \in (V_2), x(v) - x(w) \mid < q$
- 3 for  $p$  in pairs  $(u, v)$  loop
- 4 计算交换  $u, v$  的收益  $g$
- 5 if  $g > 0$  then
- 6 change  $u$  and  $v$
- 7 end if
- 8 end loop

目前好的社区结构发现算法可以在线性时间内发掘出复杂网络中包含的社区结构。使用  $k$ -路划分对加权中间图进行划分,其计算复杂度为  $O(|E'|)$ ,其中  $|E'|$  是加权中间图中的边数,而且满足  $|E'| \gg |E'|$ 。因此 CDGP 算法的整体复杂度为  $O(|E|)$ ,在最坏的情况下为  $O(|n^2|)$ ,即完全图的情况下,其中  $n$  为输入图中的节点个数。因此基于社区结构的图划分方法可以应用到大规模的社区网络划分中。

### 4 实验评估

图 2 显示了五种情况下,CDGP 算法和 Metis 划分工具进行网络划分时的割边数目。从图中可以看出,CDGP 算法整体上具有较小的割边数,特别是在划分的区域较多时,这种现象更加明显。这说明使用 CDGP 方法对大规模网络进行多区域划分时,可以得到较好的划分效果。

图 3 显示了对四种实际网络进行划分后,各个区域规模的差别情况,图中纵坐标是划分结果的均衡度。从图中可以看出,当划分区域的个数较少时,CDGP 算法和 Metis 划分的结果均衡度相差不大,有时 Metis 反而更好一些。但是随着划分区域个数的增加,均衡度的差别开始体现出来,CDGP 算法明显要优于 Metis 划分工具。因此在对大规模网络进行的多区域划分中,CDGP 算法要更有优势一些。

### 5 结束语

针对现有的多层网络划分方法对无标度网络划分结果不理想的情况,在深入研究了网络中的社区结构之后,提出了一种基于社区结构的多层网络划分改进策略。该方法首先发现网络中的社区结构,以社区结构为单位对无标度网络进行压缩,以获得更为合理的

