

一种基于正态分布的复杂网络结构划分算法

段忠祥

(广西工商职业技术学院, 广西 南宁 530007)

摘 要: 复杂网络的节点聚集呈现符合社区结构的动态、无标度和非对称的特性, 为了优化复杂网络的社区结构, 研究当前发现和优化社区结构的方法的不足, 研究用约束正态分布来改进社区结构的节点聚集归属方法, 借助信息熵, 提出了基于正太分布的复杂网络结构划分算法, 通过算法得出聚集节点的正态分布概率, 用正太分布概率作为信息熵的输入, 重新调整信息熵的变化, 根据信息熵变化的幅度, 确定节点的划分归属。本算法在确定网络社区结构划分的同时, 也能够确定社区内节点的模糊关系。

关键词: 正太分布; 复杂网络; 社区结构; 结构精简; 优化算法

中图分类号: TP393 **文献标识码:** A

A Complex Network Structural Division Algorithm Based on Normal Distribution

DUAN Zhongxiang

(Guangxi Vocational College Technology and Business, Nanning 530007, China)

Abstract: The node aggregation of complex networks presents the dynamic, fuzzy and asymmetrical characteristics of the community structure. In order to optimize the community structure of complex networks, the paper studies the current methods of discovering and optimizing community structures, and the node aggregation method to improve the community structure by constraining normal distribution. Based on information entropy, the paper proposes a complex network structure partitioning algorithm based on positive distribution. The normal distribution probability of the aggregation node is obtained by the algorithm. The positive distribution probability is used as the input of information entropy to re-adjust the information entropy. According to the magnitude of the information entropy change, the division of the node is determined. The algorithm can determine the fuzzy relationship of nodes in the community while determining the division of the network community structure.

Keywords: normal distribution; complex network; community structure; structure simplification; optimization algorithm

1 引言(Introduction)

复杂网络中社区发现的研究是当前国内外研究的热点之一, 社区结构作为复杂网络重要的特性, 对认识节点内部关联、信息传播、兴趣点推荐等具有重要的研究价值。复杂网络社区检测是指从复杂网络中挖掘出有实际意义的模块或层次结构的过程, 能够提取出网络拓扑结构特征, 有助于理解和分析网络的拓扑特性、功能特性及动力学特性, 挖掘出复杂网络中蕴含的属性关联性等深层次信息并对网络行为进行预测^[1]。

目前研究人员对复杂网络社区结构的划分和检测进行了大量的研究: 文献[2]提出一种基于k-派系渗透的动态网络社区检测算法, 该算法需要以较长的检测时间作为代价实现重叠社区的检测; 文献[3]提出一种探寻复杂网络最短路径的近似算法; 文献[4]提出一个融合贝叶斯概率的社区结构发现方法研究; 文献[5]提出动态社区的点增量发现算法; 文献[6]提

出基于单目标PSO的社区检测算法; 文献[7]提出节点相似度感知的DTN交叠社区结构检测机制; 文献[8]提出局部优先的社会网络社区结构检测算法; 文献[9]提出社交网络中基于近似因子的自适应社区检测算法。文献[10]提出一种基于常量因子近似的动态社区检测算法, 但该算法需要确定明确量化的惩罚成本, 而复杂动态网络多数情况无法预知该惩罚成本, 因此, 该算法对现实复杂网络不具有可行性。

针对上述问题, 本文提出一种基于正太分布的复杂网络结构划分算法(A Complex Network Structural division Algorithm Based on Normal Distribution, CSAN)并利用具有社区结构的已知网络现实数据检验算法的有效性。本文算法的特点是: ①算法从中心 μ 开始扩展, 通过距离密度阈值的探测, 探查社区结构, 实现社区结构的扩展; ②算法形成有效社区结构划分的同时, 也能够把社区结构中远离社区中心 μ 的节点重新归类, 减轻社区结构的节点冗余, 提高社

区结构的工作效率。

2 问题建模(Problem modeling)

