

# 多关系网络社团发现算法

黄新宇, 陈东明, 任涛  
(东北大学 软件学院, 辽宁 沈阳 110169)

**摘 要:** 分析了真实社会网络的特性,建立了节点间多关系网络模型.在此基础上定义了节点间相互作用的影响力等概念,提出了适用于多关系网络的社团发现算法.通过理论验证了相关定义的合理性,并针对多关系网络进行了对比实验.实验结果表明:所提出的多关系网络社团发现算法与其他经典算法相比具有较高的精确度和较低的时间复杂度,具有重要的研究意义及实用价值.

**关 键 词:** 在线社会网络;多关系网络;社团发现;异质网络;重叠社团

**中图分类号:** TP 391      **文献标志码:** A      **文章编号:** 1005-3026(2018)10-1375-05

## Community Discovery Algorithm for Multi-relationship Networks

HUANG Xin-yu, CHEN Dong-ming, REN Tao  
(School of Software, Northeastern University, Shenyang 110169, China. Corresponding author: CHEN Dong-ming, E-mail: chendm@mail.neu.edu.cn)

**Abstract:** The characteristics of the real social network were analyzed, and a network model with multi-relationships network between nodes was established. Based on this, the concept of influence of interaction between nodes was defined, and a community discovery algorithm for multi-relationship network (CDMN) was proposed. The rationality of the presented definitions was verified theoretically, and a series of experiments compared with other algorithms were conducted. Experimental results showed that the proposed community discovery algorithm for multi-relationship networks has higher accuracy and lower time complexity than other classical algorithms, and has important research significance and practical value.

**Key words:** online social network; multi-relationship network; community discovery; heterogeneous network; overlapping community

随着信息技术的飞速发展,社会网络已经成为人们日常生活中的一部分,国内外学者对此研究也越来越深入.传统的复杂网络研究大都对网络简化处理,形成简单的图模型进行信息挖掘.然而,社会网络在真实环境中更为复杂,个体间具有多种关系,对应网络的多个维度,也就构成了异质网络<sup>[1]</sup>.异质网络就是网络中节点和关系具有多重属性,对外表现出不同类型的节点和关系,构成了非同质不均匀的复杂网络结构.而本文研究的多关系网络是异质网络问题中的一种特例,即存在于多个网络中的节点是互相对应的同类节点,节点间的关系是多种多样的.

国内外学者主要从网络重叠结构、动态变化等方面对此类问题展开研究,研究方法多采用指数分布随机图模型、关系代数和分块模型等方法<sup>[2-3]</sup>.Xie等<sup>[4]</sup>针对网络重叠程度提出了不同方法,并对真实社会网络重叠情况进行概括,指出约30%的节点是属于2~3个社团的重叠节点,在重叠社团发现问题上具有重要的参考价值,但是在算法普适性方面有待优化.Cazabet等<sup>[5]</sup>将动态网络划分成多个时间切片,然后将每个时间切片对应的静态网络进行社团划分,再将结果进行

合并,能够给出较为准确的动态网络划分结果,但也揭示出算法依赖于处理时间先后顺序,对于大规模网络的动态划分问题存在改进空间.李慧嘉等<sup>[6]</sup>利用动态迭代技术提出了一种动态网络社团划分指标函数的一般化形式,使动态网络社团划分不需要任何参数就可以得到最优化的结果,但在多关系网络处理方面存在不足.

节点重要性测度是网络研究中的一项基本指标,国内外学者对此有很多研究基础<sup>[7-9]</sup>.就现有方法来说,最简单的办法就是考虑节点的度,即节点的度越大该节点越重要.这种方法的缺点是容易漏掉那些在多个社团之间起桥梁作用的节点,虽然这些节点的度不大,但它们对于多个社团聚合成大型社团至关重要.因此,介数的定义被提出,但因为其计算复杂性,对结构无法预知的网络往往只能采取近似计算.除了聚类系数、 $k$ -核等经典指标,也有学者提出基于 PageRank 改进的 NodeRank 指标,用于评价大规模加权网络的节点重要性测度,并验证了其可靠性<sup>[10]</sup>.大多数现有评价指标不是单纯从网络拓扑角度出发考虑节点重要性,就是从某个现实领域出发给出对节点重要性的理解.事实上节点的重要性同时受多种因素影响,从多个角度考虑节点重要性的方法能获得更好的效果.

