

ESN 中基于贪婪派系扩张的重叠社区发现

卢志刚, 吴 露

(上海海事大学 经济管理学院, 上海 201306)

摘 要: 传统局部扩张方法在对企业社会化网络(ESN)中的重叠社区结构进行识别时, 存在计算冗余与社区挖掘不彻底的问题。为此, 提出一种基于贪婪派系扩张的重叠社区发现算法 GFE。在原始 ESN 中寻找极大派系, 根据派系间的关联程度计算其链接强度, 将原始网络图转换成最大派系图。在最大化适应度函数的条件下, 贪婪扩张最大派系图中的种子派系, 以进行社区发现。在此基础上, 比较社区差异度, 合并近似重复的社区, 从而优化重叠社区的层次结构。实验结果表明, GFE 算法能有效发现 ESN 中的重叠社区结构, 且运行效率高于 CPM、LFM 等算法。

关键词: 贪婪派系扩张; 极大派系; 企业社会化网络; 社区发现; 重叠社区

中文引用格式: 卢志刚, 吴露. ESN 中基于贪婪派系扩张的重叠社区发现[J]. 计算机工程, 2019, 45(7): 32-40.

英文引用格式: LU Zhigang, WU Lu. Overlapping community discovery based on greedy factional expansion in ESN[J]. Computer Engineering, 2019, 45(7): 32-40.

Overlapping Community Discovery Based on Greedy Factional Expansion in ESN

LU Zhigang, WU Lu

(School of Economics and Management, Shanghai Maritime University, Shanghai 201306, China)

[Abstract] There are problems of computational redundancy and incomplete community mining in traditional local expansion methods for identifying overlapping community structures in Enterprise Social Network (ESN). Therefore, an overlapping community discovery algorithm GFE based on greedy factional expansion is proposed. GFE algorithm searches for maximal factions in the original ESN, calculates their link strength according to the degree of association between factions, and converts the original network graph into the maximal faction graph. Under the condition of maximizing fitness function, the seed factions in the maximal faction graph are greedily expanded for community discovery. On this basis, the community differences are compared, and the similar duplicated communities are merged to optimize the hierarchical structure of overlapping community. Experimental results show that the GFE algorithm can effectively discover overlapping community structure in ESN, and the operation efficiency is higher than those of CPM, LFM and other algorithms.

[Key words] greedy factional expansion; maximal faction; Enterprise Social Network (ESN); community discovery; overlapping community

DOI: 10.19678/j.issn.1000-3428.0052337

0 概述

企业社会化网络(Enterprise Social Network, ESN)是企业个体间为适应市场需求, 通过交互合作而形成的关系体系。在 ESN 中, 企业节点基于关系特性产生聚集效应, 呈现出模块化的社区结构。社区发现有助于分析网络中企业团体的拓扑属性、模式以及功能特性, 挖掘隐藏的企业关系与合作规律并优化供应链需求, 对探究 ESN 的结构特征具有重要意义。

传统的社区发现算法按照网络节点内在拓扑结

构的连接紧密程度, 将节点划分成若干互不相连的社区, 其典型代表有层次聚类算法^[1]、GA 算法^[2]、基于信息论算法^[3]以及模块度优化算法^[4]。然而, 在现实网络中, 有些节点的隶属关系并不唯一, 其可能同时隶属于多个社区, 社区间存在明显的重叠与嵌套特征。因此, 挖掘重叠社区结构往往更有实际意义。近年来, 重叠社区发现引起了研究者的广泛关注, 一些代表性算法相继被提出, 这些算法主要分为 3 类:

第 1 类是基于局部信息的算法, 这类算法主要依靠节点或边的局部信息来扩展探测社区的重叠结

基金项目: 上海市自然科学基金(18ZR1416900)。

作者简介: 卢志刚(1973—), 男, 教授、博士, 主研方向为大数据分析、商务智能、供应链管理; 吴 露, 硕士研究生。

收稿日期: 2018-08-07 **修回日期:** 2018-09-19 **E-mail:** 2893636745@qq.com

构。其中代表性算法有 OSLOM^[5]、LFM^[6]、GCE^[7]以及 LLCM^[8]等。

第2类是基于标签传播的算法,这类算法根据每个节点及其邻居节点的标签与隶属度,更新划分每个节点的社区归属并进行重叠社区发现。其中代表性算法有 BMLPA^[9]、LPPB^[10]以及 FNCS-LPA^[11]等。

第3类是基于链接聚类的算法,这类算法以网络中的边为对象划分每条边的社区归属,再将划分得到的链接社区转化成相应的节点社区。其中代表性算法有 LINK^[12]、Link-Comm^[13]以及 DBLINK^[14]等。

其他相关算法还包括基于非负矩阵分解模型的 LANMF^[15]、MCMOEAL^[16]、NMF^[17]等。在现实的 ESN 中,企业个体间的竞争力与地位往往不同,核心企业与配套企业间存在明显差异^[18],从而导致派系聚簇现象产生。本文结合局部信息的特点,通过种子派系来研究 ESN 中社区局部扩张的过程,提出一种基于贪婪派系扩张的社区发现算法 GFE。利用种子派系信息不断扩张、合并周围新的企业节点,得到一个理想规模的派系社区结构。重复迭代上述过程,以得到覆盖 ESN 中多个派系的重叠社区。

1 基于派系的社区结构

给定一个企业社会化网络 $G = (V, E)$, 其中, $V = \{v_1, v_2, \dots, v_N\}$ 表示企业节点的集合, N 表示节点数目, $E = \{(v_i, v_j) | v_i, v_j \in V, i \neq j\}$ 表示企业节点间合作关系边的集合。企业社区发现的目的是在 ESN 中找到符合一定条件的企业节点集合,这是典型的 NP-hard 问题。

1.1 企业社区的基本定义

因合作、联盟等原因聚在一起的企业集群,具有高度的互动性,他们以整体利益最大化为目标形成社区结构。企业社区结构通常可以被描述为企业社会化网络节点集合的若干子集,每个子集内部节点之间的连接相对稠密,而不同子集节点之间的连接相对稀疏。基于文献[19]对有关社区的描述,本文作如下定义。

