

文章编号:1008-1542(2017)02-0169-07

doi:10.7535/hbkd.2017yx02011

## 有向网络下的 CoDA 社区发现算法评估

郭松,张冬雯,许云峰,杨玉林,郑雅洁,柳晨光

(河北科技大学信息科学与工程学院,河北石家庄 050018)

**摘要:**CoDA 算法是一种基于概率模型的能识别二分结构的社区发现算法。为了验证该算法的社区划分效果,采用信息检索领域的  $F$ -measure 标准,对有向网络下重叠社区和非重叠社区的 CoDA 社区发现算法进行评估。 $F$ -measure 标准中  $F1$ -measure 值的大小能反映 CoDA 算法社区划分效果的优劣。实验所用的数据集由 LFR Benchmark 工具生成,数据集中节点数最小为 100,最大为 20 000,每增加 100 节点对 CoDA 算法社区划分效果评估一次。分析实验结果可以得出,当节点数小于 1 600 时,CoDA 算法的划分效果较好。当节点数大于 1 600 时,随着节点个数增多,CoDA 算法社区划分效果逐渐变差。由此说明,基于概率模型的 CoDA 算法适用于小规模社交网络社区的划分。

**关键词:**算法理论;社区发现;CoDA 算法;有向网络;评估; $F$ -measure

**中图分类号:**TP391.1      **文献标志码:**A

## Evaluation of the CoDA community detection algorithm based on directed network

GUO Song, ZHANG Dongwen, XU Yunfeng, YANG Yulin, ZHENG Yajie, LIU Chenguang

(School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang, Hebei 050018, China)

**Abstract:** CoDA (Communities through Directed Affiliations) algorithm is a kind of community detection algorithm which can successfully detect 2-mode communities based on probability model. The  $F$ -measure criterion, for information retrieval, is adapted to the evaluation of CoDA algorithm in directed networks with overlapping communities or non-overlapping communities. The value of  $F1$ -measure in  $F$ -measure criterion can reflect whether CoDA algorithm performs well or not. The data sets used in the experiment is generated by the LFR Benchmark tool. The minimum number of nodes in data set is 100 and the maximum is 20 000, and evaluated experiment is conducted when every 100 nodes is added. The results show that CoDA algorithm performs well when the number of nodes is bellow 1 600. CoDA algorithm's performance becomes worse with the increase of the number of nodes, which proves the CoDA algorithm based on probability model is applicable to the community detection of small-scale networks.

**Keywords:** theory of algorithm; community detection; CoDA algorithm; directed networks; evaluate;  $F$ -measure

收稿日期:2016-09-23;修回日期:2017-03-06;责任编辑:陈书欣

基金项目:国家自然科学基金(71271076)

第一作者简介:郭松(1991—),男,河北石家庄人,硕士研究生,主要从事聚类、社区发现方面的研究。

通信作者:张冬雯教授。E-mail:zdwwtx@163.com

郭松,张冬雯,许云峰,等.有向网络下的 CoDA 社区发现算法评估[J].河北科技大学学报,2017,38(2):169-175.

GUO Song,ZHANG Dongwen,XU Yunfeng,et al. Evaluation of the CoDA community detection algorithm based on directed network[J].

Journal of Hebei University of Science and Technology,2017,38(2):169-175.

社区发现是在给定网络中,将物理对象或抽象对象的集合分组为类似对象组成的多个社区。社区发现可以应用在许多领域,例如,在商业领域,社区发现可以用于对不同的客户群分组,寻找潜在市场;在生物领域,社区发现可以用于对动植物分类和对基因分类,获取对种群固有结构的认识;在因特网领域,社区发现可以用来在网上进行文档归类,修正错误信息;在保险领域,社区发现可以用来对保险单持有者的分组识别汽车保险欺诈等。因此,社区发现算法在现实生活中有着广泛的应用。

目前,许多学者从事社区发现研究,由于他们对节点集或者边集进行划分时采取了不同标准,提出了多种多样的社区发现算法。无向网络中的社区发现算法分为图论分区法<sup>[1-4]</sup>、链接分区法<sup>[5-7]</sup>、局部扩张和优化分区法<sup>[8-10]</sup>、模糊划分法<sup>[11-13]</sup>和基于代理的分区法<sup>[14-15]</sup>5类。随着无向网络中社区发现算法的逐渐成熟,一些学者尝试将无向网络社区发现算法扩展到无向网络,以期利用已有的方法解决新问题。例如,无向网络中的 DOCNet 算法采用聚集系数计算节点的重要性,FAGIOLLO<sup>[16]</sup>改进了 DOCNet 算法,并将其扩展到有向网络中。NEWMAN 等<sup>[1]</sup>在无向网络中建立了模块度优化算法——谱二分法,然后将其扩展到有向网络中,并提出有向网络中模块度的概念。NICOSIA 等<sup>[17]</sup>将有向网络中模块度概念进行了扩展,并将重叠社区考虑在内。

