

基于引力因子的加权网络重叠社区识别算法

刘冰玉 王翠荣 王 聪 苑 迎

(东北大学信息科学与工程学院 沈阳 110819)

摘 要 通过挖掘大数据来识别复杂社会网络上的社区,有利于对经济、政治、人口等方面的重要问题进行定量研究,社区的识别算法已经成为当前研究的热点问题。重点研究了重叠社区识别问题,提出了基于引力因子的加权复杂网络的重叠社区识别算法 GWCR。该算法首先选取万有引力因子大的节点为中心节点,将节点与中心节点之间的引力因子作为衡量标准,并将节点归入社区引力因子大于某一阈值的社区,最后通过识别重叠节点来识别重叠社区。在3个真实网络数据集上的实验结果表明,与传统的重叠社区识别算法相比,GWCR 算法划分的社区的模块度较高。

关键词 引力因子,社区识别,加权网络,重叠社区

中图法分类号 TP393.02 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.12.027

Overlapping Community Recognition Algorithm of Weighted Networks Based on Gravity Factor

LIU Bing-yu WANG Cui-rong WANG Cong YUAN Ying

(School of Information & Engineering, Northeastern University, Shenyang 110819, China)

Abstract The recognition of community in complex social networks by mining big data can favor the quantitative research for economic, political and demographic problems. Community recognition algorithms have become a hot topic of current research. This paper focused on the research of overlapping community discovery, and proposed the overlapping community detection algorithm GWCR, which is based on gravity factor of weighted networks. Firstly, the GWCR algorithm selects the node with the largest gravitation factor as the center node, and uses the gravitation factor between one node and the central node as a measure. The node whose gravitation factor is larger than the threshold will be included in the community. Finally, overlapping communities are discovered by identifying overlapping nodes. Experimental results on three real network datasets show that, compared with conventional overlapping community detection algorithm, GWCR has higher modularity value.

Keywords Gravity factor, Community recognition, Weighted network, Overlapping community

现实世界中,许多系统都呈现为复杂网络的形式,如人际关系网、蛋白质相互作用网、科学家协作网。这些网络一般可分为4类:生物网络、技术网络、社会网络和信息网络。这些网络都具有一些基本的统计特性:小世界特性,网络顶点间的平均最短路径相对网络的规模而言极小;无标度特性,复杂网络中的许多节点只有很少的邻接点,与此同时,一些节点却拥有大量的邻接点,从整体上看,节点的度符合幂律分布。

社区结构是复杂网络的结构特性之一,也是网络节点的组织方式。社区^[1]是指具有关联关系的个体所组成的集合。整个复杂网络由若干个社区组成,社区内部节点间的连接紧密,而不同社区之间的节点连接却比较稀疏。例如,社会网络中的社区是指根据兴趣或背景形成的社会群体。理解社区结构具有很多潜在的应用前景,可以提高网络的搜索速度、搜索的精确性和搜索引擎的命中率等。而在现实世界的复杂网络中,有些节点同时属于多个社区,这些节点称为重叠节点,包含重叠节点的社区为重叠社区。对复杂网络进行重叠社区识别更贴近真实世界。

本文借鉴物理学中的万有引力定律来解决加权社会网络

中的重叠社区识别问题,提出了基于引力因子的加权网络重叠社区识别算法,即 GWCR (Gravity Weighted network Community Recognition) 算法。GWCR 算法将节点在网络中的重要程度映射为重力学中质的概念,首先定义节点的质,将质大的节点作为初始节点,通过衡量节点与社区之间的引力因子大小来判断节点的归属,再通过识别重叠节点来识别重叠社区。

1 相关工作

目前大部分的重叠社区识别算法主要针对无权网络,算法主要分为3类:基于图论的重叠社区识别算法、基于优化的重叠社区识别算法以及基于链路的重叠社区识别算法。基于图论的算法主要包括是由 Palla 提出的 CPM (Clique Percolation Method) 算法^[2]。该算法的基本思想是社区内部的节点由于具有较高的密度而更容易形成派系。此外,文献[3]提出了一种重叠社区挖掘的两段策略算法:初始社区抽取与社区合并;基于优化的算法都是对社区的基本度量标准(如密度、模块度等)进行优化。其中 OSLOM (Order Statistics Local