定义1(结构相对度) 给定一个企业社会化网络图 $G = (V, E)$, $A = [A_{ij}]_{N \times N}$ 为图 G 中的邻接矩阵。子图 $S \subset G$, 对于 $\forall v_i \in S$, 企业节点 v_i 的内部度与外部度可以用邻接矩阵元素分别表示为:

$$d_S^{\text{in}}(v_i) = \sum_{v_j \in S} A_{ij}$$

$$d_S^{\text{out}}(v_i) = \sum_{v_j \in S, v_j \neq v_i} A_{ij}$$

则企业节点 v_i 的结构相对度 $d_S(v_i)$ 表示为:

$$d_S(v_i) = d_S^{\text{in}}(v_i) + d_S^{\text{out}}(v_i)$$

定义2(企业社区) 给定一个企业社会化网络

图 $G = (V, E)$, 子图 $S \subset G$, $\forall v_i \in S$, 如果满足 $d_S^{\text{in}}(v_i) > d_S^{\text{out}}(v_i)$, 则称 S 为强连接企业社区。若 S 仅满足 $\sum_{v_i \in S} d_S^{\text{in}}(v_i) > \sum_{v_i \in S} d_S^{\text{out}}(v_i)$, 则称其为弱连接企业社区。

在强连接企业社区中,每个节点的内部链接都多于其外部链接。在弱连接企业社区中,所有节点的内部链接之和多于其外部链接之和。

1.2 派系社区

ESN 中存在一种企业节点间两两完全互连的特殊企业子团,该子团被称为派系。 k 派系指该子团中存在 k 个企业节点,极大派系则表示企业子团规模 k 已达到最大,无法通过添加其他企业节点来扩展该派系规模。

定义3(极大派系) 给定一个企业社会化网络图 $G = (V, E)$, $S \subset G$. F 是 S 中的局部结构,若 F 中所有企业节点 v_i 的个数为 k , 且 F 的内部链接总数满足 $\sum_{v_i \in F} d_F^{\text{in}}(v_i) = \frac{k(k-1)}{2}$, 则称 F 是 S 中的派系。如果有 $\forall v_u \in S$ 且 $v_u \notin F$, 满足 $\sum d_F^{\text{in}}(v_i \cup v_u) < \frac{k(k-1)}{2}$, 则称 F 是 S 中的极大派系。

如图1所示,三角形节点表示已加入社区的邻居企业节点,圆形节点表示待加入的候选企业节点。方形节点 v_1, v_2 和 v_3 两两互连构成规模为3的派系,但该派系并非极大派系,可通过添加节点 v_4 将其扩展到规模为4的极大派系。极大派系是一种相对稳固的局部结构,派系中的企业合作关系较为紧密,且其中的节点难以被替换。

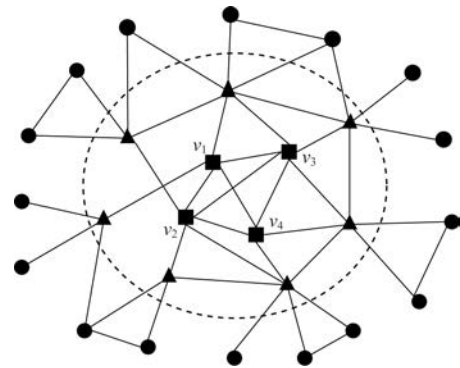


图1 派系社区示意图

派系社区是基于极大派系的一种社区结构,其包含极大派系以及周围一定区域内的邻居企业节点,该区域范围由适应度函数 f 决定。一个派系社区 S 中所包含的企业节点 v_{si} 必须满足 $f(v_{si}) > 0$ 。

1.3 重叠派系社区

在实际中,一个企业节点可能隶属于多个企业社区,派系社区间可能存在重叠、嵌套等现象。例如,某企业同时参与手机零件制造和空调零件供应,且在这2个领域都发挥着一定的作用,则该企业视

为加入了 2 个派系社区。在现实的 ESN 中,重叠派系社区与普通社区在描述上存在一定区别,稠密的重叠派系社区之间也可能存在很多外部链接。如图 2 所示,派系社区 S_1 与 S_2 相交,存在 v_1 与 v_2 2 个重叠节点。对于社区 S_1 内部节点 v_1 来说,其与 S_1 中其他节点间有 3 个链接,而与 S_2 存在 4 个链接,即外部链接多于内部链接。此外,重叠派系社区还体现了企业社区重叠性与层次性的双重特征。

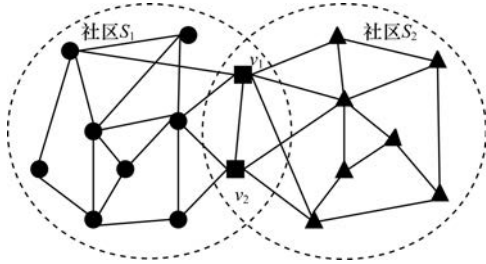


图 2 重叠派系社区示意图

定义 4(社区重叠度) 给定 2 个派系社区 S_1 和 S_2 , 社区重叠度定义为:

$$O_v(S_1, S_2) = \frac{|S_1 \cap S_2|}{\min(|S_1|, |S_2|)}$$

若 ESN 中 2 个派系社区 S_1 和 S_2 共享了 $Q_{S_1, S_2}^{O_v}$ 个企业节点, 则 S_1 和 S_2 称为重叠派系社区。此外, 在派系社区发现过程中可能会出现高度重叠、近似重复的社区对, 这就需要对社区的重叠度 O_v 进行限定。当社区对间的重叠度 O_v 超过一定范围时, 则表明这 2 个社区相似度较大, 可进行合并。

2 GFE 算法

为避免在企业社区发现过程中 2 个社区高度重叠现象的发生, 提高社区发现的效率, 本文提出一种基于贪婪派系扩张的重叠社区发现算法 GFE。

2.1 种子派系选择

GFE 算法需要选择合适的种子派系, 利用种子派系信息挖掘 ESN 中的社区结构。将图 $G = (V, E)$ 转换为最大派系图 $G^c = (V^c, E^c)$, 其中, $V^c = \{v_1^c, v_2^c, \dots, v_m^c\}$ 是派系集合, $E^c = \{(v_m^c, v_n^c) | v_m^c, v_n^c \in V^c, m \neq n\}$ 是派系间连接关系边的集合。在图的转换过程中, 需先确认每个派系中的企业节点, 然后利用派系之间的结构计算其链接强度。