随着社区发现算法的不断增多,如何衡量算法的优劣成为了一个重要的问题,因此,对社区发现算法的评估也成为了研究者关注的问题。目前常用的社区发现算法评估方法可分为外部评估法和内部评估法两大类。外部评估法是基于一种预先指定的结构,反映了人们对数据集社区结构的直观认识,主要包括正确率(Accuracy)<sup>[18]</sup>、Rand 系数(Rand Index)<sup>[19]</sup>、Jaccard 系数(Jaccard Index)<sup>[20]</sup>等指标。内部评估法是利用未知社区结构固有特征和量值来评估一个社区发现算法的社区划分效果,主要包括模块度(Modularity)<sup>[1]</sup>和 Coverage<sup>[21]</sup>等指标。

CoDA(communities through directed affiliations)社区发现算法<sup>[22]</sup>是一个新的社区发现算法,它能对一种新型的网络结构进行社区划分。本文主要用  $F$ -measure 标准对 CoDA 社区发现算法进行详细评估,以分析 CoDA 算法在不同节点数据集下社区划分结果的变化规律。

## 1 CoDA 社区发现算法的概述及实现

CoDA 社区发现算法是由斯坦福大学的 Jaewon Yang 等提出的,该算法引起关注的原因是它对社区结构提出了一个更加新颖、全面的定义:有相似的连接模式的节点,即便它们不相连也可以组成一个社区。新的社区结构的概念源于结构对等的社会网络文化,这一文化是指具有社会同质性的节点不仅能够连接到彼此的节点(即内部组连接),还包括以适当方式连接到网络中其他部分的节点(外部连接)。图 1 表示出了两种不同的社区概念。其中传统的内聚社区由社区内部链接占据主导地位,且其内部节点以双向边形式链接;而二分社区由社区间的单向链接占据主导地位。

CoDA 社区发现算法在两方面较为先进:第一,该算法不仅可以发现内聚社区还可以成功地发现二分社区;第二,该算法可以成功发现有向网络中的重叠社区。

设  $B(V, C, M)$  是一个二分图。其中  $V$  是节点集,  $C$  是社区集,  $M$  是将  $V$  和  $C$  连接在一起的边集。该模型可以通过概率  $p(u, v)$  生成从节点  $u \in V$  到节点  $v \in V$  的有向边,生成一个有向图  $G(V, E)$ , 其中  $E$  为边的集合:

$$p(u, v) = 1 - \prod_{k \in C_{uv}} (1 - p_k) \quad (1)$$

通过将有向关系网络模型拟合到  $G(V, E)$  识别社区,即,使用最大似然法找到关系图  $B$  和所有社区  $c \in C$  的概率集合  $\{P_c\}$ 。

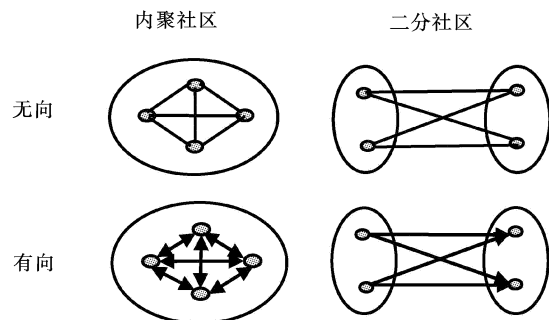


图 1 两种类型的网络和两种类型的社区结构

Fig. 1 Two types of networks and two types of community structures

$$\arg \max_{P, \{P_c\}} \prod_{(u,v) \in E} p(u,v) \times \prod_{(u,v) \notin E} (1 - p(u,v)). \quad (2)$$

取对数得:

$$\arg \max_{P, \{P_c\}} \sum_{(u,v) \in E} \log p(u,v) + \sum_{(u,v) \notin E} \log(1 - p(u,v)). \quad (3)$$