复杂网络可以用一个无向图 $G=(V, E)$ 表示, 其中, 节点数量为 $N=|V|$, 复杂网络中节点间连接边的数量为 $M=|E|$ 。节点间连接边的邻接矩阵 A 表示为 $A=(A_{ij})$, 如果节点 i 和 j 之间有一条边连接, 则 $A_{ij}=1$, 否则 $A_{ij}=0$ 。用 $C=(C_1, C_2, \dots, C_k)$ 表示复杂网络中的社区, $CS_i(i=1, 2, \dots, k)$ 表示 i 社区的社区结构。表1列出了本文涉及的符号和各自代表的含义。

表1 符号及代表的含义

Tab.1 Symbols and meanings

符号	含义
V	网络节点集合
E	网络节点间连接边集合
N	网络节点数
M	网络节点间连接边数
dv	网络节点 v 的度
μ	社区的中心位置
σ	社区探测尺度
$C(v)$	节点 v 所在的社区
$N(v)$	与节点 v 有连接边的定
X_v	点集合
	节点 v 的位置

2.1 信息熵

信息熵^[11]是信息论中用于度量信息量的一个概念, 它是有明确定义的科学名词, 不随信息的具体表达形式的变化而改变, 是抽象的概念。香农在1948年第一次提出“信息熵(Information Entropy)”的概念, 在香农的概念中, 信息是消除或减少不确定性的东西。获得信息, 意味着不确定性的减少。不确定性的减少导致“熵”减小, 即信息熵越小, 表示事件的不确定性越小, 则包含的信息量越小; 反之则大。信息熵的计算公式如下:

$$H(X)=-\sum P(x)\log_2 P(x) \quad (1)$$

其中, $x \in X$, X 表示一个离散的随机变量, $P(x)$ 表示一个概率密度函数, $-\log_2 P(x)$ 表示在状态 x 下所含信息量或者不确定性数量。

2.2 正态分布

正态分布是数理统计中非常重要的一种概率分布, 在解决与中心点距离有关的概率分布具有很广泛的应用。正态分布的概率密度函数为:

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad (2)$$

其中, $x=\mu$ 为概率密度函数的对称轴, $x=\mu$ 的位置为概率密度最大的位置, σ 与概率密度函数形状有关。

2.3 信息熵值探测模型

由式(1)和式(2)来确定信息熵值探测模型:

$$f(x)=\begin{cases} H(X)=-\sum P(x)\log_2 P(x) \\ \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \\ P(x)=f(x) \end{cases} \quad (3)$$

社区结构的节点聚集具有距离靠近聚集的大概率^[12], 为此我们通过探测社区 C 中节点 v 在正太分布概率密度下的信息熵

$H(X_v)$ 的值来确定初始社区的第一个节点, 以此节点为基准, 根据探测的尺度发现新成员, 最终形成一个网络社区。

3 基于正太分布的复杂网络结构划分算法(Complex network structure division algorithm based on normal distribution)

首先, 要确定复杂网络社区的探测“中心点”, 依据中心点计算概率密度覆盖区域内节点的信息上值, 通过信息熵值与阈值之间的对比关系来进一步确定进入社区的节点, 从而构建探测和优化社区结构。

算法1 探测复杂网络的社区“中心点”

Begin

STEP1: 初始化 X_v , $X_v=(x_0, x_1, \dots, x_{N-1})$

STEP2: $\mu \leftarrow x_0$

STEP3: $\sigma \leftarrow \sigma_0$

STEP4: 确定初始节点 x_0 周围需要探测信息熵值得节点。探测距离由两个参数决定的函数 $f(\sigma_0, x_0)$ 确定。If $x_i \in X$ and $|x_i - \mu| < f(\sigma_0, x_0)$ then $x_i \in C^1(x_0)$

STEP5: 通过(2-3)计算初始节点的信息熵 $H(x_0)$

STEP6: for($i=1$; ; $i++$){ $x_0=x_i$; $\mu=x_i$; $H(x_i)$;

if($H(x_i)*|x_i - \mu| < H(x_0)*|x_0 - \mu|$), then $CC_0=x_i$ }

STEP7: 确定社区节点聚集的中心节点序列 $CC_0, CC_1, CC_3, \dots$

End

算法2基于正态分布的复杂网络社区结构优化(CSAN)

