基于多层节点相似度的社区发现方法

张 虎 吴永科 杨陟卓 刘全明
(山西大学计算机与信息技术学院 太原 030006)

摘 要 社区发现是复杂网络研究中的一项重要研究内容,基于节点相似度的凝聚方法是一种典型的社区发现方法。针对现有节点相似度计算方法中存在的不足,提出一种基于多层节点的节点相似度计算方法,该方法既可以有效地计算节点之间的相似度,又可以解决节点相似度相同时的节点合并选择问题。进一步基于这种改进的节点相似度计算方法和团体之间的连接紧密度度量准则构建社区发现模型,并在真实世界的网络上进行社区发现实验。与GN算法、Fast Newman算法和改进的标签传播算法的实验结果相比,该模型可以更加准确地找到各个社区的成员。

关键词 节点相似度,社区发现,复杂网络
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.01.038

Community Detection Method Based on Multi-layer Node Similarity

ZHANG Hu WU Yong-ke YANG Zhi-zhuo LIU Quan-ming
(School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China)

Abstract Community detection is an important research content in complex network, and the agglomerative method based on the node similarity is a typical method of community detection. Aiming at the shortages of the existing method for calculating the node similarity, this paper proposed a novel method based on the multi-layer node similarity, which can not only calculate the similarity between nodes more efficiently, but also solve the problem of merging nodes when the node similarity is same. Furthermore, this paper constructed the community detection model based on the improved calculation method of the node similarity and the measure criteria of connection tightness between groups, and conducted the community detection experiments in real world network. Compared with the experimental results of GN algorithm, Fast Newman algorithm and the improved label propagation algorithm, the proposed model can be more accurate to find the members of each community.

Keywords Node similarity, Community detection, Complex network

1 引言

随着以互联网为代表的网络信息技术的迅速发展,人类社会已经迈入了复杂网络时代。当前,人们生活在一个充满各种类型的复杂网络世界,如www网络、电力与交通网络、经济与金融网络、社会网络等,这些网络中充斥着各种不同的关系和结构。一个社会网络是一群人或团体按照某种关系连接在一起而构成的一个系统,这里的关系可以是多种多样的,如个人之间的朋友关系、同事之间的合作关系、家庭之间的联姻关系和公司之间的商业关系等。随着对网络性质的深入研究,人们发现许多实际网络都有明显的社区结构[1],每个社区内部的节点之间的连接则相对较为紧密,各个社区之间的连接相对比较稀疏。基于这些特征,人们对不同的社会网络探索了很多实用的社区发现算法。尽管社区发现的研究在复杂网络中有很长一段时间,但到目前还没有一个严格的定义。Huang等人[2]基于局部拓扑性质、全局拓扑性质和节点属性相似度3个方面提出了一种社区的定义。Luce等人[3]将社区定义为完全子图,其要求节点都两两相连,该条件过于严格,实际意义很小。Alba等人[4-6]对限制条件进行了不同程度的弱化,给出了3个不同的社区定义,即n-cliques,n-clan与n-club。Seidman等人[7-8]通过子图内的节点邻节性来定义社区,即社区内的节点与社区内其他节点的连接数大于某个阈值。Luccio和Radicchi等人[9-10]认为仅仅考虑社区的内部属性是不够的,通过簇内与簇间的比较给出社区的限制条件,社区内任一节点的内部度大于其外部度或社区内任一节点的内部度大于其与任一其他社区的连接数。Guimera等人[11]通过社区图与其对应的零模型提出了基于全局拓扑性质的社区定义,认为社区内部边数要大于随机模型中的期望内部边数。

目前,针对社区发现问题提出的社区发现算法主要有以下几类:1)基于网络拓扑结构的算法,主要有Pothen等人[12]在20世纪90年代提出的基于谱分析方法的社区结构探测算法和Girvan等人[13]提出的GN层次聚类算法。2)基于网络