到稿日期:2015-12-08 返修日期:2016-04-23 本文受国家自然科学基金(61300195),河北省自然科学基金(F2014501078, F2016501079),河北省科技计划项目(15210146),辽宁省教育厅科学研究一般项目(L2013099),秦皇岛市科技计划项目(201401A028)资助。

刘冰玉(1978—),女,博士生,讲师,主要研究方向为数据挖掘、复杂网络;王翠荣(1963—),女,博士,教授,主要研究方向为下一代网络、数据挖掘;王 聪(1981—),男,博士,主要研究方向为网络虚拟化、复杂网络;苑 迎(1981—),女,博士,讲师,主要研究方向为网络虚拟化。

Optimization Method)^[4]算法通过增加或删除社区内的节点来增强社区的适应度函数;最著名的是 Baumes^[5]提出的一种两阶段方法,即首先将网络划分为不相交的“种子”社区,然后通过增加或移除某些邻接节点来达到社区的“密度”最大的效果;基于链路的重叠社区识别算法将原网络图转换为边图,原网络图中的边映射为边图中的节点,原网络图中边与边之间的邻接关系映射为边图中节点之间的链接。其中 Kim 等人^[6]提出了基于信息论编码的边社区识别算法,其基本思想是对网络结构进行编码并求出最小的平均编码长度对应的社区划分。

以上算法的研究对象主要是无权网络,加权网络的复杂性远高于无权网络,学者一般认为加权网络的社区识别是一个比较难的问题。目前已有的文献仅对加权网络的建模与分析等做了初步研究,文献[7]对加权网络的评价指标及相关算法进行了分析对比,文献[8]提出的基于同步的社区识别算法可以有效地识别加权网络的非重叠社区。

2 基于引力因子的重叠社区识别算法

2.1 重叠社区识别问题的提出

现实世界中许多网络都可以映射为加权网络,如在社会网络中边的权重可以反映成员之间联系的紧密程度,食物链网络中边的权重可以反映捕食与被捕食者之间能量的传递,信息网络的边权重可以反映两个节点之间带宽的大小。加权网络能够更贴切地反映真实网络结构,加权网络的社区识别有助于更进一步理解网络的结构。本文主要研究无向加权网络。

牛顿的普适万有引力定律为:任意两个质点通过连心线方向上的力相互吸引。该引力的大小与它们质量的乘积成正比,与它们的距离成反比。复杂网络中节点重要性用于度量节点在整个网络中的影响力^[9],节点的影响力与节点在网络中的重要程度成正比,与节点之间的距离成反比。借鉴万有引力定律,本文将节点在网络中的重要程度定义为网络中节点的质,将节点之间最短路径的长度作为网络中节点之间的距离。

与无权网络类似,加权网络可以抽象为邻接矩阵。不同的是,加权网络的邻接矩阵的元素不只是 0 或 1,而是节点之间连边的权重,文中假设所有边的权重都是非负数。

设网络 $G=(V,E)$, V 代表网络中的节点, E 为网络中的边。网络 G 的邻接矩阵为 W ,如果节点 i 与节点 j 之间有边直接相连,则 w_{ij} 的值为节点 i 与节点 j 之间路径的权重,否则 $w_{ij}=0$ 。

2.2 算法定义

下面对算法中的概念给出如下定义。

在复杂网络中,每个节点的重要程度不同,不同节点对其他节点的影响力也不同。越重要的节点在网络中的影响力越大,即越重要的节点影响其他节点的能力越强。节点的重要程度体现了节点在整个复杂网络中的地位,在某种程度上体现了节点在整个网络中的“分量”。节点的重要程度越大,节点对其他节点的吸引力越大;重要程度越小,吸引力越小。这正与万有引力定律中物体的质在引力中的作用相似,物理世界中的物体之间的引力与物体的质量成正比;复杂网络中节点之间的引力与节点的重要程度成正比。在无权网络中,节点的重要程度可以直接使用节点的度来衡量,在加权网络中节点的重要性可以用与节点直接相连的边的权重之和来衡量。因此,本文将加权网络节点的重要性定义为节点的质。

