

一种解决网络社区划分物理算法

王存睿^{1,2}, 段晓东¹, 刘向东¹, 李志洁¹

(1 大连民族学院 非线性信息技术研究所, 辽宁 大连 116600; 2 东北大学 研究生院, 辽宁 沈阳 110004)

摘 要: 网络结构挖掘目前是非规则数据的数据挖掘技术研究的热点之一, 其中网络结构划分在并行计算和互联网结构分析中具有重要的实用价值. 将电荷互斥和弹簧胡克定律引入网络结构划分, 用其构建了网络中节点间的各种作用力, 使得网络在力作用下产生相应运动, 然后, 在各种力的作用下达到平衡状态, 将此平衡状态映射到二维平面, 再将原先网络中的节点当作平面中的数据样本点. 这样可以将网络结构信息($n + m$) (其中 n 为网络中的节点个数, m 为网络边的个数) 压缩到 $2n$, 然后通过聚类算法对其进行划分. 构建此算法并用多种算法与传统的 GN 分裂算法进行对比, 发现新算法的划分质量与高效的 GN 算法相当, 新算法还可以通过节点代表对网络进行约减, 使得算法的速度可以根据精度调节. 本算法的构建为网络结构划分提供了一种新的途径, 同时对网络平衡状态的研究提出了相应的方法.

关键词: 复杂网络; 结构挖掘; 结构划分

中图分类号: N945

文献标识码: A

文章编号: 1000-7180(2010)09-0033-04

A Physical Community Discovery Algorithm

WANG Cun-rui^{1,2}, DUAN Xiao-dong¹, LIU Xiang-dong¹, LI Zhi-jie¹

(1 The Research Institute of Nonlinear Information Technology, Dalian Nationalities University, Dalian 116600, China;

2 The Graduate School, Northeastern University, Shenyang 110004, China)

Abstract: Community detection algorithm is very important to parallel computing and Internet structure analysis. This paper designs a new algorithm for Community Discovery based on building the various forces in network making the network move randomly by Hooke's Law and Coulomb's Law, then reach a balance by a variety of forces, this balance is used to map network to two-dimensional vector. And then the nodes of original network become the data points in 2d plane. This way can compress ($n + m$) network structure information down to $2n$ (Here n is the number of nodes in the network, m is the number of edges in the network), and then make cluster algorithm to carry out the division of these 2d data points. In this paper, the algorithm is compared with the traditional GN algorithm, the experiment indicates that new algorithm can match the efficient GN algorithm, and can get more efficient and accurate by selecting some sample of data points in the plane. The algorithm provides a new way to partition network and takes a new way to study the balance of network.

Key words: complex network; structure mining; community detection

1 引言

网络结构挖掘目前是非规则数据的数据挖掘技术研究的热点之一^[1], 网络结构诸如 Internet、代谢网络等, 其中蕴含有价值的信息. 而网络信息挖掘

中, 网络中的结构团识别是其重要的研究内容之一, 所谓结构团是指网络中的子网络内部比较致密, 子网络间链接比较松散的网络结构. 构建高效的算法来分析网络中结构团对于防备黑客攻击、提高搜索引擎质量、防治流行病和开发新药等, 都具有重要的

收稿日期: 2009-07-06; 修回日期: 2009-09-06

基金项目: 国家自然科学基金项目(60573124); 辽宁省自然科学基金(20072197); 高校科研计划项目(20060146)

意义^[2-5].

结构团划分方法著名的算法包括 Kernighan 算法、基于 Laplace 图特征值的谱平分法等算法. Kernighan 算法是一种基于贪婪算法原理^[2], 将网络分割为两个大小已知的社区的二分法, 但该算法必须已知网络社区的确切规模才能得以应用. 基于 Laplace 图特征值的谱平分法, 其利用网络结构的 Laplace 矩阵中不为零的特征值所对应的特征向量, 和同一个社区内的结点对应的元素近似相等的原理对网络社区进行划分, 在规模为 n 节点的网络中, 该算法的算法复杂度等同于求 $n \times n$ 矩阵的特征向量的复杂度为 $O(n^3)$. Wu 和 Huberman 提出了基于电阻网络电压谱的快速分割算法^[3], 它将网络看成一个电阻网络, 然后利用 Kirchhoff 定理求解各个节点的电压值, 绘制电压谱, 进而根据不同的阈值进行社区划分, 该算法的复杂度只有 $O(n+m)$, 其中 n 表示网络节点数目, m 为网络中边的数量, 但该方法须已知分属于不同结构团的两个节点.