Begin

STEP1: $n=0$ (社区的个数)。

STEP2: $k=0$ (用来遍历可能进入社区的有效节点计数变量初始化)。

STEP3: 建立社区节点栈 Si 。

STEP4: $CC_0 \rightarrow$ 压入栈。

STEP5: 根据算法1, 确定对应 CC_0 的近距离节点集合 NCC_0 。

STEP6: 计算 $|H(CC_0)-H(NCC_0_k)|$; if($|H(CC_0)-H(NCC_0_k)| < \varepsilon$), then NCC_0_k 点进入社区, 压入栈, $k++$; else NCC_0_k 点不进入社区, 弹出栈, NCC_0_k 进入 $LCC_k(CC_k$ 对应的临近社区节点集)。

STEP7: $n++$ 。

STEP8: 遍历所有的社区中心点, 得到初步的社区结构。

End

在算法2的第6步会产生现有复杂网络的一些孤立节点, 这些节点的集合 $R=\sum LCC_k(k=0, 1, 2, \dots)$ 。集合 R 的规模决定了复杂网络在基于正太分布的社区发现过程中, 剔除冗余节点的比例 R/N , 对于规模庞大的复杂网络来说, 具有一定的优化作用。

4 验证(Experimental verification)

通过实验的方式检验上述算法在复杂网络社区结构发现和优化中的可行行, 算法可以发现不能划分到社区的节点, 从而展示算法的有效性。

实验数据来源于实验机房随机选定的8台计算机, 每台计算机都能够通过级联的交换机互联, 其中编号0—3的计算机连接在1号交换机, 4—7号计算机连接在2号交换机。利用算法2计算节点之间的熵值关系矩阵如表2所示。矩阵的值表示

有序节点之间的连接概率, 矩阵值越大, 反映出节点之间的同社区可能性越大, 根据设定的阈值 $\epsilon=0.5$, 通过算法的执行, 得到节点逐步优化分割的社区结构。

通过阈值的调整, 可以进一步改进社区结构的节点聚集关系, 从而可以实现复杂网络的优化升级。

表2 节点间的扩展熵值关系

Tab.2 Extended entropy relationship between nodes

节点	C0	C1	C2	C3	C4	C5	C6	C7
C0	1	0.88	0.49	0.91	0.48	0.40	0.37	0.33
C1	0.48	1	0.40	0.95	0.22	0.43	0.45	0.33
C2	0.82	0.45	1	0.41	0.33	0.36	0.27	0.39
C3	0.80	0.92	0.66	1	0.36	0.47	0.29	0.17
C4	0.11	0.42	0.39	0.41	1	0.79	0.82	0.41
C5	0.36	0.29	0.41	0.47	0.49	1	0.93	0.36
C6	0.45	0.38	0.44	0.49	0.77	0.42	1	0.19
C7	0.33	0.29	0.45	0.41	0.39	0.42	0.37	1

算法执行过程中, 对应的节点变化走势如图1所示。

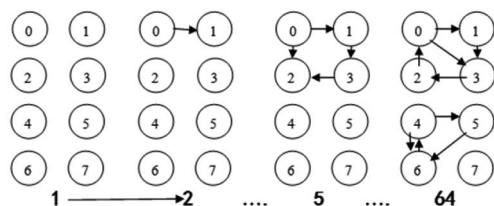


图1 节点聚集成社区结构的过程

Fig.1 The process of nodes gathering into a community structure

从算法执行的过程, 能够看出社区结构的形成是符合实际的有价值的结构, 对于信息熵值低于阈值的节点将从不隶属于任何一个社区结构, 这些点从复杂网络中独立划分, 单独处理。本算法的时间复杂度为 $O(N^2)$, 随着网络规模的急剧增长, 算法的计算时间曲线斜率线性增长, 耗费大量的系统时间。算法的计算时间同网络规模之间的关系如图2所示。