引入变量  $M_{uc}$  和  $L_{vc}$  表示节点成员资格。用  $M_{uc}$  表示节点  $u$  是否具有社区  $c$  的链出成员资格,取值为 0 或者 1;  $L_{vc}$  表示节点  $v$  是否具有社区  $c$  的链入成员资格,取值为 0 或者 1。因此,式(1)可以表示为

$$p(u,v) = 1 - \prod_{c \in C_{uv}} (1 - p_c) = 1 - \prod_c (1 - p_c)^{M_{uc} L_{vc}}, \quad (4)$$

其中  $C_{uv} = \{c | (u,c), (c,v) \in E\}$ , 如果  $C_{uv} = \emptyset$ , 则基础边缘概率为  $p(u,v) = 1/|V|$ 。

$(1 - p_c)^{M_{uc} L_{vc}}$  表示只有当节点  $u$  有社区  $c$  的链出成员资格, 节点  $v$  有社区  $c$  的链入成员资格时取值为  $(1 - p_c)$ , 否则为 1。则  $\prod_c (1 - p_c)^{M_{uc} L_{vc}}$  只要其中一个社区  $c_0$  使  $M_{uc_0} L_{vc_0} = 1$ , 则  $\prod_c (1 - p_c)^{M_{uc} L_{vc}} \neq 1$ , 则  $p(u,v) \neq 0$ 。只要存在一个社区  $c_0$ , 节点  $u$  有其链入成员资格, 节点  $v$  有其链出成员资格, 就有可能存在有向边  $\langle u,v \rangle$ , 且其概率可用式(4) 表示。

令  $1 - p_c = \exp(-a_c)$ , 其中  $a_c \geq 0$ 。代入式(4) 得,

$$p(u,v) = 1 - \exp\left(-\sum_c M_{uc} a_c L_{vc}\right). \quad (5)$$

令  $\tilde{M}_{uc} = \sqrt{a_c} M_{uc}$ ,  $\tilde{L}_{vc} = \sqrt{a_c} L_{vc}$  则可进一步化简等式(5) 为

$$p(u,v) = 1 - \exp\left(-\sum_c \tilde{M}_{uc} \tilde{L}_{vc}\right). \quad (6)$$

因为  $M_{uc} \in \{0,1\}$ ,  $L_{vc} \in \{0,1\}$ , 所以  $\tilde{M}_{uc} \in \{0, \sqrt{a_c}\}$ ,  $\tilde{L}_{vc} \in \{0, \sqrt{a_c}\}$ 。

$\tilde{M}_{uc}$  表示节点  $u$  对社区  $c$  的链出成员关系的强度, 若节点  $u$  属于社区  $c$ , 则其将以因子  $\sqrt{a_c}$  同社区  $c$  中其他成员节点相链接,  $\sqrt{a_c}$  的值决定概率  $p_c$  的大小。同理,  $\tilde{L}_{vc}$  表示节点  $v$  对社区  $c$  的链入成员关系的强度, 若节点  $v$  属于社区  $c$ , 则其将以因子  $\sqrt{a_c}$  同社区  $c$  中其他成员节点相链接。

用非负连续取值的成员关系  $F_{uc}$  和  $H_{vc}$  分别代替  $\tilde{M}_{uc}$  和  $\tilde{L}_{vc}$ 。优点是此时可以选择每个节点与给定社区的成员关系强度:  $F_{uc}$  取值越大, 节点  $u$  指向社区  $c$  中其他成员的链出边越多。同理,  $H_{vc}$  取值越大, 则节点  $v$  指向社区  $c$  中其他成员的链入边越多, 因此可以将式(6) 转换成:

$$P(u,v) = 1 - \exp(-F_u H_v^T). \quad (7)$$

式(3) 可转化为连续优化问题:

$$\{\hat{F}, \hat{H}\} = \arg \max_{F, H \geq 0} l(F, H), \quad (8)$$

其中

$$l(F, H) = \sum_{(u,v) \in E} \log(1 - \exp(-F_u H_v^T)) - \sum_{(u,v) \notin E} F_u H_v^T. \quad (9)$$

使用最大似然法求出  $l(F, H) = \log P(G | F, H)$ , 取得最大值时的非负关系矩阵, 来拟合有向关系网络模型, 以达到社区划分的目的。