到稿日期:2016-12-06 返修日期:2017-05-23 本文受国家高技术研究发展计划(863计划)项目(2015AA015407),国家自然科学基金项目(61432011,61502287,61673248),山西省自然科学基金项目(201601D102030),山西省高等学校科技创新项目(2015104,2015105)资助。

张 虎(1979-),男,博士,讲师,主要研究方向为社会计算、自然语言处理,E-mail:zhanghu@sxu.edu.cn(通信作者);**吴永科**(1992-),男,硕士生,主要研究方向为社会计算;**杨陟卓**(1983-),男,博士,讲师,主要研究方向为自然语言处理;**刘全明**(1973-),男,博士,高级工程师,主要研究方向为社会计算。

动力学的算法，主要包括 Reichardt 和 Bornholdt[14-15] 将物理学中的 Potts 模型运用到社区发现中的超顺磁聚类算法，以及 Wu 等人[16] 将网络类比成电路的电流算法。3）基于模块度优化的算法，主要包括贪婪算法[17] 和极值优化算法[18-19] 等。4）基于聚类的社区发现算法，这类算法是寻找社会网络中社团结构的一类传统方法，它基于各个节点之间连接的相似性或者强度，将网络自然地划分为各个子群，根据向网络中添加边或者是从网络中移除边，该类算法可以分为凝聚方法和分裂方法[20] 两类。

基于节点相似度的社区发现算法是一种典型的基于聚类的方法，常用的节点相似性度量方法有共同邻居、Jaccard 指标、大度节点有利指标、Salton 指标和 Sorenson 指标等。Qiu 等人[21] 提出了共有邻居相似度的定义，把共有邻居相似度作为衡量两个节点相似度的唯一标准，若两个节点的共有邻居相似度高于预先设定的阈值，则认为这两个节点相似，属于同一个社区。Xu 等人[22] 提出了一种综合考虑用户交互行为和相似度的社区发现方法，该方法将网络中用户之间的多维关联概括为交互行为和相似度，使用加入相似性惩罚因子的相似模块度作为目标函数来指导社区的划分。Wei 等人[23] 提出了一种基于用户紧密度的微博网络社区发现算法，根据微博网络中用户间的交互度与共有邻居的相似度来计算用户的紧密度，并通过与传统的 GN 算法相结合来对微博网络进行社区划分。

目前，现有基于节点相似性的社区划分方法存在以下不足：1）如果多个节点之间的相似性相同，首先应该将哪两个节点合并成小团体；2）构成不同的小团体后应如何进一步合并这些小团体。针对上述不足，本文提出了一种基于多层节点的节点相似度计算方法和团体之间的连接紧密度度量准则，并在此基础上构建了社区发现模型。

2　节点相似度度量准则

2.1　单层节点相似性指标

社会网络分析中经典的三元闭包原则指出，如果两个人 A 和 B 拥有一个共同的朋友 C，那么这两个人今后也很有可能成为朋友，从而使得 3 个节点构成一个闭合的三角形 ABC。对于一般的网络，可以把这一原则进一步推广如下：两个节点的共同邻居的数量越多，这两个节点就越相似，从而更倾向于相互连接。节点相似度旨在度量两个节点 v_i 和 v_j 之间的相似程度，它通过节点的邻居节点之间的交叉度来计算，在社区划分中一般认为相似度大的节点属于同一个社区。假设 $N(v_i)$ 和 $N(v_j)$ 分别是 v_i 和 v_j 的邻居节点，通过图 1 可以看出 v_i 和 v_j 具有公共邻节点的个数。基于共同邻居的节点相似性指标一般可定义为：

$$\sigma_1(v_i,v_j)=|N(v_i)\bigcap N(v_j)| \tag{1}$$

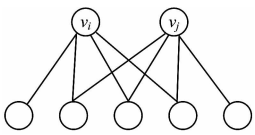

图 1　两个节点的公共邻居节点

Fig. 1　Common adjacent nodes of two nodes

但在实际复杂网络中，多组节点的节点相似度经常相等，如图 2 所示，v_i 与节点 v_j 和 v_n 的节点相似度均为 3，面对该情形，确定 v_i 与哪个节点先合并就成为一个亟需解决的问题。