定义 1(节点的质) 加权网络中节点 i 的质 $M(i)$ 为与其直接相连的边的权重之和。

$$M(i) = \sum_{j=1}^{D(i)} w_{ij} \quad (1)$$

其中, $D(i)$ 为节点 i 的度数,即与节点 i 连接的边的个数。 w_{ij} 为与节点 i 连接的第 j 条边的权重。

定义 2(节点之间的引力因子) 复杂网络中两个不同节点之间的引力,表示网络中节点之间联系的紧密程度。

$$F_{ih} = \frac{m_i m_h w_{ih}}{d^2(i, h)} \quad (2)$$

其中, i 和 h 分别为网络中的两个节点, $m_i = M(i)$ 和 $m_h = M(h)$ 分别为节点 i 和节点 h 的质, w_{ih} 为节点 i 与节点 h 之间最短路径上各条边的权重之和。

定义 3(节点的万有引力因子) 复杂网络中节点与其他节点的引力之和定义为这个节点的万有引力因子,节点 i 的万有引力因子为:

$$F_i = \sum_{j=1}^{n-1} \frac{m_i m_j w_{ij}}{d^2(i, j)} \quad (3)$$

其中, n 为网络中节点的数量; m_i (m_j) 为节点 i (j) 的质; $d(i, j)$ 为节点 i 与节点 j 之间距离,即节点 i 与节点 j 之间最短路径的长度; w_{ij} 为节点 i 与节点 j 之间的权重,定义为节点 i 与节点 j 之间最短路径中各条边的权重之和。节点的万有引力反映节点与网络中其他所有节点的吸引力的强弱。

定义 4(社区引力因子) 某个节点和某一特定社区中各节点之间的引力之和:

$$F_c = \sum_{j=1}^{n_c} \frac{m_i m_j w_{ij}}{d^2(i, j)} \quad (4)$$

其中, n_c 为社区 c 中节点的个数, j 为社区 c 中的节点。社区引力因子用来衡量节点与某个社区之间连接的紧密程度。 m_i 和 m_j 分别为节点 i 和节点 j 的质, w_{ij} 为节点 i 与节点 j 之间最短路径中各条边的权重之和。

定义 5(节点的邻接社区) 如果社区 c_i 中某节点 i 的邻居节点 j 属于另一个社区 c_j ,则称社区 c_j 为节点 i 的邻接社区。

定义 6(归属感) 归属感表示某一节点属于某邻接社区的程度,本算法中将归属感定义为节点与某一邻接社区的社区引力因子与节点和所有邻接社区的社区引力因子之和的比值。

$$ms(ic) = \frac{F_{ic}}{\sum_{j=1}^T F_{ij}} \quad (5)$$

其中, F_{ic} 为节点 i 与其相邻的某个社区之间的引力因子, T 为与节点 i 邻接的社区个数, F_{ij} 为节点和所有与其相邻的社区之间的引力因子之和。

可以使用重叠社区的模块度 EQ(Extend Modularity)^[10]对重叠社区划分的质量进行评价。EQ 通过扩展 Newman 等人提出模块度函数,使之可以用于评判重叠社区识别结果的好坏,模块度的值越高,社区划分的效果越好。2010 年,中科院计算所的沈华伟等人提出了模块度指标用来评价重叠社区划分的效果,其定义如下:

$$EQ = \frac{1}{2m} \sum_{c \in C} \sum_{u,v} \alpha_{cu} \alpha_{cv} (A_{uv} - \frac{k_u k_v}{2m}) \quad (6)$$

其中, A 是网络的邻接矩阵,如果 u 和 v 之间存在一条连边,那么 $A_{uv}=1$,否则 $A_{uv}=0$; m 是网络的总边数; c 是对应划分的社区集合; k_u 和 k_v 分别是节点 u 和 v 的度; α_{cu} , α_{cv} 分别是节点 u 和节点 v 归属于社区 c 的程度。