## 2 评估方法

评估方法借鉴了信息检索领域中的  $F$ -measure 标准<sup>[23]</sup>。  $F$ -measure 包含准确率和召回率两方面。其中准确率是检索出相关文档数与检索出的文档数的比率, 衡量系统的查准率; 召回率是指检索出的相关文档数和文档库中所有相关文档数的比率, 衡量系统的查全率。本文将  $F$ -measure 标准中的文档数替换为节点对个数, 进而将  $F$ -measure 标准应用到社区发现算法的评估领域。  $F$ -measure 可以很好地刻画某社区发现算法对社交网络的社区划分结果与标准数据集中社区划分结果的差异, 在此用  $F$ -measure 标准对 CoDA 社区发现算法进行评估。

社区中节点对的归属情况见表 1。其中:  $a$  表示原始数据集中属于同一个真实社区, 社区划分后也属于同一个社区中的节点对数目;  $b$  表示原始数据集中不属于同一个真实社区, 社区划分后却划分到同一个社区中的节点对数目;  $c$  表示原始数据集中属于同一个真实社区, 社区划分后却未被划分到同一个社区中的节点

表1 社区划分结果

Tab.1 Results of community division

节点对归属	原始数据集中属于真实社区的节点对数目	原始数据集中不属于真实社区的节点对数目
社区划分后属于某社区的节点对数目	$a$	$b$
社区划分后不属于某社区的节点对数目	$c$	$d$

对数目; $d$ 表示原始数据集中即不属于同一个真实社区,社区划分后也不属于同一个社区中的节点对数目。

准确率(Precision)是输出结果中被划分到同一社区的2个节点组成的节点对在数据集中属于同一真实社区的比例,计算公式如式(10)所示:

$$P = \frac{a}{a+b} \quad (10)$$

召回率(Recall)是输出结果中被划分到同一社区中的2个节点且在数据集中属于同一真实社区的节点对占划分结果中社区内所有节点对的比例,计算公式如式(11)所示:

$$R = \frac{a}{a+c} \quad (11)$$

准确率与召回率都只能反映算法单方面的性质,且这2个参数往往有相反的变化趋势,而F1-measure将准确率与召回率进行综合考虑,能够反映算法的综合性能。其定义如式(12)所示:

$$F1\text{-measure} = \frac{2 \times P \times R}{P + R} \quad (12)$$

## 2.1 评估程序流程图

评估程序流程图如图2所示,本程序采用Java语言实现。该评估方法为外部评估法,其特点是标准社区结构已知。首先,读取数据集中的标准社区结构文件,得到标准社区结构中社区编号与其包含节点的对应关系HashMap<String, LinkedList<String>>communities。其次,根据得到的标准社区情况communities分社区遍历,构造处于同社区的节点对nodePairInReal。然后,读取被某算法处理后的社区