2.1.1 派系节点确定

为找出极大派系, 首先需要确定派系中的每个企业节点。由于每个极大派系均是图 G 中企业节点所隶属的派系之一, 因此本文将派系中的节点确认过程转化成找出 G 中每个节点所隶属的所有极大派系的过程, 如算法 1 所示。

算法 1 确定 G^c 中的派系节点

输入 初始图 $G = (V, E)$

输出 派系节点集合 V^c

```

1.  $V^c \leftarrow \emptyset$ 
2. Calculate the degree  $k(v_i)$  of each node  $v_i \in V$ ;
3.  $k_{\max} \leftarrow \max_{v_i \in V} k(v_i)$ 
4. Sort the nodes in descending order of the degree;
5. for  $k = k_{\max} + 1$  to 1 do
  //节点度不小于  $(k - 1)$ 
6. for each node  $v_i \in V$  do
7. if  $k(v_i) < k - 1$ 
8. no more  $k$ -factions exist and goto Outer loop;
9. end if
  //该节点没有被分配给任何派系
10. if  $v_i$  has been assigned to one faction node
11. goto Inner loop;
12. end if
  //至少  $(k - 1)$  个相邻节点的度数不小于  $(k - 1)$ 
13.  $\text{Neigh}(v_i) \leftarrow \{v_j | (v_j \text{ is adjacent to } v_i \text{ and } k(v_j) \geq k - 1)\}$ ;
14. if  $|\text{Neigh}(v_i)| < k - 1$ 
15.  $v_i$  cannot constitute  $k$ -faction and goto Inner loop;
16. end if
  //转换为搜索由该节点邻居构成的  $(k - 1)$  派系问题
17. if the nodes in  $\text{Neigh}(v_i)$  can constitute  $q(k - 1)$ -faction ( $q \geq 1$ )
  //将所有发现的  $k$  派系添加到派系节点集合中
18.  $V^c \leftarrow V^c \cup \{v_i, ((k - 1)\text{-faction})_1\} \cup \dots \cup \{v_i, ((k - 1)\text{-faction})_q\}$ ;
19. end if
20. Inner loop;
21. end for
22. Outer loop;
23. end for
  
```

在算法 1 中, 派系中的节点根据其所在派系大小以降序排列。企业节点的“择优连接”偏好性表明企业更愿意与关系连接数目较多的企业节点建立联系, 即 2 个企业节点建立链接关系的边越多, 其构成极大派系的机率越大。因此, 本文首先计算 G 中每个节点的度数, 然后按度数的降序将节点进行排序。

假设 G 中的节点数为 N , 最大节点度为 k_{\max} , 则 G 中的派系规模不大于 $k_{\max} + 1$ 。随着派系规模 k 的值从 $k_{\max} + 1$ 下降到 1, 搜索每个节点所隶属的 k 派系。因为派系中的节点两两完全互连, 所以 k 派系中所有节点的度数必须大于 $k - 1$ 。此外, 若一个节点已经被分配给较大的派系, 则该节点寻找到相对较小的子团后停止搜索过程。因此, 搜索节点所隶属的 k 派系需满足以下 3 个条件:

- 1) 该节点度数不小于 $k - 1$ 。
- 2) 该节点没有被分配给任何派系。
- 3) 该节点至少有 $k - 1$ 个相邻节点的度数不小于 $k - 1$ 。

若满足上述所有条件, 查找该节点的所有 k 派系的问题, 可转换为确认其与邻居节点是否具有 $k - 1$ 个连接关系的问题, 最后将发现的所有 k 派系添加到派系集合中。

图3(a)所示为一个具有9个节点和14条连接边的ESN原始图,图3(b)给出了相应的派系节点,其中,包括一个规模为4的派系和3个规模为3的派系。由图3可以看出,节点 v_7 、 v_8 和 v_9 属于派系 v_m^c ,而节点 v_4 和 v_5 同时属于派系 v_m^c 和 v_n^c 。原始图的每个节点至少被分配给对应的最大派系图中的一个派系。

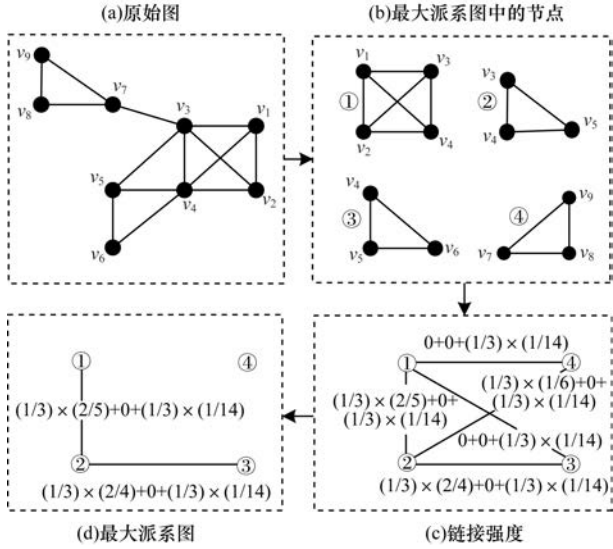


图3 从原始图到最大派系图的构造过程示例

2.1.2 派系间的链接强度

2个派系之间的链接强度大小取决于重叠节点、重叠边以及联通边的比例。给定2个派系,重叠节点为这2个派系的共同节点,其连接边被称为重叠边,联通边指2个派系中不同节点间的连接边。如图4所示,图4(a)中的 v_3 为重叠节点,图4(b)中节点 v_2 与 v_3 间的边为重叠边,图4(c)中节点 v_3 与 v_4 间的边为联通边。

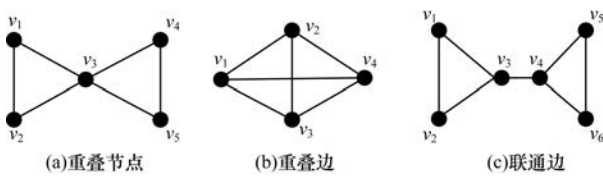


图4 重叠节点、重叠边、联通边示例

定义5(重叠节点比率) 给定2个派系 v_m^c 和 v_n^c ($m \neq n$),其之间的重叠节点比率 $l_{on}(v_m^c, v_n^c)$ 记为:

$$l_{on}(v_m^c, v_n^c) = \frac{N(v_m^c \cap v_n^c)}{N(v_m^c) + N(v_n^c) - N(v_m^c \cap v_n^c)}$$