将式(6)进行变换,使之适用于加权网络,将邻接矩阵 A 转换为加权网络的邻接矩阵 W 。如果节点 i 与节点 j 之间有

边相连,则 w_{ij} 的值为这两个节点连边的权重,否则为 0。 $ms(ic)$, $ms(jc)$ 分别是式(5)定义的节点 i 与节点 j 对于社区 c 的归属度; $M(i)$, $M(j)$ 分别为节点 i 和节点 j 的质; W_A 表示全网络中边的总权重之和, $W_A = \sum_{i,j} W_{i,j}$ 。

$$Q_{wo} = \frac{1}{W} \sum_{c \in C} \sum_{i,j} ms(ic)ms(jc)(W_{ij} - \frac{M(i)M(j)}{W_A}) \quad (7)$$

2.3 GWCR 算法的总体步骤

该算法主要由 3 部分组成:1)初始化建模并寻找中心节点;2)识别社区;3)识别重叠节点。

在初始化阶段,算法将所有节点标记为 F ,表明该节点未被划分到任何社区,也未进行过社区归属的计算。使用式(1)计算所有节点的质 $M(i)$,根据式(3)计算所有节点的万有引力因子,并按从大到小的顺序进行排序,万有引力因子大的节点的影响力较大,其可以作为社区的中心节点来进行社区识别。

在社区识别阶段,主要将归属度作为衡量标准来识别节点的社区归属。首先选取标记为 F 的节点中万有引力因子最大的节点作为初始社区 c_i ,并将节点划分标记改为 T 。其次计算社区的邻居节点与社区之间的社区引力因子,将节点划分标记改为 T 。计算节点的归属度,如果归属度大于归属度阈值 β ,则将节点归入该社区,并将划分标记改为社区标记 c_i 。遍历网络中所有节点,直到没有划分标记为 F 的节点。最后计算标记为 T 的节点与其相邻社区的归属度,将其归入归属度最大的社区,并将划分标记改为 c_j 。

经过上述步骤,识别的社区为非重叠社区,为了识别重叠社区,必须要进一步识别重叠节点。

为了方便识别重叠节点,需要记录社区的边界节点。本文将存在边直接连接到其他邻接社区的节点称为边界节点。算法中使用社区邻接矩阵 B 来记录节点所归属的社区。 B 中元素 B_{ij} ($i \neq j$) 表示节点 i 的邻居节点 j 所在的社区。矩阵 B 的对角线元素 B_{ii} 表示节点 i 所在的社区。由于矩阵 B 用于识别社区边界节点,可对其进行简化,即假设节点 j 为节点 i 的邻居节点,如果节点 j 与 i 在同一社区,则 $B_{ij} = 0$,否则 $B_{ij} = c_j$ (c_j 为节点 j 所属社区标识)。在矩阵 B 中,如果 B_{ij} ($i \neq j$) 不等于 0,则节点 i 和节点 j 为社区边界节点。对于边界节点,重新计算其归属度。由于社区划分的主要评价指标为 Q_{wo} ,根据文献[11]中的理论,在无权网络中,当节点在两个非重叠社团中的邻接点数比较均匀时,将其变为重叠节点后社团结构的模块度将不会受到明显影响。对于加权网络,除了节点的邻接节点数外,边的权重也是重要的衡量指标,本文将邻边的权重之和作为衡量标准,将在两个非重叠社区中邻边权重之和比较均匀的节点作为重叠节点。节点 i 的重叠度 $p_{ij}(v)$ 定义为节点 v 和在社区 c_i 中的节点的连边权重之和与节点 v 在邻接社区 c_j 中的节点连边权重之和的比。

$$p_{ij}(v) = \gamma \frac{w_j(v)}{w_i(v)} + (1-\gamma) \frac{w_i(v)}{w_j(v)} \quad (8)$$

其中, $w_i(v)$ 为节点 v 与同一社区节点连边的权重之和, $w_j(v)$ 为节点 v 与邻居社区 c_j 中节点连边的权重之和。如果 $w_i(v) > w_j(v)$,则 $\gamma = 1$,否则 $\gamma = 0$ 。将 $p_{ij}(v)$ 的值小于阈值 α 的节点视为重叠节点。识别重叠节点的算法 OLND 描述如下。