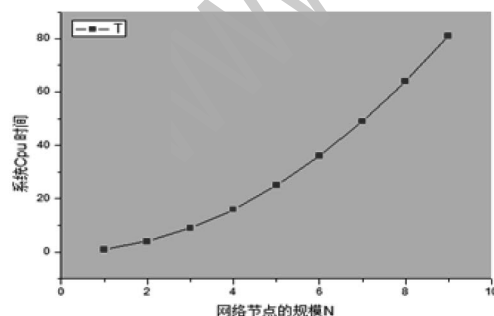


图2 算法执行的系统时间随N的变化

Fig.2 The system time of the algorithm execution varies with N

5 结论(Conclusion)

复杂网络的节点聚集呈现符合社区结构的动态、无标度和非对称的特性, 为了优化复杂网络的社区结构, 本文提出的提出一种基于正态分布的复杂网络结构优化算法(CSAN), 本算法既能发现并重构社区结构, 也能发现社区内节点之间

的动态模糊指向关系, 采用了信息熵的变化来发掘社区节点, 同时, 能够发掘网络中的信息相对孤立的冗余节点, 这些节点不划归任何一个社区, 从而优化了社区结构, 精简了社区规模, 这是本研究的创新之处。现实中, 复杂网络节点之间的关系复杂, 存在多种关系形态, 如关联关系, 因果关系, 相反关系等, 深入研究这种关系对于复杂网络的优化具有十分重要的意义。本研究只用简单的局域网进行了检验, 检验算法的数据有限, 在实际复杂网络环境下本算法的具体性能研究还需要做进一步的检验工作。对根据信息熵的变化和社区结构具有位置相对靠近聚集的特点, 采用正态分布概率密度作为信息熵值估计, 从而判断节点是否属于一个社区, 熵阈值确定很重要, 如何确定更好的阈值, 使得发现的社区结构有更好的可用性, 这也是目前所有社区发现方法面临的问题, 也是下一步的重点研究方向。本研究所涉算法的复杂度偏高, 如何降低社区结构发现和优化的算法时间复杂度也是下一步研究的方向。

参考文献(References)

- [1] Qiao Shao-Jie, Guo Jun, Han Nan, et al. Parallel Algorithm for Discovering Communities in Large-Scale Complex Networks[J]. Chinese Journal of Computers, 2017, 3(40): 687-700.
- [2] Roy S D, Lotan G, Zeng Wenjun. The Attention Automaton: Sensing Collective User Interests in Social Network Communities[J]. IEEE Transactions on Network Science and Engineering, 2015, 2(1): 40-52.
- [3] TANG Jin-Tao, WANG Ting, WANG Ji. Shortest Path Approximate Algorithm for Complex Network Analysis[J]. Journal of Software, 2011, 22(10): 2279-2290.
- [4] 王刚. 一个融合贝叶斯概率的社区结构发现方法研究[J]. 计算机技术与发展, 2018(12): 1-5.
- [5] 顾炎, 熊超. 动态社区的点增量发现算法[J]. 计算机技术与发展, 2017, 27(6): 81-85.
- [6] 杨令兴, 张喜斌. 基于单目标 PSO 的社区检测算法[J]. 计算机科学, 2015, 42(6A): 57-60.
- [7] 王峰, 王于波, 朱承治. 节点相似度感知的 DTN 交叠社区结构检测机制[J]. 微电子学与计算机, 2017, 34(12): 74-78.
- [8] 李春英, 汤志康, 汤庸, 等. 局部优先的社会网络社区结构检测算法[J]. 计算机科学与探索, 2018, 12(08): 1263-1274.
- [9] 阙建华. 社交网络中基于近似因子的自适应社区检测算法[J]. 计算机工程, 2016(05): 134-138.
- [10] Tantipathananandh C, Berger-Wolf T. Constant-factor Approximation Algorithms for Identifying Dynamic Communities[C]. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, June 28-July 1, 2009: 827-836.
- [11] Peng CG, Ding HF, Zhu YJ, et al. Information entropy models and privacy metrics methods for privacy protection[J]. Journal of Software, 2016, 27(8): 1891-1903.
- [12] 李兆南, 杨博, 刘大有. 复杂网络社区挖掘的距离相似度算法[J]. 计算机科学与探索, 2011, 5(4): 336-344.

作者简介:

段忠祥(1978-), 男, 硕士, 高级实验师. 研究领域: 软件设计, 计算机应用技术, 高职教育.