压缩;然后对压缩后的规模较小的网络进行初始划分;最后把划分的结果还原为对原网络的划分。文中对该方法进行了详细的介绍,并通过实验和 Metis 多层网

络划分工具进行对比,发现基于社区结构的网络划分可以获得更好的划分效果。

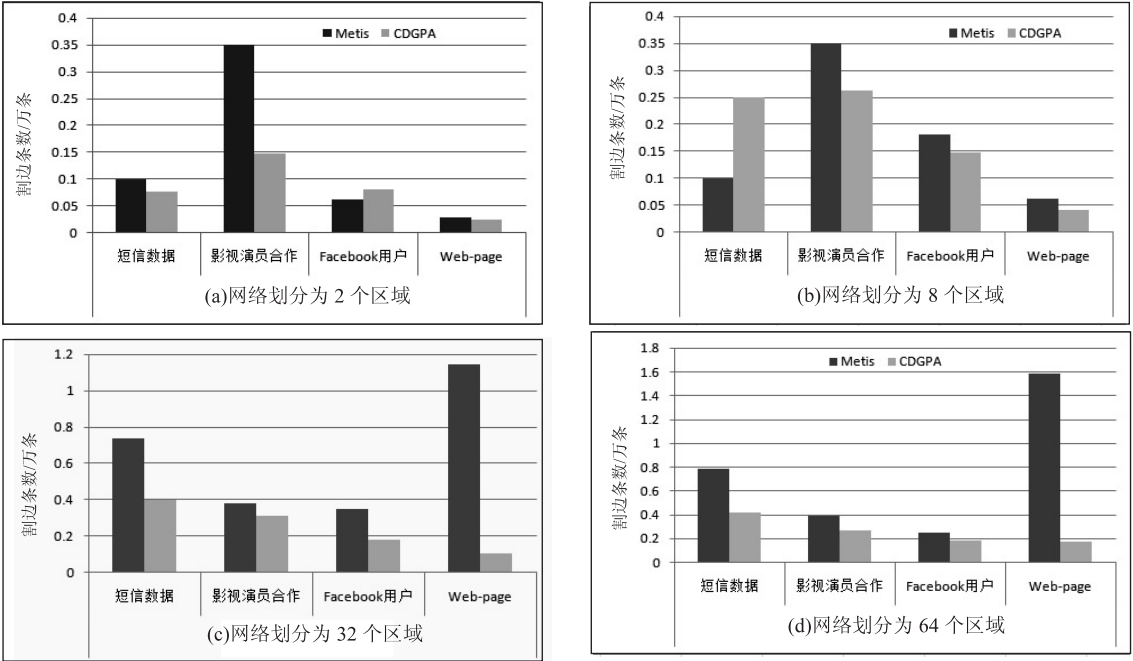


图 2 CDGP 算法和 Metis 划分工具进行网络划分时的割边数目

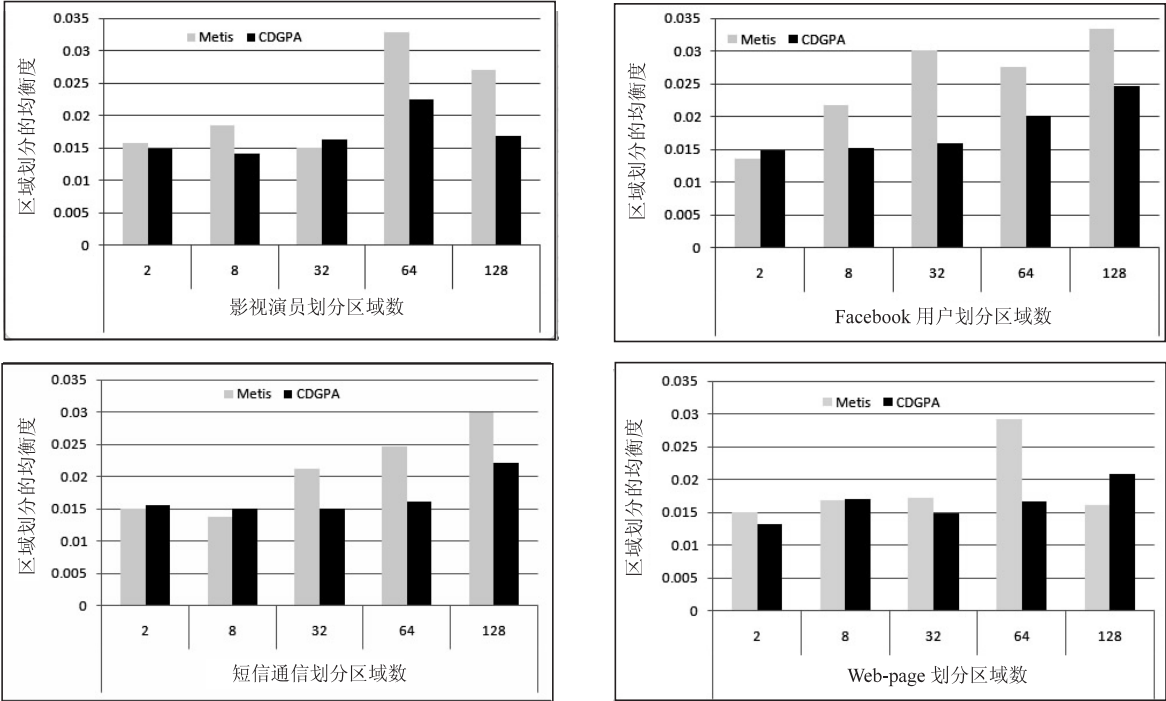


图 3 对四种实际网络进行划分后各个区域规模的差别情况

在无标度网络的划分中,文中解决了多层网络划分方法的局限性,但是依然有一些极端的情况没有解决。比如:在社区发现之后,得到的社区结构的个数小于要求划分的区域数,或者发现的社区结构非常不均衡,存在一个规模很大的社区,它的规模超过了进行划分时的均衡节点数,或者最为极端的情况,网络中节点的联系足够紧密,以至于整个网络只有一个社区结构。

参考文献:

[1] de Silva V, Tenenbaum J B. Global versus local methods in nonlinear dimensionality reduction[C]//Proc of the 15th advances in neural information processing systems. Cambridge: MIT Press, 2003: 721-728.

[2] Belkin M, Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation[J]. Neural Computation,



5 结束语

正是由于 OFDM 技术能有效对抗多径效应,消除符号间干扰,对抗频率选择性衰落,并且具有很高的频谱利用率,这才使它符合 4G 对于在无线条件下稳定可靠地实现高速率数据传输的需求,使它成为 4G 的核心技术。OFDM 技术的应用潜力是非常巨大的,但还有诸多问题需要解决,只有对 OFDM 技术进行更加深入的研究,很好地克服 OFDM 的缺陷,才能使其在 4G 移动通信等诸多领域发挥巨大的作用。

参考文献:

[1] 李荣秀,王心水. 4G 中的关键技术[J]. 甘肃科学学报, 2006,18(3):87-90.

[2] Steendam H, Moeneclaey M. Sensitivity of orthogonal frequency division multiplexed systems to carrier and clock synchronization errors [J]. Signal Processing, 2000, 80(7):1217-1229.

[3] 陈良明,韩泽耀. OFDM—第四代移动通信的主流技术[J]. 计算机技术与发展,2008,18(3):184-187.

[4] 覃凤清,郭洪容. 正交频分复用系统及其关键技术研究[J]. 宜宾学院学报,2008(6):58-60.