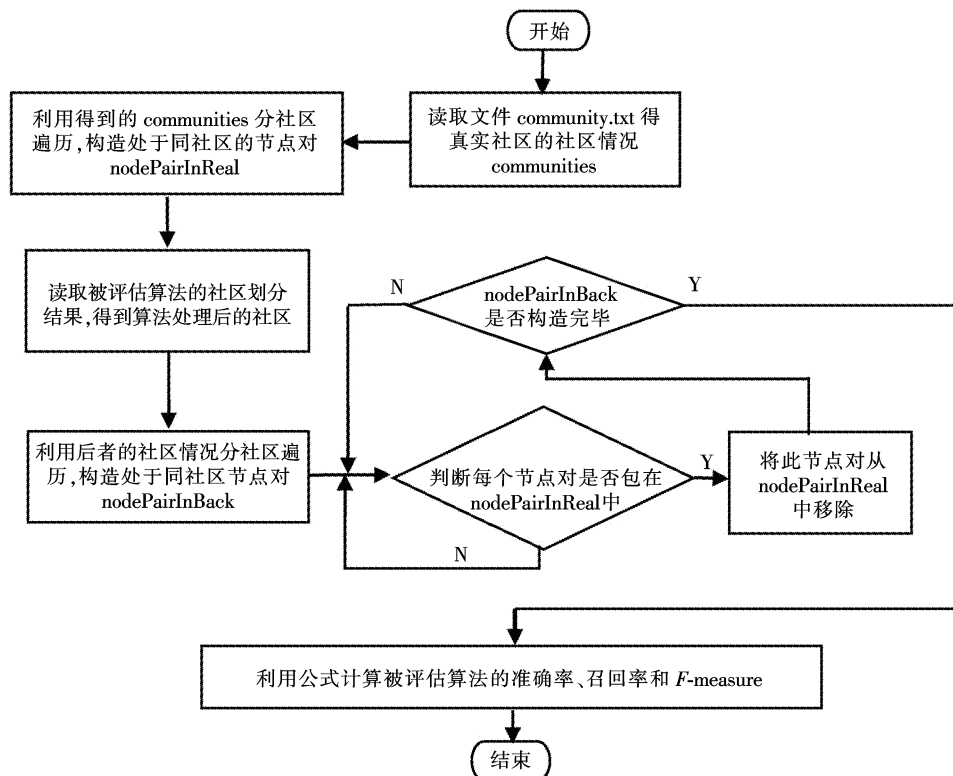


图2 评估程序流程图

Fig.2 Flow chart of evaluation procedure

情况,构造处理后社区中的节点对  $\text{nodePairInBack}$ ,在此过程中判断每个节点是否包含在真实社区的节点对  $\text{nodePairInReal}$  之中,如果存在,则构造既包含在真实社区中又在某算法处理后的社区中的节点对  $\text{nodePairCorrect}$ ,同时,将此节点对从  $\text{nodePairInReal}$  中移除。其中节点对指位于同一社区中的 2 个节点的组合。其中  $\text{nodePairCorrect}$  对应表 1 中  $a$ ,  $\text{nodePairInBack}$  对应表 1 中  $b$ ,  $\text{nodePairInReal}$  对应表 1 中  $c$ 。最后根据式(10)~式(12)依次计算出准确率、召回率和  $F1$ -measure 的值。

## 2.2 有向网络下重叠社区 CoDA 算法的评估

实验所用有向网络下重叠社区的标准数据集由 LFR Benchmark<sup>[24]</sup> 生成。数据集中节点平均度数为 10,最大度数为 50,混合参数为 0.2,社区最小尺寸为 20,最大尺寸为 100,40%的重叠节点,且每个重叠节点同时属于 4 个社区。节点数从 100 到 20 000,每 100 个节点进行一次测试。评估结果如图 3 和图 4 所示。其中图 3 表示链入成员的评估结果,图 4 表示链出成员的评估结果。

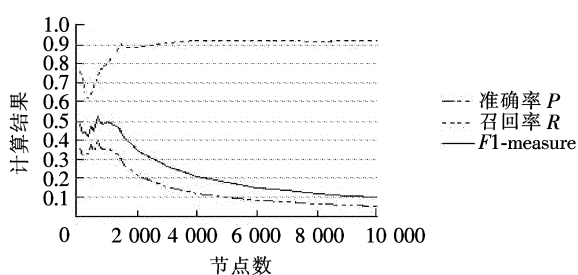


图 3 链入成员评估结果图

Fig. 3 Evaluation result of incoming members

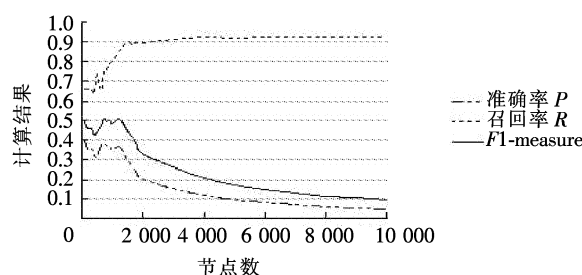


图 4 链出成员评估结果图

Fig. 4 Evaluation result of outgoing members

通过实验结果可以发现,在有向网络带重叠社区的情况下,相同节点数据集的链入成员资格和链出成员资格之间的准确率、召回率和  $F1$ -measure 几乎相等。数据集节点数在 1 600 以内时,CoDA 社区发现算法的社区划分效果比较好, $F1$ -measure 在 0.4 以上。链入成员在节点为 700 时 CoDA 社区划分效果最好,此时  $F1$ -measure 达到 0.524;链出成员在节点为 1 200 时 CoDA 社区划分效果最好,此时  $F1$ -measure 达到 0.506。当数据集节点数超过 1 700 时,随着节点的不断增多,召回率不断升高,且趋近于 1;但是准确率不断下降,且趋近于 0.1; $F1$ -measure 呈明显下降趋势,其值趋近与 0.1。当节点数超过 10 000 时, $F1$ -measure 一直处于 0.1 以下,故图中未给出评估结果。