综上所述,到目前为止异质网络的研究方兴未艾,随着大数据时代的到来,复杂多元化网络分析具有非常重要的研究意义.

## 1 模型与方法

在社会网络中,每个人在不同时期担负着不同的角色,并且大部分肩负多种角色,各个角色都可以看作一种社会关系,多种社会关系构成的多个网络通过个人的参与映射成了一个多关系网络.在多关系网络中,人们受到社会环境的影响,却能够在社会环境中保持稳定状态,正如经典物理学中力的合成,如图 1 所示.

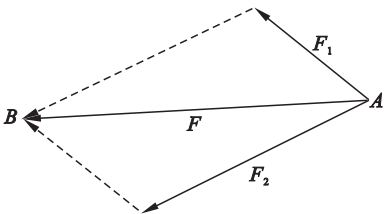


图 1 物理学中力的合成  
Fig. 1 Synthesis of forces in physics

在多关系网络中,假设节点  $A$  和节点  $B$  代表

2 个人,既是同学关系也是同事关系,那么在同质网络中可能仅仅能够表示出  $A$  和  $B$  之间存在连边,但在多关系网络中就能够呈现出更细致的关系.本文建立的网络模型具有多种关系类型和对应的权值,表述如下:

$$G = (V, E, R, W). \tag{1}$$

式中: $G$  为网络模型; $V$  为节点集合; $E$  为关系集合; $R$  为关系类型; $W$  为关系对应的权值.

为了描述节点间多种关系的相互作用,本文定义节点影响力  $F$  的计算式为

$$F_i = \frac{N(i)}{N(R)} \cdot \sum_{j=1}^{d_i} \left( \frac{w_{ij}}{d_j} \right). \tag{2}$$

式中: $N(i)$  为节点的连边中具有的关系种类个数; $N(R)$  为关系种类总数; $w_{ij}$  为由  $i$  和  $j$  构成的关系权值; $d_i$  表示节点  $i$  的度; $j$  表示节点  $i$  的邻居节点; $d_j$  表示节点  $j$  的度.在单一关系网络构成的简单图中,节点影响力定义可简化为

$$F_i = \sum_{j=1}^{d_i} \left( \frac{1}{d_j} \right). \tag{3}$$

节点的影响力取决于其邻居节点的度及节点自身的度.社团由节点构成,因此,社团  $C$  的影响力  $F_C$  为

$$F_C = \sum_{i=1}^N (F_i). \tag{4}$$

式中: $N$  为社团内所包含的节点数; $F_C$  为当前社团内所有节点影响力之和.为了描述社团稳定程度,本文给出社团进化的加速度  $A_C$ :

$$A_C = \frac{F_C}{N}. \tag{5}$$

由于完全图构成的社团结构最为稳定,可以通过数学归纳法得出:社团内节点数不超过 3 个时,即  $N \leq 3$ ,社团进化加速度取值为 1 时最稳定;通常情况下,社团内节点数大于 3 时,社团进化加速度取值趋向  $\frac{1}{N-1}$ ,社团结构最为稳定.

此外,重叠节点连接多个社团,在社团划分问题中尤为重要.对于网络中连接多个社团的重叠节点  $o$ ,与社团  $C$  之间的判定规则  $R$  如下:

$$R = \min \left( \frac{\Delta A_{o,C}}{I(o,C)} \right). \tag{6}$$

式中: $\Delta A_{o,C}$  为节点  $o$  加入社团  $C$  之后社团进化加速度的增量; $I(o,C)$  表示节点  $o$  与社团  $C$  之间已有的连边数.判定规则  $R$  可以解释为:重叠节点应划分到这样的社团,使得节点  $o$  加入该社团后引起进化加速度增量与节点和社团间连边数的比值最小.下面将给出多关系网络社团划分算法

CDMN 的具体过程.

## 2 算法描述

假设给出参数  $k$  用于表示需要得到的社团数目,本文设计的 CDMN 算法执行步骤如下.