算法 1 识别重叠节点算法 OLND

INPUT: $G(V, E)$; 阈值 α ; 网络的权重向量 W
OUTPUT: 社区集合 $C = \{c_1, c_2, \dots, c_t\}$

1. 计算社区邻接矩阵 A
2. FOR (each node v in G)
3. IF (v 为社区 c_i 的边界节点,并与社区 c_j 中的节点相连)

4. calculate $p_{ij}(v)$
5. IF $p_{ij}(v) \leq \alpha$
6. $v \in c_j$
7. END IF
8. END IF
9. END FOR

2.4 GWCR 算法描述

下面对算法进行具体描述。

算法 2 GWCR 算法

INPUT: 网络 $G(V, E)$; 阈值 β ; 网络 G 中边的权重向量 W ; 网络 G 中节点个数 N ; 边的个数 M

OUTPUT: 网络划分的社区集合 $C = \{c_1, c_2, \dots, c_t\}$

1. 计算网络中节点的质 $M(i)$,并按从大到小的顺序进行排序
2. $S(i) = 'F'$ //将所有节点标记为 ' F '
3. $M(i) = \sum_j w_{ij}$ //计算所有节点的质
4. $F_i = \sum_{j=1}^n \frac{m_i m_j w_{ij}}{d^2(i, j)}$ //计算所有节点的万有引力因子
5. 按照 F_i 从大到小的顺序进行排序
6. FOR (标记为 ' F ' 的节点)
7. 选取 F_i 最大的节点 i 作为初始社区 c_i
8. $S(i) = 'T'$
9. FOR (节点 i 的所有邻居节点 j)
10. $F_{ic} = \sum_{j=1}^n \frac{m_i m_j w_{ij}}{d^2(i, j)}$ //计算社区引力因子
11. $ms(ic) = \frac{F_{ic}}{\sum_{j=1}^M F_{ij}}$ //计算归属度
12. $S(j) = 'T'$
13. IF $ms(ic) \geq \beta$
14. $S(j) = c_i$
15. END IF
16. END FOR
17. END FOR
18. FOR (标记为 ' T ' 的节点 k) //未被划分到任意一个社区的节点
19. FOR (节点 k 的任一邻接社区 c_j)
20. $ms(kc) = \frac{F_{kc}}{\sum_{j=1}^M F_{kj}}$
21. IF $ms(kc)$ 最大
22. $S(k) = c_j$
23. END IF
24. END FOR
25. END FOR
26. OLND() //识别重叠节点,输出重叠社区

3 实验结果与分析

本节从多个角度对算法的性能进行评测,首先利用人工生成网络分析算法的参数特性,其次在不同规模的 3 个真实网络上进行算法效果的对比实验。

3.1 评价指标验证

为了验证本文提出的加权网络重叠社区识别算法的评价指标 Q_{wo} 的正确性,需要与其他已验证的评价指标做比较。由于无权网络可以看作是所有边的权重都为 1 的加权网络,将本文所提出的评价指标 Q_{wo} 与已有的无权网络的重叠社区评价指标 Q_{on} 做对比。

实验采用 Newman 和 Girvan 给出的计算机生成的仿真网络。该网络有 128 个节点,分成 4 个社区,每个社区的节点数均为 32 个。对每个任意的节点对分别放置一条边,这 2 个节点位于 2 个不同的社区内的概率为 p_{out} ,而位于同一个社

区内的概率 $1 - p_{out}$ 。构建网络使得每个节点的度的期望值都为 16。本文设 p_{out} 以 0.05 为步长, 由 0.05 变到 0.50。实验结果如图 1 所示, 两个评价指标的最大值所对应的社区划分恰好完全相同, 证明 Q_{wo} 的评价指标是有效的。