其中, $N(v_m^c)$ 、 $N(v_n^c)$ 分别表示派系 v_m^c 、 v_n^c 的节点数目, $N(v_m^c \cap v_n^c)$ 表示派系 v_m^c 与 v_n^c 间的重叠节点数目。

定义6(重叠边、联通边比率) 设 $A = [A_{ij}]_{N \times N}$ 为原始图 G 中的相邻矩阵,若 $(v_i, v_j) \in E$,则 $A_{ij} = 1$,否则 $A_{ij} = 0$ 。重叠边比率 $l_{oe}(v_m^c, v_n^c)$ 、联通边比率

$l_{je}(v_m^c, v_n^c)$ 分别表示为:

$$l_{oe}(v_m^c, v_n^c) = \frac{\sum_{v_i, v_j \in (v_m^c \cap v_n^c)} A_{ij}}{\sum_{v_i, v_j \in V} A_{ij}}$$

$$l_{je}(v_m^c, v_n^c) = \frac{\sum_{v_i \in (v_m^c - v_n^c), v_j \in (v_n^c - v_m^c)} A_{ij}}{\sum_{v_i, v_j \in V} A_{ij}}$$

定义7(派系间链接强度) 基于 l_{on} 、 l_{oe} 、 l_{je} ,派系 v_m^c 和 v_n^c 间的链接强度为:

$$L(v_m^c, v_n^c) = \beta l_{on}(v_m^c, v_n^c) + \gamma l_{oe}(v_m^c, v_n^c) + \omega l_{je}(v_m^c, v_n^c)$$

其中, $\beta, \gamma, \omega \in [0, 1]$ 且 $\beta + \gamma + \omega = 1$, β, γ, ω 分别控制重叠节点、重叠边和联通边的权重。 $L(v_m^c, v_n^c) \in [0, 1]$,当2个派系 v_m^c 和 v_n^c 完全独立时, $L(v_m^c, v_n^c) = 0$ 。

如图3(c)所示,派系 v_1^c 和 v_2^c 间有2个重叠节点和1条重叠边,没有联通边,这2个派系的总节点数为5,总边数为14,则 $l_{on} = 2/5$, $l_{oe} = 1/14$, $l_{je} = 0$ 。故派系 v_1^c 和 v_2^c 间的链接强度为:

$$L(v_1^c, v_2^c) = \frac{1}{3} \left(\frac{2}{5} + \frac{1}{14} + 0 \right)$$

派系间的链接强度有强有弱,有些甚至为0。在派系扩张的过程中,链接相对紧密的派系间极有可能发展成近似重复的企业社区,这不仅会造成计算的浪费,还对社区层次结构的划分产生不利影响。为避免上述问题,可以设置一个链接强度阈值 Δ ,当派系间链接强度超过 Δ 时,这些派系被认为是“可疑种子”。采用一个简单优化方法合并这类“可疑种子”,然后进行扩张并发现企业社区,从而避免近似重复种子的膨胀现象,且不会影响社区发现的结果。

2.2 贪婪扩张

在选取了适当的种子派系后,通过贪婪算法不断扩张种子派系,选择、合并其邻居节点,最终得到一个候选企业社区 C' 。在该过程中,需要定义一个适应度函数^[6]来指导种子派系的发展,使其扩张成理想的社区。

定义8(结构适应度) 给定图 $G = (V, E)$,子图 $S \subset G$, d_{in}^S 、 d_{out}^S 分别表示 S 的内部度、外部度,其中, d_{in}^S 是2个 S 内部节点的链接数。则社区结构适应度函数定义为:

$$f_S = \frac{d_{in}^S}{(d_{out}^S + d_{in}^S)^\alpha}$$

其中, α 是控制社区规模大小的参数,其在一定程度上避免了扩张过程落入局部稠密的现象。

定义9(节点适应度) 给定图 $G = (V, E)$,子图 $S \subset G$,给定候选企业节点 $v_\mu \in V$,则节点 v_μ 对于 S 的结构适应度(即节点适应度)为:

$$f_{(S, v_\mu)}^{v_\mu} = f_{S \cup \{v_\mu\}} - f_{S - \{v_\mu\}}$$

其中, $f_{S \cup \{v_\mu\}}$ 与 $f_{S - \{v_\mu\}}$ 分别表示 S 中包含节点 v_μ 和不

包含节点 v_u 的结构适应度。

适应度函数返回的值反映了 S 内部及 S 与外部之间的连接紧密程度,增加不同的企业节点可能会引起适应度函数值的增大或缩小,函数值越高,则表明增加节点后社区结构越好。在 2.1 节提供的初始种子派系的基础上,利用适应度函数不断选取合适的成员节点,贪婪扩张种子派系,直到添加任何企业节点都不能增大适应度函数值为止。

2.3 社区差异度

通过适应度函数贪婪扩张可以有效挖掘 ESN 中社区的重叠结构,且通过种子派系的贪婪扩张过程,也在一定程度上展现了社区的良好层次结构。但在此过程中,可能产生高度重叠的社区,同一种社区结构被高度相似的社区重复发现了多次,这不仅会造成计算成本的浪费,而且不符合 ESN 的社区结构特征。为避免上述问题,本文在不影响社区发现结果的前提下,采用基于社区重叠度的阈值合并方法进行过程优化。

定义 10(社区差异度) 给定 2 个企业社区 S_1 和 S_2 ,通过社区重叠度的定义, S_1 和 S_2 间的差异度 δ 为:

$$\delta_E(S_1, S_2) = 1 - \frac{|S_1 \cap S_2|}{\min(|S_1|, |S_2|)}$$

社区差异度表示较小社区未嵌入较大社区中的节点所占的比率,比率越低,这 2 个社区的相似性越高。本文引入一个社区差异度阈值 ε ,当 $\delta_E(S_1, S_2) \leq \varepsilon$ 时,表明社区对的相似性过高,可将这 2 个社区进行合并。

给定一组已被接受的企业社区 W 和一个候选企业社区 C' ,将 C' 的近似重复社区定义为 W 中与 C' 的差异度不超过 ε 的所有社区。通常设置 ε 的默认值为 0.25,但是如果在输出过程中发现太多近似重复的社区,应适当增大 ε 的值。

2.4 算法实现

GFE 算法的整体步骤如下:

步骤 1 根据种子派系选择策略,将 ESN 原始图 $G = (V, E)$ 转换为极大派系图 $G^c = (V^c, E^c)$,选择未扩张的种子派系 F_0 作为初始社区 C_i 。

步骤 2 计算种子派系所有邻接节点对其的节点适应度,利用社区适应度函数贪婪扩张,直到没有节点可以增加适应度函数值,得到一个候选企业社区 C' 。

步骤 3 计算候选企业社区 C' 与已经被接受的社区 C 的差异度,若差异度不超过阈值 ε ,则 C' 与 C 近似重复,将 C' 与 C 进行合并;否则,接受 C' 。

步骤 4 返回步骤 1,进行新的社区扩张。若没有未扩张的种子派系,则算法终止,最后输出社区。

GFE 算法描述如下:

算法 2 GFE 算法

输入 初始图 $G = (V, E)$,种子派系 F_0

输出 生成的社区 C_i

```

1. for all  $v_u \in F_0$  neighbors
2. calculate  $v_u$ 's node fitness  $f_{(s, v_u)}^{v_u}$ ;
3. if  $f_{(s, v_u)}^{v_u} > 0$  then
4. add  $v_u$  to  $C_i$ ;
5. end if
6. end for
7. create a candidate community  $C'$ ;
8. for candidate community  $C'$ 
  Calculate the  $\delta_E$  for  $C'$  with already accepted community  $C$ 
9. if  $C'$  within  $\varepsilon$  of any already accepted community  $C$  then
10.  $C'$  and  $C$  are near-duplicates;
11. merge the candidate community  $C'$ 
12. else
13. Accept  $C'$ ;
14. end if
15. end for

```

3 实验结果与分析

为验证 GFE 算法的性能,将其与 CPM 算法^[20]、LFM 算法、基于层次聚类的 HC-PIN 算法^[21]以及基于标签传播的 COPRA 算法和 SLPA 算法^[22]进行比较。实验环境为 Lenovo PC Intel(R) Core(TM) i5-2520M CPU @ 2.50 GHz,内存为 8 GB,Windows 操作系统,编程语言为 Python。实验将在模拟 ESN 的 LFR 基准网络^[23]以及公开的社会网络数据集上进行算法比较。

3.1 评价指标

不同算法在同一网络上可能会划分出不同的社区结构,需要对划分出的社区质量进行评估。本文采用目前重叠社区研究中比较惯用的 4 种评价指标:标准化互信息(Normalized Mutual Information, NMI),相对标准化互信息(relative Normalized Mutual Information, rNMI), F_1 值,扩展模块度 EQ 。

3.1.1 NMI 和 rNMI

NMI 利用信息熵来衡量算法划分的社区结构与预先已知社区结构之间的差异,其是基于混合矩阵 N 进行计算的数字指标,定义为:

$$NMI = \frac{-2 \sum_{i,j} N_{ij} \lg \left[\frac{N_{ij} n}{N_{i.} N_{.j}} \right]}{\sum_i N_{i.} \lg \left[\frac{N_{i.}}{n} \right] + \sum_j N_{.j} \lg \left[\frac{N_{.j}}{n} \right]}$$

其中, n 表示网络节点的数量, N_{ij} 表示社区 i, j 中公共的节点数, $N_{i.}$ ($N_{.j}$) 表示混合矩阵 N 中第 i 行(第 j 列)之和。若算法发现的社区结构与预先已知的社区结构完全一致,则 NMI 值为 1;如果两者完全独立,则 NMI 值

为 0。NMI 值越接近 1, 则表明算法发现的社区质量越好。

当网络规模有限时, NMI 易受到系统误差的影响。rNMI 能够克服网络规模的影响, 其通过将 NMI 与随机分区的预期 NMI 进行比较来考虑统计的显著性^[24]。rNMI 定义为:

$$rNMI(A, B) = NMI(A, B) - \langle NMI(A, C) \rangle$$

其中, C 是与检测的分区大小相同的空模型。

3.1.2 F_1 值

F_1 值通过召回率 (*recall*) 和精确度 (*precision*) 的调和平均值, 度量算法划分的社区与预先已知社区结构间的匹配程度。 F_1 值定义为:

$$F_1 = \frac{2recall \times precision}{recall + precision}$$

其中, *recall*, *precision* 在网络二分后进行定义。

3.1.3 扩展模块度 EQ

扩展模块度 EQ 是建立在模块度 Q 基础上的评价指标。模块度指标 Q 用于衡量网络社区划分结果的优劣, 其思想是社区内的节点间连接边越多, 与社区外部连接边越少, Q 值越大, 社区结构划分效果越好。但在重叠社区研究中发现, 重叠节点可能与外部连接较多, 这会降低 Q 的值。 EQ 在此基础上进行改进, 其在处理重叠节点时, 会考虑这类节点的“重叠度”, 将“重叠度”对 Q 的影响值除以该节点所隶属社区的个数, 以减少这类节点对模块度值的影响, 使其可用于对重叠社区进行度量。 EQ 定义为:

$$EQ = \frac{1}{2m} \sum_c \sum_{i,j \in c} \frac{1}{Q_i Q_j} \left[A_{ij} - \frac{k_i k_j}{2m} \right]$$

其中, Q_i 表示节点 i 所属的社区数, k_i 为与节点 i 建立链接关系的边数, A_{ij} 是整个网络对应的邻接矩阵的元素, 若节点 i 与节点 j 间存在连接, 则 $A_{ij} = 1$, 否则 $A_{ij} = 0$, m 表示网络中节点间的连接边总数。

3.2 结果分析

3.2.1 LFR 基准网络

本文利用 LFR 基准网络模拟生成 ESN。LFR 基准网络具有如下 2 个优点:

1) 可以模拟 ESN 中的节点度和社区大小无标度特性。

2) 具有预先已知企业社区的结构, 可将其与算法发现的社区进行比较, 以评估算法的性能。

构建 LFR 基准网络图, 须定义一些参数。其中, N 表示企业节点的个数, k 表示企业节点的平均度数, k_{\max} 表示企业节点的最大度, O_n 表示网络图中重叠企业节点的数量, O_m 表示每个重叠节点所隶属的企业社区个数, τ_1 、 τ_2 分别表示节点度和社区大小所遵循的分布参数, C_{\min} 、 C_{\max} 分别表示最小社区规模和最大社区规模所包含的节点数量, 混合参数 μ 控