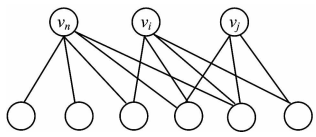

图 2　多个节点的公共邻节点

Fig. 2　Common adjacent nodes of multiple nodes

2.2　多层节点相似性指标

通过式(1)可以计算出两个节点之间的相似度，为了解决多组节点的节点相似度相等时的合并选择问题，考虑进一步使两个节点 v_i 和 v_j 的相似度度量更加精准，据此提出了基于多层节点的节点相似性度量指标。式(1)是通过节点邻居节点之间的交叉程度计算的，那么是否可以考虑节点邻居的邻居节点之间的交叉程度呢？基于这个思路，本文定义了二层节点相似度，主要有两种情况。

(1)节点 v_i 和 v_j 通过两个节点连接。图 3(a)给出了节点 v_i 的邻居节点的邻居节点与节点 v_j 的邻居节点的交叉程度，将 $N_2(v_i)$ 表示为节点 v_i 的邻居节点的邻居节点，这时可以将这种情况下的节点相似度表示为：

$$\sigma_{21}(v_i,v_j)=|N_2(v_i)\bigcap N(v_j)| \tag{2}$$

(2)节点 v_i 和 v_j 通过 3 个节点连接。图 3(b)给出了节点 v_i 的邻居节点的邻居节点和节点 v_j 的邻居节点的邻居节点的交叉程度，将 $N_2(v_j)$ 表示为节点 v_j 的邻居节点的邻居节点，这时可以将这种情况下的节点相似度表示为：

$$\sigma_{22}(v_i,v_j)=|N_2(v_i)\bigcap N_2(v_j)| \tag{3}$$

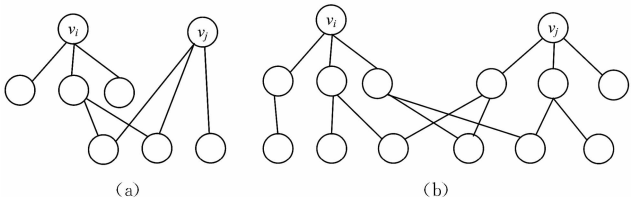

图 3　两个节点的二层公共邻节点

Fig. 3　Two-layer common adjacent nodes of two nodes

将以上两种形式称为二层节点相似度，可表示为式(4)：

$$\sigma_2(v_i,v_j)=\sigma_{21}(v_i,v_j)+\sigma_{22}(v_i,v_j) \tag{4}$$

当计算两个节点之间的相似度时，如何将二层节点相似度合并到节点相似度中？这需要考虑二层节点相似度所占的权重。本文通过网络的平均度来衡量这个权重，可表示为式(5)：

$$W=m/((n*(n-1))/2) \tag{5}$$

其中，W 表示二层节点相似度的权值，m 表示网络的实际边数，n 表示网络的节点数。

通过以上公式可以重新定义节点 v_i 和节点 v_j 之间的节点相似性，如式(6)所示：

$$\sigma(v_i,v_j)=\sigma_1(v_i,v_j)+W*\sigma_2(v_i,v_j) \tag{6}$$

利用式(6)计算两个节点之间的相似度，不仅可以使节点间的相似性更加精确，而且还能解决节点合并选择问题。显然，通过这样的计算仍会存在节点相似度相等的问题，因此实验还尝试计算了三层节点相似度，但结果表明三层节点相似度在小规模的社区网络上的效果不明显，反而会增加大量模型的计算。