## 2.3 有向网络下非重叠社区 CoDA 算法的评估

数据集生成过程及数据集节点数的选取同 2.2,节点平均度数、最大度数、混合参数、社区最小尺寸和最大尺寸等没有改变,只将重叠节点的个数和每个重叠节点属于的社区数调整为 0。评估结果如图 5 和图 6 所示。

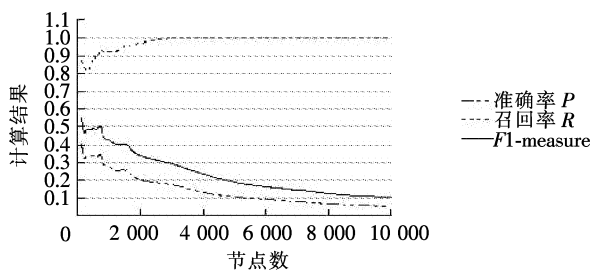


图 5 链入成员评估结果图

Fig. 5 Evaluation result of incoming members

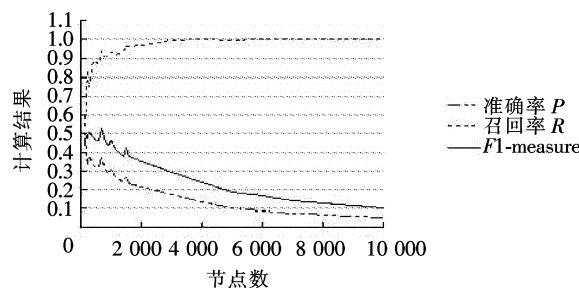


图 6 链出成员评估结果图

Fig. 6 Evaluation result of outgoing members

通过对有向网络下非重叠社区 CoDA 算法的评估可以发现,在相同节点数据集下,链入成员资格和链出成员资格之间的准确率、召回率和  $F1$ -measure 几乎相等。随着节点的不断增多,召回率不断升高,准确

率和  $F1$ -measure 在不断下降。这说明随着节点的增多,标准社区结构中不在同一社区的节点对经过 CoDA 算法处理后,被划分到同一社区的数量越来越多。标准社区结构中在同一社区的节点对经过 CoDA 算法处理后,被划分到 2 个不同社区的数量越来越少。到达 10 000 节点后,几乎所有属于同一  $c_i$  社区的节点对都不会被划分到不同的社区内。数据集节点数量较小时,其综合性能较好;随着数据集节点数量的增多,其综合性能在逐渐变差。

分别对比有向网络下的重叠社区和非重叠社区的 CoDA 社区发现算法实验结果可知,相同节点的数据集下,重叠社区和非重叠社区之间的准确率、召回率和  $F1$ -measure 相差很小,这说明 CoDA 社区发现算法在分别处理重叠社区和非重叠社区时,其稳定性非常好。

### 3 结 论

1) 将信息检索领域中的  $F$ -measure 评价标准应用于社区发现算法的评估领域。既属于标准社区又属于某算法划分后,社区中的节点对数对应  $F$ -measure 评价标准中检索出的相关文档数替换为算法划分后,社区中的节点对数对应该标准中检索出的文档数,真实社区中的节点对数对应该标准中文档库中所有相关文档数。通过实验可知, $F$ -measure 可以很好地刻画某社区发现算法对社交网络的社区划分结果与标准数据集中社区划分结果的差异,因此可以将该标准应用到有向网络下重叠社区和非重叠社区的挖掘算法评估领域。

2) 在研究社交网络结构的过程中,一个准确的社区概念至关重要。传统模型认为“社区”是密集连接节点的集合。CoDA 算法考虑了二分社区结构,它将一组在网络中没有直接链接,但链接在同一个坐标的节点组合在一起。该算法能识别有向网络下的重叠二分社区。当数据集节点数小于 1 600 时,CoDA 算法社区划分效果较好且稳定,但是当节点数超过该值时,随着数据集节点数的增多,社区划分效果逐渐变差。因此,基于概率模型的 CoDA 社区发现算法适用于数据集规模较小时社交网络的划分,不适用于大规模社交网络的社区划分。