制企业社区内的节点与外部连接的比率, 随着 μ 值的增大, 企业社区的结构减弱, 社区发现的难度越大。本文根据实验需要设置不同的参数, 以生成不同类型的企业社会化网络。下文实验结果图 5 ~ 图 8 中对应的参数配置如表 1 所示。

表 1 LFR 基准网络参数设置

参数名称	参数配置		
	图 5	图 6	图 7、图 8
N	$10^3 \sim 10^4$	5 000	2 000
k	10	20 ~ 200	18 ~ 90
k_{\max}	30	200	120
μ	0.4	0.4	0.2
C_{\min}	10	k	60
C_{\max}	50	500	100
τ_1	-2	-2	-2
τ_2	-1	-1	-2
O_n	10	0	500
O_m	2	1	1 ~ 5

各算法在 LFR 基准网络上运行效率与节点数、节点平均度间的关系分别如图 5、图 6 所示。从图 5 可以看出, GFE 算法运行效率不仅与结构适应度参数 α 有关, 还与网络中的企业节点数成正比。CPM 算法由于要对网络中的全局极大子团进行定位, 因此其时间复杂度相对较高。LFM 算法采用局部贪婪优化策略, 与 GFE 算法相似, 但是 LFM 算法随机选择种子节点, 所以其时间复杂度也较高。SLPA 算法改进了原始标签传播算法 COPRA, 效率相对较高, 但仍低于 GFE 算法 ($\alpha = 0.9$)。从图 5 还可以看出, 随着节点数量的增加, 与其他重叠社区发现算法相比, GFE 算法运行效率的优势呈增长趋势。在图 6 所示的一系列实验中, 企业节点的数量保持固定值 5 000。从图 6 可以看出, 与本实验中运行效率较快的 SLPA 和 COPRA 算法相比, 随着平均度的增加, GFE 算法的运行时间增长很快, 即使在规模较小的图上, 节点平均度对 GFE 算法运行时间的影响也很大。但综合观察可以看出, GFE 算法在重叠社区发现的运行效率方面整体表现良好。

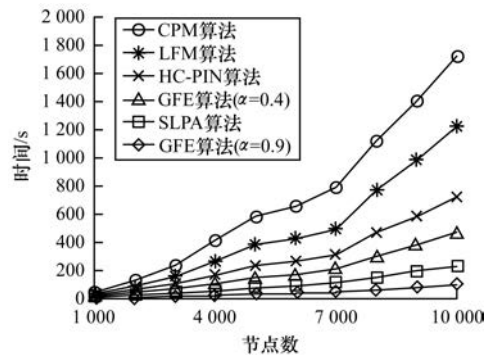


图 5 算法运行时间与节点数的关系

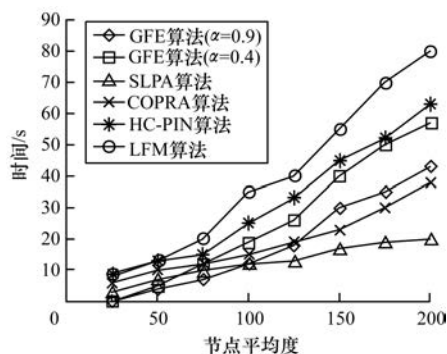


图 6 算法运行时间与节点平均度的关系

在社区重叠程度较高的情况下进一步比较各算法所发现社区的质量,分别用指标 NMI 和 rNMI 评价算法的性能。本次实验分别生成 5 个模拟网络。在第 1 个网络图中,每个节点隶属于唯一社区,设置节点平均度为 $k=18$ 。为了让企业节点更有可能同时隶属多个社区,本文设置更高的节点度。在第 2 个网络图中,每个重叠节点隶属于 2 个社区,节点平均度设置为 $k=36$ 。以此类推,在第 5 个网络图中,每个重叠节点隶属于 5 个社区,设置节点平均度为 $k=90$ 。其他参数设置情况如表 1 所示,实验结果如图 7、图 8 所示。

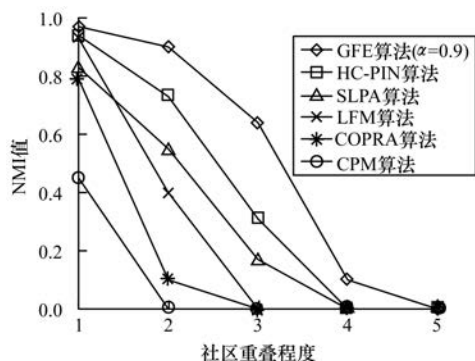


图 7 算法 NMI 值与社区重叠程度的关系

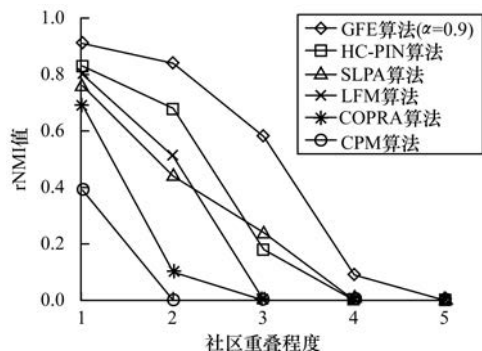


图 8 算法 rNMI 值与社区重叠程度的关系

从图 7、图 8 可以看出,即使社区的重叠程度保持在中等水平,CPM 算法和 COPRA 算法也不能有效地挖掘重叠企业社区结构。随着社区重叠程度的增加,SLPA 算法和 HC-PIN 算法表现较好,而 GFE 算法表

现最佳。虽然 LFM 算法和 GFE 算法使用相同的适应度函数和相似的贪婪扩张方式,但它们的测试结果差别很大。随着社区重叠程度的加深,LFM 算法的性能下降,原因是随机选择种子可能会导致其过早地放弃尝试扩张未被识别出的社区的图区域。

基于 GFE 算法 ($\alpha=0.6$) 发现的 ESN 重叠社区的可视化结果如图 9 所示。由图 9 可以看出,GFE 算法在 ESN 中发现了 3 个很明显的不同社区结构,图中用圆圈进行标注。尽管存在一些其他节点的干扰,但并未影响社区的结构。圆圈交叉部分内的企业节点就是社区间的重叠部分,可以看出,有些节点隶属于 2 个社区,而有些节点同时隶属于 3 个社区,社区间存在交叉重叠现象。虽然本次实验是利用 LFR 基准网络模拟 ESN,但从重叠社区的挖掘效果来看,这种做法具有一定的实际意义。