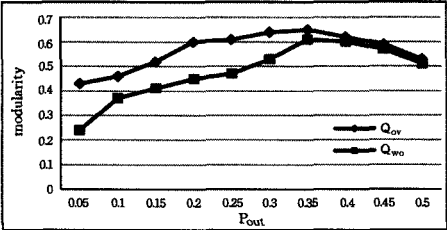


图 1 评价指标 Q_{wo} 的验证

3.2 参数测试

利用 LFR 生成的网络对算法参数 β 的特性进行分析。算法的结果利用反映社区划分准确度的 NMI (Normalized Mutual Information) 指标进行测试。NMI 通常用来衡量算法在已知社区结构的网络中所识别的社区的准确率, 定义如式 (9) 所示:

$$NMI(A, B) = \frac{-2 \sum_{i=1}^{C_A} \sum_{j=1}^{C_B} C_{ij} \log(\frac{C_{ij} C_i}{C_i C_j})}{\sum_{i=1}^{C_A} C_i \log(\frac{C_i}{C}) + \sum_{j=1}^{C_B} C_j \log(\frac{C_j}{C})} \quad (9)$$

其中, A, B 是网络 C 的两种不同的划分, $C_A (C_B)$ 是 $A (B)$ 中社区的个数, C_{ij} 表示同时在 A 中的 i 社区和 B 中的 j 社区的节点的数量, C_i 表示在 A 中第 i 个社区节点的个数, C 表示网络中节点的数量。互信息表示的是算法识别的社区与真实社区结构这两个集合之间的相似性程度, 取值范围为 $[0, 1]$, 取值越趋近于 1, 表明两个集合越相似。如果算法的结果与真实的网络社区结构完全一致, 那么 NMI 的值为 1。

LFR 基准网络假设网络中节点的度的分布和社区的大小的分布都符合幂律分布。该模型生成的网络中, 将每个节点在社区内部的边设为其节点度数的 $1 - \mu$ 倍, 与社区外的节点相连的边的个数是其度数的 μ 倍, 在这里 μ 为混合参数, 取值范围为 0 到 1。除此之外, LFR 模型还有其他表征网络结构的参数: 参数 N 表示网络中的节点个数, 参数 $\langle k \rangle$ 表示网络节点的平均度数, 参数 k_{min} 表示网络中节点的最小度数, 参数 k_{max} 表示网络中节点的最大度数, 参数 C_{min} 表示规模最小的社区中节点的个数, 参数 C_{max} 表示规模最大的社区中节点的个数; 为了构建重叠社区, 该模型使用参数 O_m 表示网络中的重叠节点所属社区的个数, 使用参数 O_n 表示网络中重叠节点的个数。

为了测试算法的效果, 选择使用 LFR 构建 3 类人工生成网络: 1) 标准网络 S (Standard Networks); 2) 大社团网络 LC (Large Community Networks); 3) 大规模网络 LN (Large Scale Networks)。3 类网络的参数设置如表 1 所列。

表 1 3 类 LFR 网络参数			
	S	LC	LN
N	1000	1000	4000
k	10	10	10
k_{max}	30	30	30
C_{min}	10	30	10
C_{max}	50	100	50
μ	0.1	0.1	0.1
O_n	100	100	400
O_m	2	2	2

通过上述实验可以看到, 在这 3 类不同特性的网络上, GWCR 算法的参数 β 取值在 0.6~0.65 间, 算法往往得到比较好的效果, 如图 2 所示。算法的另一个参数是重叠度的阈值, 该阈值的选取已经在文献[12]中进行了验证, 本文在此不再赘述, α 取值为 0.55。

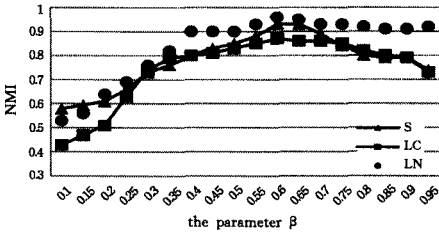


图 2 归属度阈值 β 参数的实验

3.3 实验数据集

本文选取 3 个经典的真实网络进行算法的验证。第一个数据集是来自美国一所空手道俱乐部成员间的关系网络^[13]。它是 20 世纪 70 年代初期 Zachary 观察空手道俱乐部, 基于俱乐部成员之间的关系建立的 karate 网络。该网络有 34 个节点, 代表俱乐部中 34 个成员, 网络的所有边的权重从 1 变化到 11, 反映各成员之间联系的紧密程度。第二个数据集是从 Arxiv 网络抓取的论文数据集^[14], 将该数据集抽象为无向加权网络, 网络中的节点代表数据集中论文的作者, 节点之间的连边表示作者的论文之间的引用关系, 连边的权重表示作者之间引用的次数。该网络具有 242 个节点和 684 条边。第三个数据集是 Facebook-like social network 数据集^[15]。该数据集描述加州大学欧文分校学生的在线社会网络。它共包含 1899 个节点, 节点表示学生, 边表示学生之间的信息接受和发送, 边上的权重表示接受和发送信息的数量。将 Facebook-like social network 数据集转化为无向加权图进行实验。