为了解决基于多层公共邻节点的节点相似度相等问题，在本文的实验中对式(1)－式(3)给出了另外的计算方法。一般来说，当公共邻节点数相同，且节点本身的邻节点较多时，相似度会较小。基于此，本文定义式(7)来精确化多层节点相似度，以进一步解决多层节点相似度相等的问题。

$$\sigma_m(v_i,v_j)=\frac{N(v_i)\bigcap N(v_j)}{N(v_i)\bigcup N(v_j)} \tag{7}$$

其中，$N(v)$表示节点 v 的邻节点集合；σ_m 的范围为 0～1，σ_m 越接近 1 表示两节点的相似度越高。在本文实验中，当式(6)计算的两个节点相似度相等时，可用式(7)的计算方式替换式(1)－式(3)，并重新计算式(6)的结果。

3 基于节点相似度的社区划分模型

3.1 团体之间的紧密度

当单个节点合并成小的团体后，需要将这些小团体合并成大的团体，直到形成一个个社区。而在合并小团体时，需要考虑各个小团体之间的紧密度，然后使紧密度高的团体之间优先合并，直到最后形成合适的社区。团体之间的紧密度可定义为团体之间的连边数与团体之间可能形成的边数的比值，可用式(8)表示：

$$D(C_i,C_j)=e(C_i,C_j)/(n(C_i)*n(C_j)) \tag{8}$$

其中，$e(C_i,C_j)$表示团体 C_i 和 C_j 之间边的连接数量；$n(C_i)$表示团体 C_i 内部的成员数量；$n(C_j)$表示团体 C_j 内部的成员数量。

在式(8)中，如果两个团体间计算的结果相同，可以通过计算两个团体间有连边节点的平均相似度，选择较大的优先合并。平均相似度可用式(9)表示：

$$aveS(C_i,C_j)=\sum\sigma(v_i,v_j)/e(C_i,C_j) \tag{9}$$

其中，$\sum\sigma(v_i,v_j)$表示小团体 C_i 和 C_j 之间有连边节点的相似度之和；$e(C_i,C_j)$表示小团体 C_i 和 C_j 之间边的连接数量。

3.2 算法描述

算法思想：首先计算任意两个节点之间的相似度，按照设定的阈值合并不同的节点，并形成一个个小团体；然后通过团体间的紧密度不断合并节点或小团体，迭代该过程直到找到合适的社区为止；最后通过评价指标对社区划分结果进行评价。

算法流程：假设 $G(V,E)$是一个由 n 个节点、m 条边组成的复杂网络，其中 V 是网络节点的集合，E 是网络边的集合。

(1)计算各个节点间的相似度，表示为 S，同时根据 S 的大小设定一个阈值 T。

(2)初始化社区，对于每个节点，将每个节点的社区分别表示为 $C\{i\}$。

(3)找到相似度最大且相似度值大于阈值 T 的节点对，将其合并成小团体。若节点 v_i 和节点 v_j 满足条件，则将节点 v_j 合并到节点 v_i 对应的社区中，即 $C\{i\}=C\{i,j\}$，并将 $C\{j\}$清空，对应的相似度 S 置为 0。在此过程中，若节点 v_i 已经合并过则不再进行处理。

(4)重复步骤(3)直到没有找到满足相似度最大且相似度值大于阈值 T 的节点对为止。

(5)计算由步骤(4)得到的各个小团体之间的紧密度 D。找到最大 D 值所对应的团体进行合并，若 D 值相同，则计算团体间节点的平均相似度值，选择较大的优先合并，合并操作与步骤(3)一样。

(6)重复步骤(5)，直到划分的社区个数等于 k 时迭代终止，并计算其相应的社区划分的评价指标。

(7)对于不同类型的复杂网络，分别尝试设定不同的 k 值，通过相应的评价指标来分析最优的社区划分结果。

4 实验测试与分析

4.1 评价指标

为了测试所提方法的性能，本文利用真实网络数据集进行了社区发现实验。所用的数据集有 Karate 跆拳道俱乐部网络、Dolphins 网络、Football 网络和论文作者合作网络，这些都是通过实际统计得到的数据集[25]。针对社区划分的实验结果评价，本文运用了两种典型的方法：模块度函数和归一化互信息。