图 9 ESN 中重叠社区的可视化结果

综上,在运行效率与挖掘效果方面,相比 SLPA、LFM 等算法,GFE 算法具有一定优势。

3.2.2 真实社会网络

本节利用 5 个公开的标准社会网络数据集对算法进行测试。社会网络数据集包括 3 个已知社区结构的常用真实网络:空手道俱乐部网络 (Zachary's Karate Club, 简称 Karate)^[25],海豚网络 (Dolphin's Associations, 简称 Dolphin)^[26],NCAA 大学橄榄球联盟比赛网络 (College Football League, 简称 Football)^[25],以及 2 个大规模网络:科学合作者网络 (Netscience)^[27],单词关联网络 (Word Association)^[28]。这些社会网络的详细描述如表 2 所示,使用 F_1 值和扩展模块度 EQ 作为评价标准,以衡量各算法所划分社区的质量以及合理性。

表 2 真实社会网络数据集信息

网络	节点数	边数	已知社区数
Karate	34	78	2
Dolphin	62	160	2
Football	115	613	12
Netscience	1 589	2 742	—
Word Association	10 617	63 785	—

图 10 所示为采用 GFE 算法发现的 Karate 的重叠社区。2 个社区分别以节点 1、2、4 和节点 24、33、

34 为种子派系进行贪婪扩张, 中间的节点 3、9、10、14、31 是 2 个社区的重叠部分。由图 10 可以看出, GFE 算法在真实网络中同样能有效地发现较为理想的社区。

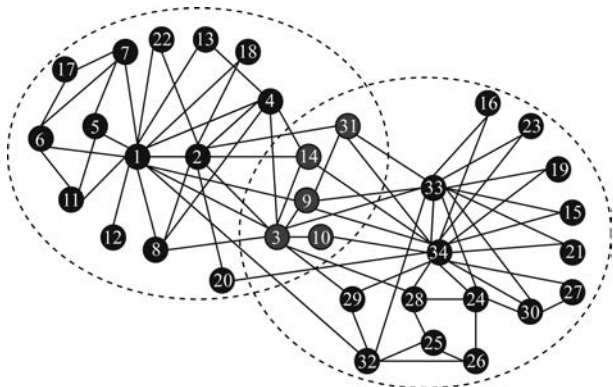


图 10 Karate 数据集重叠社区发现结果

表 3 所示为 6 种算法的性能比较结果。由表 3 可以看出, 本文 GFE 算法在 3 个数据集上都能找到较理想的社区, 且精度较高, 同时 F_1 值最优。从社区发现的角度看, GFE 算法将 3 个数据集进行准确地划分, 尤其是在 Karate 数据集中, GFE 算法始终保持正确。因此, GFE 算法总体上呈现出较高的平均准确度。

表 3 6 种算法性能比较结果

算法	数据集	召回率	精确度	F_1 值	社区数
GFE	Karate	0.843	1.000	0.915	2
	Dolphin	0.461	0.976	0.620	2
	Football	0.902	0.878	0.890	12
COPRA	Karate	0.703	0.923	0.798	2
	Dolphin	0.379	0.903	0.534	2
	Football	0.726	0.898	0.803	11
LFM	Karate	0.726	0.912	0.808	2
	Dolphin	0.331	0.981	0.494	2
	Football	0.911	0.830	0.868	13
CPM	Karate	0.581	0.941	0.719	2
	Dolphin	0.402	0.943	0.564	2
	Football	0.823	0.893	0.856	11
SLPA	Karate	0.735	0.952	0.830	2
	Dolphin	0.373	0.947	0.535	2
	Football	0.873	0.852	0.863	12
HC-PIN	Karate	0.855	0.961	0.904	2
	Dolphin	0.443	0.958	0.606	2
	Football	0.861	0.887	0.874	12

在 Word Association 网络中, 以 bright 为例划分出其所属的社区, 如图 11 所示。bright 单词涉及 emotion、color 和 light 3 个社区, 每个社区代表 bright 的一种含义。除 bright 外, GFE 算法还成功检测到 color 和 light 之间的其他重叠节点, 即 pale 和 gray。

实验结果表明, GFE 算法在 Word Association 中能够合理并有效地发现社区。

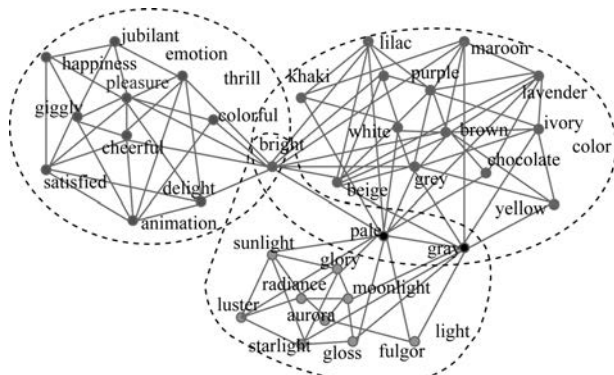


图 11 Word Association 网络分区示例

6 种算法在 5 个真实网络上的 EQ 值结果对比如表 4 所示。由表 4 可以看出, 在 Karate 和 Dolphin 中, GFE 算法的 EQ 值相对偏低, 但准确率较高, 而在 Football、Netscience 和 Word Association 中, GFE 算法的 EQ 值都偏高。原因是 GFE 算法采用种子派系作为初始社区, 具有良好的邻域鲁棒性, 且在一定程度上避免了由初始节点落入稀疏边界所带来的计算冗余与误差。综上, GFE 算法在真实数据集的社区发现中表现良好, 在规模较大的网络中优势更为明显。

表 4 6 种算法在真实网络上的 EQ 值结果对比

数据集	EQ 值					
	GFE	COPRA	LFM	CPM	SLPA	HC-PIN
Karate	0.323	0.364	0.314	0.342	0.392	0.407
Dolphin	0.392	0.385	0.386	0.490	0.427	0.397
Football	0.594	0.483	0.472	0.517	0.498	0.554
Netscience	0.632	0.584	0.573	0.591	0.641	0.662
Word Association	0.891	0.842	0.856	0.857	0.865	0.867