步骤 1 数据预处理,进行网络初始化并计算各节点的影响力  $F_i$ ,得到节点影响力的有序列表;

步骤 2 获取网络中节点影响力较大的  $k$  个节点,初始化  $k$  个社团,并将这些节点的非重叠邻居节点划入社团  $C_i$ ;

步骤 3 将网络内其他非重叠节点划分到与已有社团直接相连的社团内;

步骤 4 针对网络中剩余的重叠节点,采用式(6)中的判定规则  $R$  进行判定,划分到相应的社团.

CDMN 算法执行的伪代码如下所示.

```
Algorithm: CDMN
INPUT: 网络数据集
OUTPUT: 社团划分之后的节点集合
// 初始化节点,计算影响力
For n in Nodes do
    Calculate  $F_i$ 
End For
// 按节点影响力大小排序,初始化社团
sortedNodes = Sort Nodes by  $F_i$ 
initCommunities = sortedNodes[0:k]
For i in k:
    Put n into Community i
    Put n's non-overlapping neighbors into
Community i
End For
// 将非重叠节点划入临近社团
For n in non-divided sortedNodes do
    Calculate  $A_c$  of Communities.
    Divide n into adjacent Community.
End For
// 将重叠节点按式(6)的  $R$  规则划分
For o in overlapping non-divided sortedNodes:
    Divide o into Community according rule  $R$ 
End For
```

首先,CDMN 算法所建立的数学模型  $G = (V, E, R, W)$  更加贴近真实社会网络环境,能够较好地利用网络原始信息;其次,算法参考经典物理学中力的合成等方法,定义了节点的影响力,并通过数学归纳法计算出社团结构最稳定状态的情

况,更有说服力;最后,本文所提出的算法在社团划分中避免使用模块度最优化方法,减少了计算量并避免了模块度限制的分辨率问题,并对重叠节点着重考虑,社团划分结果更为合理.

假设网络中有  $N$  个节点,CDMN 算法对给定网络中节点只需要执行一次遍历即可计算出各个节点的影响力,在社团划分过程中需要再执行一次遍历,得到社团个数为  $k$ ,考虑重叠节点的划分问题,总体时间复杂度为  $O(N + N + kN)$ ,由于  $k$  通常远远小于  $N$ ,因此可近似化简为  $O(N)$ ,优于 FastNewman 算法时间复杂度.由于处理重叠节点的划分问题,CDMN 算法的时间复杂度高于 LPA 算法.在空间复杂度方面,CDMN 算法存储网络使用树结构,空间复杂度为  $O(N)$ ,最终转化为邻接矩阵形式,与 FastNewman 算法、LPA 算法的矩阵存储无明显差异.

## 3 实验分析

实验操作系统环境为 Windols 10 × 64 教育版,内存为 8 G DDR4, CPU 为 Inter (R) Core (TM) i7 - 4790 @ 3.6 GHz,编程语言为 Python 2.7.13.

为了验证算法的有效性,本文首先采用经典网络数据集空手道俱乐部网络<sup>[11]</sup>进行实验,通过 CDMN 算法读取网络中的节点,计算其影响力,结果如表 1 所示.

表 1 节点影响力计算结果			
Table 1 Calculation results of nodes' influence			
Node_id	$F_i$	Node_id	$F_i$
33	5.766 7	16	0.500 0
0	5.194 4	13	0.499 1
32	3.725 5	30	0.453 3
1	2.362 5	7	0.440 3
⋮	⋮	⋮	⋮
8	0.554 7	11	0.062 5

设定参数  $k = 2$ ,执行 CDMN 算法得到 2 个以 0 号节点(Mr. Hi)和 33 号节点(Officer)为核心的 2 个社团,模块度为 0.210 6,略高于实际情况的模块度<sup>[12]</sup>(0.204 0),划分结果如图 2 所示.

由图 2 可以看出:网络划分结果与俱乐部后来的发展趋势一致,验证了算法的有效性.选取“9·11”恐怖袭击网络数据集<sup>[13]</sup>验证 CDMN 算法,结果如图 3 所示.