3.4 实验结果分析

实验在一台配置为 Intel i5 2.27GHz CPU 和 4GB DDR3 内存的笔记本上进行, 操作系统为 Win 7, 使用 Java 语言编程实验。

针对上述 3 个真实网络, 将本文所提出的算法与经典的重叠社区识别算法进行比较, 使用 Q_m 作为加权网络的社区模块度来衡量算法的效果, 与经典的重叠社区识别算法 COPRA (Community Overlapping Propagation Algorithm)^[16], LFM^[17] 以及 CFinder^[18] 进行比较。COPRA 算法是 Steve Gregory 于 2010 年提出的重叠社团识别算法, 其将标签传播算法应用到重叠社区识别中; LFM 算法是 Lancichinetti 等人于 2009 年提出的既能识别重叠节点又能给网络结构分层次的一种优化适应度函数 (Fitness Function) 的社区识别算法, 该方法可用于大规模的网络社区划分; CFinder 是一种基于 CPM 的可视化分析软件, 它能够在网络中寻找指定大小的全连集, 并通过全连集中共享的节点和连边构建更大的节点集合。通过对上述重叠社区识别算法进行扩展, 使之适用于加权网络。

首先在小规模网络 karate 上进行实验, 该网络是广泛应用于社团结构研究的标准网络, 网络结构已知。本算法将 karate 网络划分为 3 个社区, 分别是 $c1 = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 13, 14, 18, 19, 20, 22, 30, 31\}$, $c2 = \{8, 9, 10, 15, 16, 19, 21, 23, 24, 27, 28, 30, 33, 34\}$, $c3 = \{24, 25, 27, 28, 31\}$ 。实

验结果如图 3 所示,GWCR 算法的划分结果的 Q_{ov} 值比 COPRA, LFM 及 CFinder 都有明显的提高。

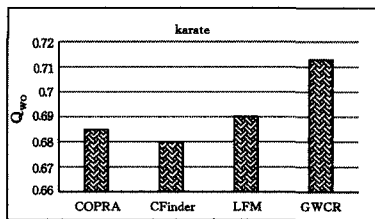


图 3 karate 网络上执行不同算法的 Q_{ov}

其次,在中型网络 Arxiv 数据集上进行实验,实验结果如图 4 所示。GWCR 算法所识别社区的重叠社区模块度明显高于其他 3 个算法。

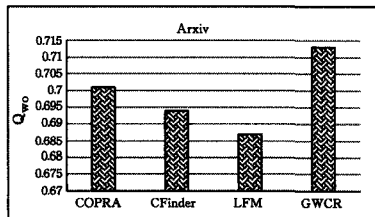


图 4 Arxiv 网络上执行不同算法的 Q_{ov}

最后,对于较大型的网络 Facebook,实验结果如图 5 所示,GWCR 算法的 Q_{ov} 值与 LFM 算法的基本相同,证明该算法也适用于较大规模的网络。

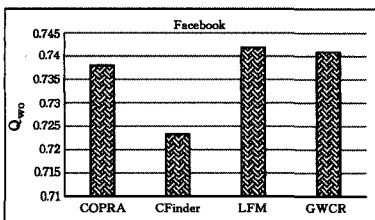


图 5 Facebook 网络上执行不同算法的 Q_{ov}

对于真实的加权社会网络,都进行了 50 次实验,实验结果基本一致,证明所提算法具有较强的稳定性;实验结果与网络的真实情况相同,证明所提算法的准确性较高。

结束语 本文提出一种在加权网络中识别重叠社区的方法,该方法受到物理学中的万有引力的启发,将社会网络中的引力分为 3 类进行计算,通过引力因子大小判断节点属于哪个社区。该算法较现有的经典加权重叠社区识别算法能更准确地地区分出重叠社区,并且也能计算出影响力大的中心节点,为进一步对社会网络进行监督控制提供了有利的依据。算法也可以通过将网络中边的权重进行归一化处理转换为无权网络,应用于无权网络的重叠社区识别。下一步可以对算法进行改进,降低其复杂度,使之更适应于大型加权网络的重叠社区识别。