(1)模块度，它是 Newman 和 Girvan 等人[24]提出的用于衡量网络划分质量的标准。其物理意义是：网络中连接两个同种类型的节点的边的比例，减去在同样的社区结构下任意连接这两个节点的边的比例的期望值。模块度函数通常用式(10)表示：

$$Q=\sum(e_{ij}-a_i^2) \tag{10}$$

其中，e_{ij}表示网络中链接两个不同社区的节点的边在所有边中所占的比例，这两个节点分别位于第 i 个社区和第 j 个社区；每行(或列)中各元素之和 $a_i=\sum e_{ij}$，它表示与第 i 个社区中的节点相连的边在所有边中所占的比例。

(2)归一化互信息[26](Normalized Mutual Information，NMI)，常用来度量原始社区划分和实验算法得到的社区划分的差别。其定义如式(11)所示：

$$NMI(C_0,C_e)=\frac{H(C_0)+H(C_e)-H(C_0,C_e)}{\sqrt{H(C_0)H(C_e)}} \tag{11}$$

其中，C_0 表示原始的社区划分，即真实的社区；C_e 表示算法得到社区划分；$H(C)$表示划分 C 的香农信息熵。当 C_e 与 C_0 完全一致时，$NMI(C_0,C_e)=1$；当 C_e 与 C_0 完全不同时，$NMI(C_0,C_e)=0$。

4.2 基于 Karate 跆拳道俱乐部的社区划分实验

Karate 跆拳道俱乐部网络是对一个美国大学空手道俱乐部进行观测而构建的一个社会网络。该网络包含 34 个节点和 78 条边，其中个体表示俱乐部中的成员，而边表示成员之间存在的友谊关系。

实验 1 采用单层节点相似度，通过不断调整阈值的大小对 Karate 网络进行了划分，当阈值 T 设定为 5，k 设定为 2 时，该方法得到了最优的结果，结果如图 4 所示，与真实社区相比，实验中编号为 10 的节点被划分错误。

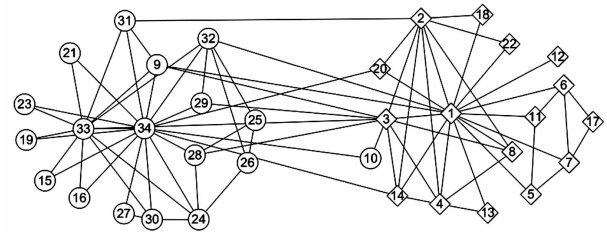

图 4 Karate 跆拳道俱乐部基于单层节点社区的划分结果

Fig. 4 Detection result based on the single-layer nodes community on Karate data

实验 2 使用多层节点相似度方法进行社区划分，发现基于两层相似度时仍会有一些相同的节点相似度，因此实验进一步使用了三层节点相似度。在阈值 T 为 5，k 为 2 时，该网络的社区划分结果最好，结果如图 5 所示。不同的社区分别用不同的形状来表示，由图 5 可以看出，该社区划分结果与真实网络划分结果完全一致，即 $NMI=1$。

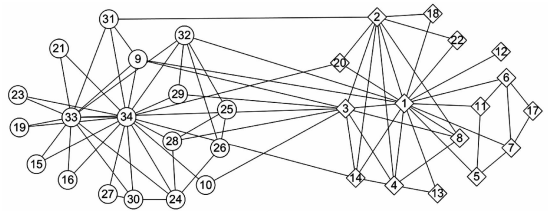

图 5　Karate 跆拳道俱乐部基于多层节点社区的划分结果

Fig. 5　Detection result based on the multi-layer nodes community on Karate data

同时由图 5 可知，划分的社区分别是以节点 34 和节点 1 为代表的小团体，这与真实网络中由于俱乐部的主管和校长意见不合，最后导致俱乐部分裂成两个以他们为核心的不同俱乐部是一致的。