除上述算法之外, 还有利用分级聚类概念提出的分裂算法和凝聚算法. 其中的 GN 算法^[4]通过不断地从网络中移除介数(网络中经过每条边的最短路径的数目)最大的边, 逐步将整个网络分成越来越小的各部分. GN 算法的不足在于对网络结构团的划分优劣没有一个定量描述. 因此, Newman 等人经过研究提出了一种度量网络结构团划分质量的标准, 成功地解决了这个问题^[6]. Radicchi 等人在 GN 算法基础上提出快速分裂算法^[7], 通过计算局部变量, 减小了全局的运算量, 但该算法本身很大程度依赖于网络中存在的三角形的数量. 网络如果存在的三角形的数目很少, 那么相应边的聚类系数就会很小, 此算法就无法正确地识别网络结构团. Duch J 提出了一种网络结构划分的极值优化算法^[8], 其通过调整局部极值来优化全局最优的变量, 进而提高算法效率.

如果赋予网络以物理特性, 使得网络在各种力的作用下产生平衡状态, 这个状态能否体现网络中的结构团呢? 本文将自组织网络运动引入到结构团的发现中, 通过自组织行为对网络结构进行划分, 构建一种能够解决大规模网络划分问题的新方法.

2 基于作用力的网络结构自组织运动

网络结构能否在运动中达到相对平衡状态? 本文将电荷互斥定律和弹簧的胡克定律^[9]引入网络, 假定网络中的各个节点两两互斥, 产生排斥力, 网络

中的边作为具有最小长度的理想弹簧.

定义 1 网络节点之间的斥力.

对于网络中的任意两个节点 i 和 j , 其在平面上对应的坐标分别为 (x_i, y_i) 和 (x_j, y_j) , 假设网络中的各个节点的质量相同, 则他们的斥力为 $F_{ij} = k / ((x_i - x_j)^2 + (y_i - y_j)^2)$, 将其分解为 x 和 y 两个方向上的作用力.

则节点 i 的 x 坐标受到的节点 j 的斥力为

$$F_{ij}^x = \frac{k(x_i - x_j)}{[(x_i - x_j)^2 + (y_i - y_j)^2]^{\frac{3}{2}}} \quad (1)$$

节点 i 的 y 坐标受到的节点 j 的斥力为

$$F_{ij}^y = \frac{k(y_i - y_j)}{[(x_i - x_j)^2 + (y_i - y_j)^2]^{\frac{3}{2}}} \quad (2)$$

式中, k 为斥力常数, 主要调节斥力与边的弹性拉力之间的平衡. 上述斥力的构建主要参照库仑定律, 力的大小与距离的平方成反比. 在本文定义的虚拟空间中, 各个网络节点是相互排斥的, 而且质量相同. 网络结构中的节点受到彼此之间的斥力会彼此分开, 在空间无限扩展. 其中, 网络中的个体还受到网络边的拉力的作用, 这里采用胡克定律来定义节点之间的受力情况.

定义 2 网络节点之间的弹簧边的力作用.

假定边弹簧长度为 x , 其自然长度为 S , 其连接节点 i 和 j , 则节点 i 和 j 受到的拉力为

$$F_{ij} = m(x - S) \quad (3)$$

式中, $x = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$, 对于节点 i 的 x 坐标受到边弹簧的力为

$$F_{ij}^x = m(1 - \frac{S}{x})(x_j - x_i) \quad (4)$$

节点 i 的 y 坐标受到的边弹簧的力为

$$F_{ij}^y = m(1 - \frac{S}{x})(y_j - y_i) \quad (5)$$

3 算法过程

网络可以在动态运动中达到平衡状态, 这需要弹簧的伸缩保持一定的惯性, 其次是在节点运动中增加随机扰动, 使得网络处于振荡状态. 网络平衡状态由网络边的长度和进行度量. 下面介绍算法详细过程.

算法 1 二维平面的网络动态运动算法

Step 1: 设定弹簧的原始长度 S , 弹簧的劲度系数 m , 斥力系数 k ;

Step 2: 在给定范围内, 随机初始化每个网络节点的位置;

Step 3: 计算每个弹簧边的状态, 判断其收缩趋

势或扩张趋势,然后对其连接的两个节点的位置在趋势下进行随机扰动调整。

Step 4:遍历每个节点,计算每个节点的受力情况,根据不同方向上的受力大小对于节点的位置进行调整,调整过程中,对于每个节点 i 有一个描述惯性的速度 (v_{xi}, v_{yi}) ,节点的位置移动首先影响该节点的速度量,然后在其原有位置上进行移动.如式(6)、式(7)所示,对网络中的各个节点的 x 轴位置进行调整,其中, l_1 和 l_2 为各分力的系数, w 为惯性因子, y 轴的位置调整与此相同进行调整。

$$v_{xi} = wv_{xi} + l_1 F_{x\text{边弹性拉力}} + l_2 F_{x\text{斥力}} \quad (6)$$

$$x_{xi} = x_{xi} + v_{xi} \quad (7)$$

Step 5:计算所有边的长度和为 P ;