参考文献

[1] Liu Y, Liu Q, Qin Z. Community detecting and feature analysis in real directed weighted social networks [J]. Journal of Networks, 2013, 8(6): 1432-1439

[2] Palla, Gergely, et al. Uncovering the overlapping community structure of complex networks in nature and society [J]. Nature, 2005, 435(7043): 814-818

[3] Chen Duan-bing, Shang Ming-sheng, Li Xia. Two-phase Strategy on Overlapping Communities Detection [J]. Computer Science, 2013, 40(1): 225-228 (in Chinese)

陈端兵, 尚明生, 李霞. 重叠社区发现的两段策略 [J]. 计算机科学, 2013, 40(1): 225-228

[4] Lancichinetti A, Radicchi F, Ramasco J J, et al. Finding statistically significant communities in networks [J]. PloS one, 2011, 6(4): 3360338

[5] Baumes J, Goldberg M, Magdon-Ismael M. Efficient identification of overlapping communities [C]// Intelligence and Security Informatics. Springer Berlin Heidelberg, 2005: 27-36

[6] Kim Y, Jeong H. Map equation for link communities [J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2011, 84(2): 1402-1409

[7] Lv T Y, Xie W, Zheng W, et al. Analysis of community evaluation criterion and discovery algorithm of weighted complex network [J]. Acta Physica Sinica, 2012, 61(21): 210511 (in Chinese)

吕天阳, 谢文艳, 郑伟民, 等. 加权复杂网络社团的评价指标及其发现算法分析 [J]. 物理学报, 2012, 61(21): 210511

[8] Lou X, Suykens J A. Finding communities in weighted networks through synchronization [J]. Chaos: An Interdisciplinary Journal of Nonlinear Science, 2011, 21(4): 744-783

[9] Huang Jia-xin, Guo Kun, Guo Hong. Label Propagation Algorithm for Community Detection Based on Vertex Significance and Label Influence [J]. Journal of Chinese Computer Systems, 2015, 36(6): 1171-1175 (in Chinese)

黄佳鑫, 郭昆, 郭红. 融入节点重要性和标签影响力的标签传播社区发现算法 [J]. 小型微型计算机系统, 2015, 36(6): 1171-1175

[10] Cheng S X, Cai K, Hu M. Detect overlapping and hierarchical community structure in networks [J]. Physica A: Statistical Mechanics and its Applications, 2009, 388(8): 1706-1712

[11] Newman M E J, Girvan M. Finding and evaluating community structure in networks [J]. Physical review E, 2004, 69(2): 26113

[12] Wu Zhi-hao, Lin You-fang, Wan Huai-yu, et al. Efficient overlapping community detection in huge real-world networks [J]. Physica A: Statistical Mechanics and its Applications, 2012, 391(7): 2475-2490

[13] Zachary W W. An information flow model for conflict and fission in small groups [J]. Journal of Anthropological Research, 1977, 33(4): 452-473

[14] Leskovec J, Kleinberg J, Faloutsos C. Graphs over time: densification laws, shrinking diameters and possible explanations [C]// Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, 2005

[15] Opsahl T, Panzarasa P. Clustering in weighted networks [J]. Social networks, 2009, 31(2): 155-163

[16] Gregory S. Finding overlapping communities in networks by label propagation [J]. New Journal of Physics, 2009, 12(10): 2011-2024

[17] Lancichinetti A, Fortunato S, Kertesz J. Detecting the overlapping and hierarchical community structure in complex networks [J]. New Journal of Physics, 2009, 11(3): 19-44

[18] Adamcsek B, Palla G, Farkas I J, et al. CFinder: locating cliques and overlapping modules in biological network [J]. Bioinformatics, 2006, 22(8): 1021-1023