4.3　基于 Dolphins 网络的社区划分实验

Dolphins 数据集是 Lusseau 等人使用长达 7 年的时间观察新西兰 Doubtful Sound 海峡 62 只海豚群体的交流情况而得到的海豚社会关系网络。该网络具有 62 个节点，159 条边，节点表示海豚，边表示海豚间的频繁接触。

实验分别基于单层节点和多层节点相似度对 Dolphins 网络数据集进行划分。不断调整阈值 T 和 k 的大小，当阈值 T 设定为 2，k 也设定为 2 时，两种方法均得到了最优的结果。该算法基于单层节点和多层节点相似度的社区的划分结果分别如图 6 和图 7 所示，不同的社区分别用不同的形状表示。

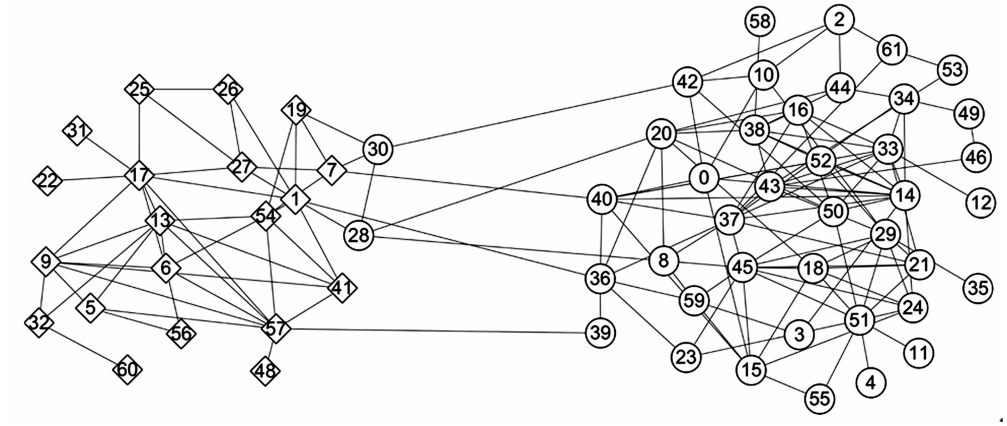

图 6　Dolphins 网络基于单层节点社区的划分结果

Fig. 6　Detection result based on the single-layer nodes community on Dolphins data

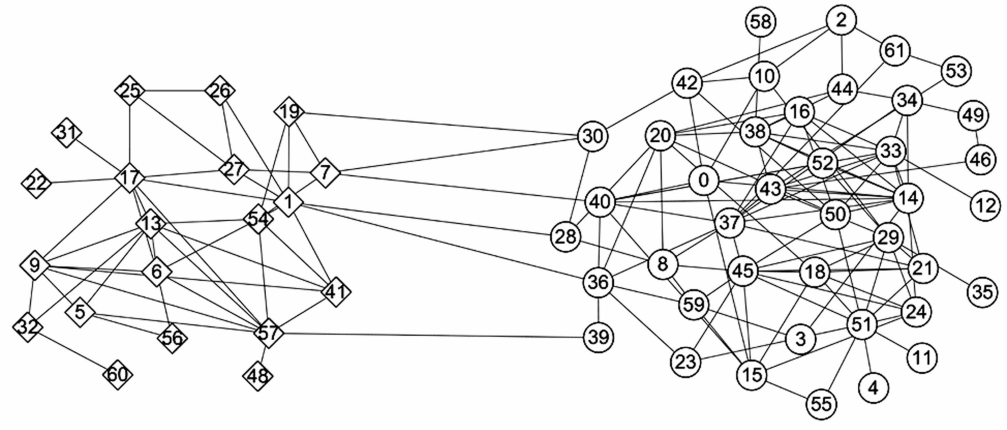

图 7　Dolphins 网络基于多层节点社区的划分结果

Fig. 7　Detection result based on the multi-layer nodes community on Dolphins data