Step 6:判断停止条件:在 n 次迭代中其变动幅度小于某个阈值 θ ;如果满足则跳转到 Step7,如果不满足跳转到 Step3;

Step 7:输出节点的编号和横纵坐标;

如图1所示为网络在不同时间步的变化,从图1中可以看出,网络初始状态时聚成一团,然后,在各种力的作用下产生运动,各网络中的聚类团在各种力的作用下产生运动,同时一个聚类团对其他聚类团形成斥力合力,使得不同聚类团在运动中彼此分离,当网络在运动中动态平衡时,边总和振幅减小,并呈周期性变化。

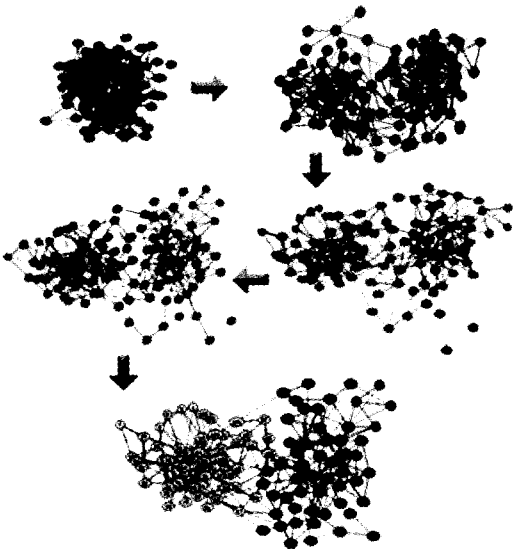


图1 网络在不同时间步力作用下的变化情况

4 实验及结果分析

本文利用该算法对不同规模的网络结构数据进行测试,规模分别为100、200、300、400个节点,如图

2所示为网络边长度和的变化曲线,可以发现100和200个节点的网络数据在17~18次迭代的时候就已经达到稳定状态,300和400个节点的网络在3541次迭代左右进入稳态,可以看出进入稳态的时间随网络规模增大呈线性增长。

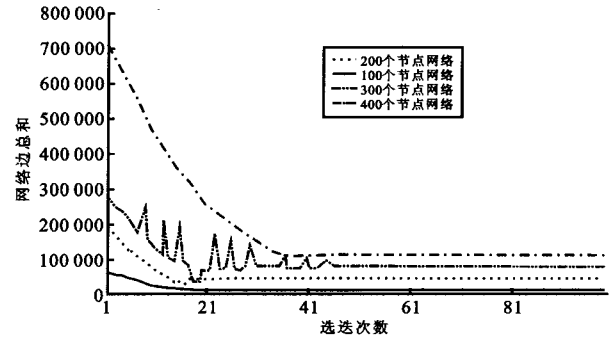


图2 不同规模网络边的总和的变化情况

通过在网络中构建力作用,使得网络达到稳态,进而将网络数据映射为平面上的一组二元的点数据.通过这种映射,可直观看出网络中的不同结构团在网络中的状态,为更为精确地对网络结构团进行划分,可以采用聚类算法对其进行快速分析,由于这种二元映射中边弹簧全部相同,所以形成的二元映射在平面中的密度比较均匀,因此,传统的 K -Means 和 K -Medoid^[10] 就可以对其进行分析,密度聚类算法如 DBSCAN^[11] 等也可对其进行分析.如果网络的规模较大,可以通过将平面划分为相应区域,同区域的点用中心点替代来进行约减,进而降低计算复杂度。

本文采用 Newman 提出的一个函数对网络结构团划分优劣进行度量^[6]. 假设其将网络划分为 k 个团,定义一个 $k \times k$ 的对称矩阵 $E = (e_{ij})$,其中元素 e_{ij} 表示网络中连接 i 团和 j 团的边的数量占所有边的比例.这里所说的边是指在原始网络中的,是利用完整的原始网络计算的,而非在计算过程中被算法破坏的网络.设矩阵中对角线上的各元素之和为 $\text{Trace}(E) = \sum_i e_{ii}$,它给出了网络中连接某一个团内部各节点的边在所有边的数目中所占比例.定义每行中各元素之和为 $a_i = \sum_j e_{ij}$,它表示与第 i 个团中的节点相连的边在所有边中所占的比例.在此基础上,用式(8)来定义网络结构团划分的衡量标准。

$$Q(y) = \sum (e_{ij} - a_i^2) = \text{Trace}(E) - \|E^2\| \quad (8)$$

$\|E^2\|$ 表示矩阵 e^2 中所有元素之和,该度量标准为结构划分质量的全局度量函数.我们比较了4

种算法对网络的聚类团的发现,并将其与其他算法进行对比,如表 1 所示.