[5] 江秀萍. OFDM 中的峰均比降低技术[D]. 西安:西安理工大学,2007.

[6] 陈明. 基于流星余迹通信的 OFDM 实现研究[D]. 西安:

西安电子科技大学,2008.

[7] 吴胜楠. OFDM 系统峰均比降低算法研究[D]. 西安:西安理工大学,2009.

[8] 苏红卫. OFDM 系统中降低峰平比问题的研究[D]. 无锡:江南大学,2008.

[9] 李晓明,段红梅,李强. 宽带无线通信的核心技术 OFDM [J]. 电信工程技术与标准化,2006,19(2):64-67.

[10] 郭建英. OFDM 关键技术及应用[J]. 山西电子技术,2011(2):60-61.

[11] Wang X, Tihuang T T, Ng C S. Reducing of peak to average power ratio of OFDM System using a companding technique [J]. IEEE Trans on Broadcasting,1999,45(3):303-307.

[12] 王林,江秀萍. 降低 OFDM 信号峰均比的 PTS 技术[J]. 计算机工程与应用,2007,43(5):153-158.

[13] Atallah J G, Ismail M. Future 4G front-ends enabling smooth vertical handovers [J]. IEEE Circuits & Devices Magazine, 2006,22(1):6-15.

[14] 田春阳,刘泉,梁学俊. 第四代移动通信 IP 网络关键技术[J]. 武汉理工大学学报(信息与管理工程版),2005,27(5):112-115.

[15] Natarajan N. On systems beyond 3G: requirements and approaches[C]//Proc of international conference on communication technology. [s.l.]:IEEE,2003:1305-1309.

[16] 姚成凤,葛万成. OFDM 原理及其在现代高速无线数据传输中的新应用[J]. 现代电视技术,2005(1):89-92.

(上接第 237 页)

2003,15(6):1373-1396.

[3] Xu Qianfang, Xiao Bo, Guo Jun. A mining algorithm with alarm association rules based on statistical correlation[J]. Journal of Beijing University of Posts and Telecommunications, 2007,30(1):66-70.

[4] Knorr E M, Ng R T, Tucakob V. Distance-based outliers in near linear time with randomization and a simple pruning rule [C]//Proc of SIGKDD. Washington, DC, USA: [s. n.], 2003.

[5] Hammer M A, Acar U A, Chen Yan. CEAL: a C-based language for self-adjusting computation[C]//Proceedings of the 2009 ACM SIGPLAN conference on programming language design and implementation. Dubin: ACM, 2009.

[6] Logothetis D, Trezzo C, Webb K, et al. In-situ MapReduce for log processing[C]//Proc of USENIXATC. [s. l.]: [s. n.], 2011.

[7] Olston C, Chiou G, Chitnis L, et al. Nova: continuous pig/Hadoop workflows [C]//Proc of SIGMOD. [s. l.]: [s. n.], 2011.

[8] Peng D, Dabek F. Large-scale incremental processing using distributed transactions and notifications[C]//Proc of OSDI. [s. l.]: [s. n.], 2010.

[9] Tan Pangning, Steinbach M, Kumar V. Introduction to data mining[M]. Beijing: Post & Telecom Press, 2006:403-405.

[10] Shekhar S, Lu C T, Zhang P. A unified approach to spatial outliers detection[J]. Geoinformatica, 2003,7(2):139-166.

[11] Shekhar S, Lu C T, Zhang P. Detecting graph-based spatial outliers[J]. International Journal of Intelligent Data Analysis, 2002,6(5):451-468.

[12] Lu Chang-Tien, Chen Dechang, Kou Yufeng. Algorithms for spatial outlier detection [C]//Proc of 3rd IEEE international conference on data mining. Blacksburg, VA, USA: IEEE, 2003:597-600.

[13] Lu Chang-Tien, Chen Dechang, Kou Yufeng. Detecting spatial outliers with multiple attributes [C]//Proc of the 15th IEEE international conference on tools with artificial intelligence. [s. l.]: IEEE, 2003:122-128.

[14] 文俊浩,吴中福,吴红艳. 空间孤立点检测[J]. 计算机科学, 2006,33(5):186-187.

社区发现在复杂网络划分中的应用

作者：[谭莹](#)，[张然](#)，[朱东生](#)，[TAN Ying](#)，[ZHANG Ran](#)，[ZHU Dong-sheng](#)

作者单位：[长沙理工大学 计算机与通信工程学院, 湖南 长沙, 410000](#)

刊名：[计算机技术与发展](#)

英文刊名：[Computer Technology and Development](#)

年，卷(期)：2014(11)

引用本文格式：[谭莹](#).[张然](#).[朱东生](#).[TAN Ying](#).[ZHANG Ran](#).[ZHU Dong-sheng](#) [社区发现在复杂网络划分中的应用](#)[期刊论文]-[计算机技术与发展](#) 2014(11)