### 参考文献/References:

- [1] NEWMAN M E J, GIRVAN M. Finding and evaluating community structure in networks[J]. *Physical Review E*, 2004, 69:026113.
- [2] YANG J, LESKOVEC J. Overlapping community detection at scale: A nonnegative matrix factorization approach[C]// *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*. New York:ACM,2013:587-596.
- [3] MAKINO K, UNO T. New algorithms for enumerating all maximal cliques[C]//9th Scandinavian Workshop on Algorithm Theory. Berlin:Springer, 2004:260-272.
- [4] PALLA G, DERÉNYI I, FARKAS I, et al. Uncovering the overlapping community structure of complex networks in nature and society [J]. *Nature*, 2005, 435:814-818.
- [5] EVANS T S, LAMBIOTTE R. Line graphs, link partitions, and overlapping communities[J]. *Physical Review E*, 2009, 80:016105.
- [6] AHN Y Y, BAGROW J P, LEHMANN S. Communities and hierarchical organization of links in complex networks[J]. *Nature*, 2009, 466:761-764.
- [7] WU Zhihao, LIN Youfang, WAN Huaiyu, et al. A fast and reasonable method for community detection with adjustable extent of overlapping[C]// 2010 International Conference on Intelligent Systems and Knowledge Engineering (ISKE). [S. l.]: IEEE Xplore, 2010:376-379.
- [8] ANDREA L, FILIPPO R, RAMASCO J J, et al. Finding statistically significant communities in networks[J]. *Inorganic Materials*, 2002, 38(4):336-338.
- [9] LANCICHINETTI A, FORTUNATO S, KERTÉSZ J. Detecting the overlapping and hierarchical community structure of complex networks[J]. *New Journal of Physics*, 2008, 11(3):19-44.
- [10] SHEN Huawei, CHENG Xueqi, CAI Kai, et al. Detect overlapping and hierarchical community structure in networks[J]. *Physica A: Statistical Mechanics & Its Applications*, 2009, 388(8):1706-1712.
- [11] LU Yinghua, MA Tinghui, YIN Changhong, et al. Implementation of the fuzzy C-means clustering algorithm in meteorological data[J]. *International Journal of Database Theory & Application*, 2013, 6(6):1-18.
- [12] NEPUSZ T, PETRÓCZI A, NÉGYESSY L, et al. Fuzzy communities and the concept of bridgeness in complex networks[J]. *Physical Review E Statistical Nonlinear & Soft Matter Physics*, 2008, 77:016107.
- [13] PSORAKIS I, ROBERTS S, EBDEN M, et al. Overlapping community detection using Bayesian non-negative matrix factorization[J]. *Physical Review E Statistical Nonlinear & Soft Matter Physics*, 2011, 83:066114.

- [14] XIE J, SZYMANSKI B K, LIU X M. SLPA: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process[C]//2011 IEEE 11th International Conference on Data Mining Workshops (ICDMW). Vancouver: IEEE, 2011: 344-349.
- [15] RHOUMA D, ROMDHANE L B. An efficient algorithm for community mining with overlap in social networks[J]. Expert Systems with Applications, 2014, 41(9):4309-4321.
- [16] FAGIOLO G. Clustering in complex directed networks[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2007, 76:026107.
- [17] NICOSIA V, MANGIONI G, CARCHIOLO V, et al. Extending the definition of modularity to directed graphs with overlapping communities[J]. Journal of Statistical Mechanics Theory & Experiment, 2009 (3):3166-3168.
- [18] STEINHAEUSER K, CHAWLA N V. Identifying and evaluating community structure in complex networks[J]. Pattern Recognition Letters, 2010, 31(5):413-421.
- [19] RAND W M. Objective criteria for the evaluation of clustering methods[J]. Journal of the American Statistical Association, 1971, 66: 846-850.
- [20] BEN-HUR A, ELISSEEFF A, GUYON I. A stability based method for discovering structure in clustered data[J]. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing, 2002(7):6-17.
- [21] BRANDES U, GAERTLER M, WAGNER D. Experiments on Graph Clustering Algorithms[M]. Berlin: Springer, 2003:568-579.
- [22] YANG J, MCAULEY J, LESKOVEC J. Detecting cohesive and 2-mode communities in directed and undirected networks[C]// Proceedings of the 7th ACM International Conference on Web Search and Data Mining. New York:WSDM, 2014:323-332.
- [23] 万里. 时间序列中的知识发现:基于频繁模式发现的分类和聚类方法研究[D]. 北京:北京邮电大学, 2009.  
WAN Li. Knowledge Discovery in Time Series:Researches on Classification and Clustering Based on Frequent Pattern Discovery[D]. Beijing: Beijing University of Posts and Telecommunications, 2009.
- [24] LANCICHINETTI A, FORTUNATO S. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2009, 80:016118.