图 6 中菱形形状节点的节点簇中含有圆形形状节点（节点编号分别为：28，30）属于划分错误的节点，这表明基于单层节点的社区划分结果与真实网络相比有不一致的地方。通过图 7 可以看出，该社区无划分错误节点，这说明基于多层节点的社区划分方法的结果与真实网络划分的结果完全一致，即 $NMI=1$。

4.4　基于 Football 网络的社区划分实验

Football 网络是 Newman 根据美国大学生足球联赛创建的一个复杂的社会网络，该网络包含 115 个节点和 613 条边，其中网络中的结点代表足球队，两个结点之间的边表示两只球队之间进行过一场比赛。

实验分别基于单层节点和多层节点相似度对 Football 网络数据集进行划分。不断调整阈值 T 和 k 的大小，当阈值 T 设定为 5，k 设定为 11 时，两种方法均得到了最优的结果。该算法基于单层节点和多层节点相似度的社区的划分结果分别如图 8 和图 9 所示。

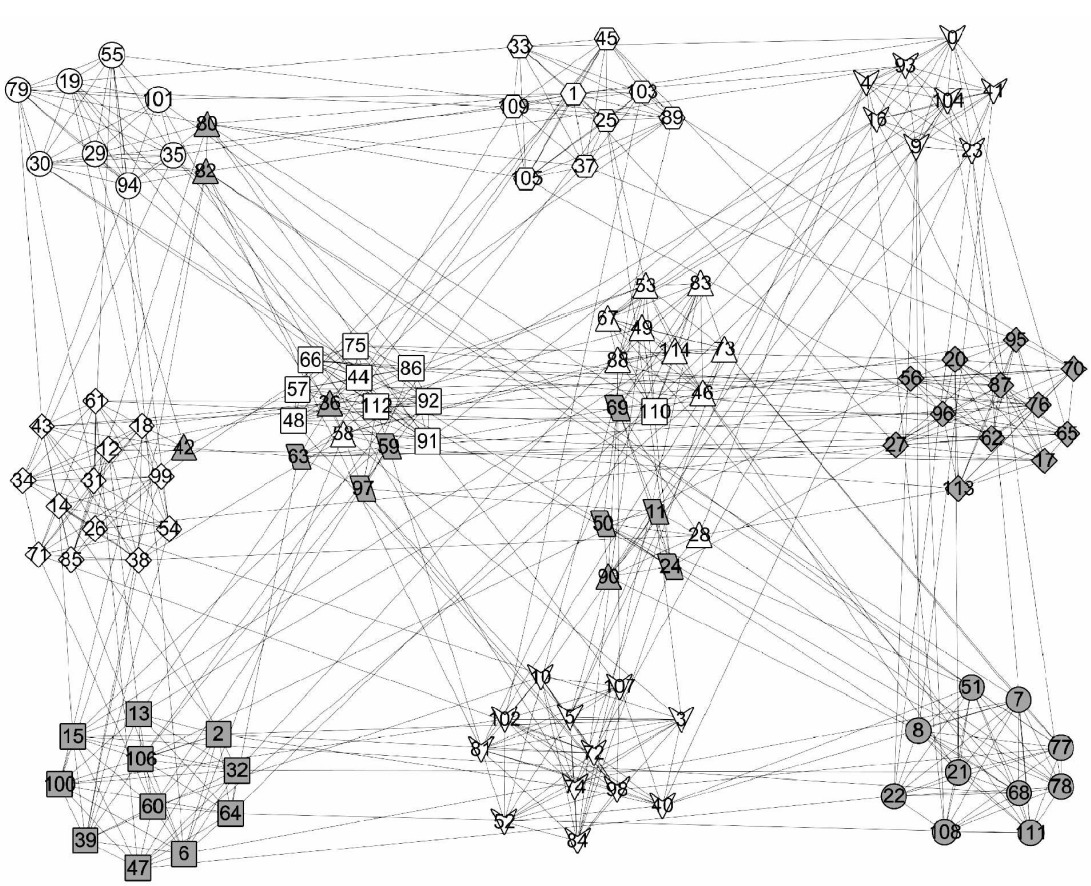

图 8 Football 网络基于单层节点社区的划分结果

Fig. 8 Detection result based on the single-layer nodes community on Fottball data