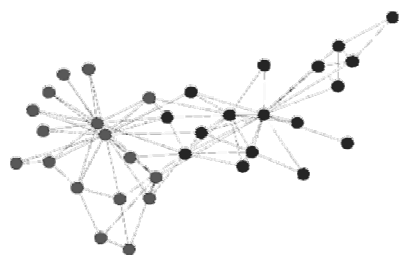


图 2 空手道俱乐部网络实验结果  
Fig. 2 Experiment results of karate club network

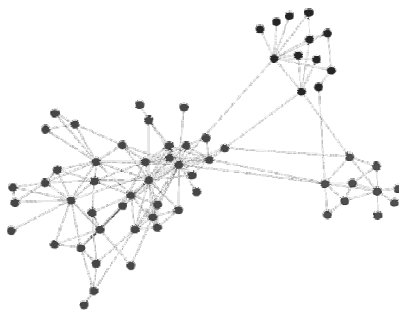


图 3 “9·11”袭击网络实验结果  
Fig. 3 Experiment results of “9·11” attack network

通过计算,网络数据集中用户的影响力计算结果如表 2 所示.

表 2 “9·11”袭击网络节点影响力计算结果  
Table 2 Calculation results of nodes' influence in “9·11” attack network

Node_id	Node_name	$F_i$
0	Djamal Beghal	5.628 8
1	Essid Sami Ben Khemais	5.159 9
2	Mohamed Atta	4.158 3
3	Ahmen Hannan	2.973 6
4	Nawaf Alhazmi	2.777 5
5	Mamoun Darkazanli	2.594 7
6	Usman Bandukra	2.000 0
7	Raed Hijazi	1.970 5
8	Karim Koubriti	1.868 9
9	Ahmed Alnami	1.790 9
⋮	⋮	⋮
6	Zakariya Essabar	0.045 5

通过表 2 给出网络中影响力较大的用户在各  
自所在的社团内处于核心位置,更容易策动反动  
活动,在实际应用中具有重要意义.

在目前公开的网络数据集中,很少有涉及网  
络中节点存在多关系的情况.作者构建了 1 个在  
线社交网络平台,实现了注册登录、组织活动、交  
流分享的平台,并邀请了 200 多名学生注册使用.

该平台中多种不同社交活动表示网络中的多种关  
系,运行 CDMN 算法结果如图 4 所示.

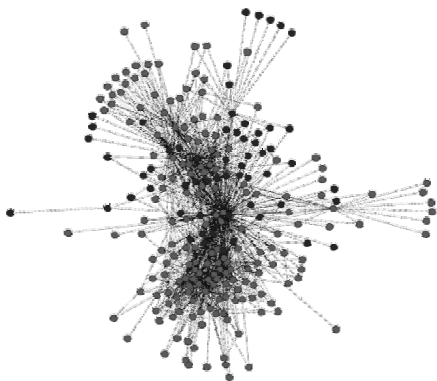


图 4 本文构建的社交网络实验结果  
Fig. 4 Experiment results of the established social  
network proposed in this paper

通过与真实网络对比,发现网络中影响力较  
大的节点大都是平台中负责推广的老师和学生,  
因为其邀请很多其他用户,在网络中具有较高的  
影响力,因此更容易形成社团的核心.

此外,本文选取 LPA 算法、FastNewman 算法  
与 CDMN 算法进行对比,设定相同参数及数据  
集,为了兼顾算法运行时间与划分效果,LPA 算  
法迭代次数参数设定为 1000,CDMN 算法与其他  
算法运行时间对比如图 5 所示.

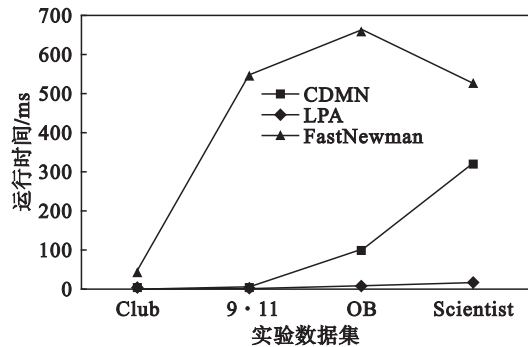


图 5 算法运行时间对比  
Fig. 5 Comparison of algorithms' running time

从图 5 可知,CDMN 算法与 LPA 算法运行时  
间基本在同一数量级,由于对重叠节点的判定,故  
运行时间略多于 LPA 算法,但是远少于  
FastNewman 算法的运行时间. CDMN 算法与其  
他算法的模块度对比如图 6 所示.