表 1 该算法对随机网络结构团划分的实验结果

节点数 参数	边数 Cluster = 2	k -means $k = 2$	k -medoid $k = 2$	DBSCAN $r = 11.5$	GN 算法 $g = 3$
100	342	0.450 8	0.450 3	0.421 2	0.452 1
200	656	0.431 5	0.429 0	0.431 0	0.438 0
300	1 215	0.440 1	0.458 6	0.398 5	0.422 1
400	897	0.434 5	0.424 5	0.429 7	0.439 3
500	941	0.381 0	0.363 5	0.331 1	0.360 1
600	1 807	0.454 9	0.444 9	0.361 0	0.448 0
700	1 762	0.483 1	0.448 2	0.391 8	0.453 6
800	1 888	0.468 6	0.457 2	0.418 5	0.451 7
900	2 264	0.451 0	0.460 1	0.421 7	0.434 9
1 000	2 825	0.478 5	0.481 0	0.432 9	0.482 1

本文利用 3 个聚类算法与传统的网络划分方法比较,算法参数如表 2 所示,迭代次数 1 000 次,每个算法运行 10 次取平均值,其中, K-Means 和 K-Medoid 需要指定聚类数设为 2,基于密度聚类的 DBSCAN 算法的对象半径为 11.5,核心对象为 5. 与传统网络结构划分算法中的 GN 快速分裂算法(多边形边数 3)进行比较.实验结果可以看出除基于密度聚类算法对网络进行算法划分结果较差外, K-Means 与 K-Medoid 与传统的 GN 分裂算法相当, K-Medoid 在样本平均值略优于 K-Means 算法.分析其中的原因,主要是在映射过程中,网络间的距离过于接近,而且很多连接结构团之间的中间节点也较多,所以基于密度聚类的方法的效果较差.由此可见,此种映射结合相应的聚类算法可以达到与传统算法相当的结果.从处理规模上看,此映射可以对网络的规模进行约减,通过在图上利用代表点来对网络进行简化,进而降低复杂度.

5 结束语

从实验结果可以看出,该网络结构团构建算法是有效的,可以用于网络结构团划分算法的相应研究,为其提供更多不同参数下的测试数据.本文通过构建网络中的各种作用力,使得网络产生随机运动,在各种力的作用下达到平衡状态,并将此平衡状态映射到二维平面,然后将原先网络中的节点当作平面中的数据样本点.这样可以将网络结构信息($n + m$)(其中 n 为网络中的节点个数, m 为网络边的个数)压缩到 $2n$,然后通过聚类算法对其进行划分.本文构建此算法并用多种算法与传统的 GN 分裂算法进行对比,发现新算法的划分质量与高效的 GN 算

法相当,新算法还可以通过节点代表对网络进行约减,使得算法的速度可以根据精度进行调节.本算法的构建为网络结构划分提供了一种新的途径,同时对网络平衡状态的研究提出了相应的方法.

参考文献:

- [1] 杨楠,弓丹志,李欣,等. Web 社区发现技术综述[J]. 计算机研究与发展, 2005(42):439-447.
- [2] Kernighan B W, Lin S. An efficient heuristic procedure for partitioning graphs[J]. Bell System Technical Journal, 1970(49):291-307.
- [3] Wu F, Huberman B A. Finding communities in linear time: a physics approach[J]. Euro. Phys. J B, 2003 (38):331-338.
- [4] Girvan M, Newman M E J. Community structure in social and biological networks[J]. Natl Acad, 2002, 99(7):821-7826.
- [5] 石新景,汤小春,闫磊. 基于孟然论潜在语义分析的 Web 使用挖掘[J]. 微电子学与计算机, 2008, 25(6):225-229.
- [6] Newman M E J, Girvan M. Finding and evaluating community structure in networks[J]. Physical Review E, 2004, 69(2):026113.
- [7] Radicchi F, Castellano C, Cecconi F. Defining and identifying communities in networks[J]. National Academy of Science, 2004, 101(9):2658-2663.
- [8] Duch J, Arenas A. Community detection in complex networks using extremal optimization[J]. Physical Review E, 2005, 72(2):027104.
- [9] 陈治,陈祖刚,刘志刚. 大学物理[M]. 北京:清华大学出版社, 2007:57-131.
- [10] Han J W, Kamber M. Data mining: concepts and techniques. [M]. 2nd ed. San Francisco: Morgan Kaufmann Publishers, 2001:223-250.
- [11] Ester M, Krieger H P, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]// Proc. of the 2nd Int'l Conf. on Knowledge Discovery and Data Mining. Portland: AAAI Press, 1996: 226-231.

作者简介:

王存睿 男, (1980-), 博士研究生, 讲师. 研究方向为智能优化和数据挖掘.

段晓东 男, (1963-), 博士, 教授. 研究方向为非线性信息处理技术.

刘向东 男, 博士, 教授. 研究方向为非线性信息与控制技术.

李志洁 女, 博士, 讲师. 研究方向为网络计算.