4 结束语

本文提出一种基于贪婪派系扩张的重叠社区发现算法 GFE。采用种子派系选择策略, 利用社区适应度函数贪婪扩张种子派系以得到一个候选社区, 将候选社区与已接受社区进行差异度比较, 判断是否接受该候选社区。在种子派系的选择过程中, GFE 算法将链接强度过高的派系进行合并, 能够在一定程度上避免网络社区的过度重叠情况, 降低计算复杂度。实验结果表明, 在运行效率与社区划分质量方面, GFE 算法相对 LFM、CPM 等算法具有一定的优势。利用 GFE 算法在 ESN 中发现具有重叠性的社区结构, 能够为企业分析相互间的合作关系、优化供应链需求提供一种新的方法和途径。但本文主要以静态网络为研究对象, 忽略了网络以及社区结构的动态变化, 下一步将分析动态网络中的社区演化模式。

参考文献

- [1] BLONDEL V D, GUILLAUME J L, LAMBIOTTE R, et al. Fast unfolding of communities in large networks[J]. *Journal of Statistical Mechanics*, 2008(10):155-168.
- [2] GUIMERÀ R, NUNES AMARAL L A. Functional cartography of complex metabolic networks[J]. *Nature*, 2005, 433(7028):895-900.
- [3] 邓小龙,王柏,吴斌,等. 基于信息熵的复杂网络社团划分建模和验证[J]. *计算机研究与发展*, 2012, 49(4):725-734.
- [4] SHANG Ronghua, BAI Jing, JIAO Licheng, et al. Community detection based on modularity and an improved genetic algorithm[J]. *Physica A: Statistical Mechanics and Its Applications*, 2013, 392(5):1215-1231.
- [5] LANCICHINETTI A, RADICCHI F, RAMASCO J J, et al. Finding statistically significant communities in networks[J]. *PLoS One*, 2011, 6(4):e18961.
- [6] LANCICHINETTI A, FORTUNATO S, KERTÉSZ J. Detecting the overlapping and hierarchical community structure of complex networks[J]. *New Journal of Physics*, 2009, 11(3):19-44.
- [7] LEE C, REID F, MCDAID A, et al. Detecting highly overlapping community structure by greedy clique expansion[EB/OL]. [2018-07-25]. <https://arxiv.org/pdf/1002.1827.pdf>.
- [8] 潘磊,金杰,王崇骏,等. 社会网络中基于局部信息的边社区挖掘[J]. *电子学报*, 2012, 40(11):2255-2263.
- [9] WU Zhihao, LIN Youfang, GREGORY S, et al. Balanced multi-label propagation for overlapping community detection in social networks[J]. *Journal of Computer Sciences and Technology*, 2012, 27(3):468-479.
- [10] 刘世超,朱福喜,甘琳. 基于标签传播概率的重叠社区发现算法[J]. *计算机学报*, 2016, 39(4):717-729.
- [11] 顾军华,霍士杰,王守彬,等. 基于节点中心性和社区相似性的快速标签传播算法[J]. *计算机应用*, 2018, 38(5):1320-1326.
- [12] AHN Y Y, BAGROW J P, LEHMAN S. Link communities reveal multiscale complexity in networks[J]. *Nature*, 2010, 466(7307):761-764.
- [13] KIM Y, JEONG H. Map equation for link communities[J]. *Physical Review E*, 2011, 84(2):026110.
- [14] 朱牧,孟凡荣,周勇. 基于链接密度聚类的重叠社区发现算法[J]. *计算机研究与发展*, 2013, 50(12):2520-2530.
- [15] 贺超波,汤庸,刘海,等. 一种集成链接和属性信息的社区挖掘方法[J]. *计算机学报*, 2017, 40(3):601-616.
- [16] WEN Xuyun, CHEN Weineng, LIN Ying, et al. A maximal clique based multiobjective evolutionary algorithm for overlapping community detection[J]. *IEEE Transactions on Evolutionary Computation*, 2017, 21(3):363-377.
- [17] KANNAN R, ISHTEVA M, PARK H. Bounded matrix factorization for recommender system[J]. *Knowledge and Information Systems*, 2014, 39(3):491-511.
- [18] 吴松强,孙波,王路. 集群中核心企业网络权力对配套企业合作行为的影响——关系资本的调节效应[J]. *科技进步与对策*, 2017, 34(13):81-88.
- [19] RADICCHI F, CASTELLANO C, CECCONI F, et al. Defining and identifying communities in networks[J]. *Proceedings of the National Academy of Sciences of the United States of America*, 2004, 101(9):2658-2663.
- [20] PALLA G, DERENYI I, FARKAS I, et al. Uncovering the overlapping community structure of complex networks in nature and society[J]. *Nature*, 2005, 435(7043):814-818.
- [21] WANG Jianxin, LI Min, CHEN Jianer, et al. A fast hierarchical clustering algorithm for functional modules discovery in protein interaction networks[J]. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2011, 8(3):607-620.
- [22] XIE Jierui, SZYMANSKI B K, LIU Xiaoming. SLPA: uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process[C]// *Proceedings of IEEE International Conference on Data Mining Workshops*. Washington D. C., USA: IEEE Computer Society, 2011:344-349.
- [23] LANCICHINETTI A, FORTUNATO S, RADICCHI F. Benchmark graphs for testing community detection algorithms[J]. *Physical Review E*, 2008, 78(4):046110.
- [24] ZHANG Pan. Evaluating accuracy of community detection using the relative normalized mutual information[EB/OL]. [2018-07-25]. <https://arxiv.org/pdf/1501.03844.pdf>.
- [25] GIRVAN M, NEWMAN M E J. Community structure in social and biological networks[J]. *Proceedings of the National Academy of Sciences of the United States of America*, 2002, 99(12):7821-7826.
- [26] LUSSEAU D. The emergent properties of a Dolphin social network[EB/OL]. [2018-07-25]. <https://arxiv.org/ftp/cond-mat/papers/0307/0307439.pdf>.
- [27] NEWMAN M E J. Finding community structure in networks using the eigenvectors of matrices[J]. *Physical Review E*, 2006, 74(3):036104.
- [28] NELSON D L, MCEVOY C L, SCHREIBER T A. The university of south Florida free association, rhyme, and word fragment norms[J]. *Behavior Research Methods Instruments and Computers*, 2004, 36(3):402-407.

编辑 吴云芳