从图 6 可知,本文所提出的 CDMN 算法划分  
效果优于 FastNewman 算法和 LPA 算法. 综上述,  
本文所提出的算法具有较好的划分效果和较  
低的时间复杂度.



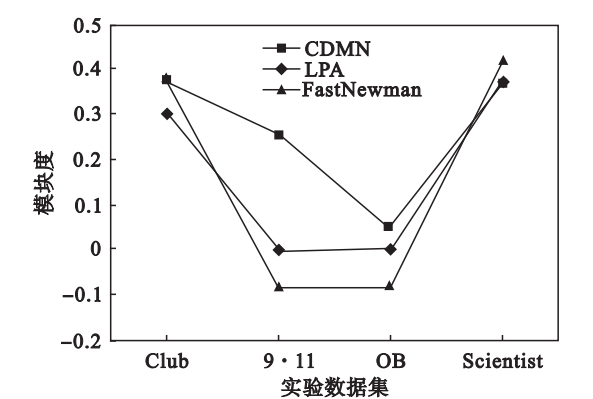


图 6 算法模块度对比实验

Fig. 6 Comparison of algorithms' modularity

## 4 结 语

本文建立了节点间多关系的网络模型,在此基础上,定义了节点的影响力等概念,并通过数学归纳法给出社团稳定状态的评判标准,对于网络中的重叠节点着重考虑,给出了较好的划分结果.为了深入研究多关系网络,本文设计了社交网络实验平台,并设计多种交互方式用于收集多关系网络数据进行实验.对比实验结果表明本文所提出的算法在较低的时间复杂度下获得了较好的划分效果.

### 参考文献：

[ 1 ] Du S, Niu K, He Z, et al. Community detection analysis of heterogeneous network [ C ]// International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery. Xi'an, 2015: 509 – 512.

[ 2 ] Zhang H, Wang C D, Lai J H, et al. Modularity in complex multilayer networks with multiple aspects: a static perspective [ J ]. *Applied Informatics*, 2017, 4(1) : 1 – 7.

[ 3 ] Bródka P, Kazienko P. Multilayer social networks [ M ]. New York: Springer, 2017.

[ 4 ] Xie J, Kelley S, Szymanski B K. Overlapping community detection in networks: the state-of-the-art and comparative study [ J ]. *ACM Computing Surveys*, 2011, 45(4) : 1 – 35.

[ 5 ] Cazabet R, Amblard F. Dynamic community detection [ M ]. New York: Springer, 2014.

[ 6 ] 李慧嘉, 李爱华, 李慧颖. 社团结构迭代快速探测算法 [ J ]. *计算机学报*, 2017, 40(4) : 970 – 984.

( Li Hui-jia, Li Ai-hua, Li Hui-ying. Fast community detection algorithm via dynamical iteration [ J ]. *Chinese Journal of Computers*, 2017, 40(4) : 970 – 984. )

[ 7 ] Ai X. Node importance ranking of complex networks with entropy variation [ J ]. *Entropy*, 2017, 19(7) : 303 – 310.

[ 8 ] Lü L, Chen D, Ren X L, et al. Vital nodes identification in complex networks [ J ]. *Physics Reports*, 2016, 650: 1 – 63.

[ 9 ] 韩忠明, 陈炎, 刘雯, 等. 社会网络节点影响力分析研究 [ J ]. *软件学报*, 2017, 28(1) : 84 – 104.

( Han Zhong-ming, Chen Yan, Liu Wen, et al. Research on node influence analysis in social networks [ J ]. *Journal of Software*, 2017, 28(1) : 84 – 104. )

[ 10 ] Pan Y, Tan W, Chen Y. The analysis of key nodes in complex social networks [ C ]// Cloud Computing and Security of the Third International Conference. Nanjing, 2017: 829 – 836.

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

[ 12 ] Newman M E, Girvan M. Finding and evaluating community structure in networks [ J ]. *Physical Review E*, 2004, 69(6) : 026113.

[ 13 ] SNABook. 9\_11\_edgelist. txt [ EB/OL ]. ( 2011 – 08 – 17 ) [ 2017 – 06 – 27 ]. [https://github.com/maksim2042/SNABook/blob/master/chapter4/9\\_11\\_edgelist.txt](https://github.com/maksim2042/SNABook/blob/master/chapter4/9_11_edgelist.txt).