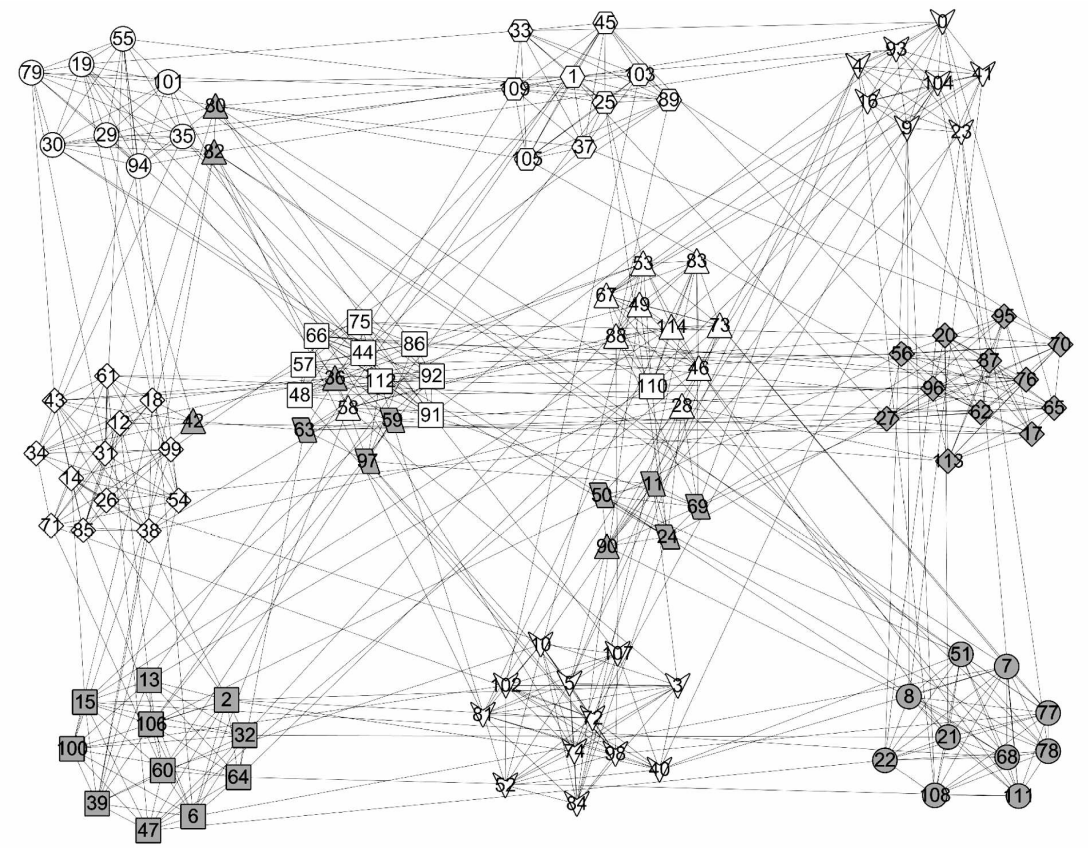

图 9 Football 网络基于多层节点社区的划分结果

Fig. 9 Detection result based on the multi-layer nodes community on Football data

图 8 表明，基于单层节点相似度的社区划分结果与真实网络相比有 12 个划分错误的节点（节点编号分别为：28，36，42，58，59，63，69，80，82，90，97，110）。图 9 显示出了基于多层节点相似度的社区划分结果：在相同形状及阴影形状表示的簇中共有 10 个不同类型的节点，其中 5 个阴影三角形的节点（节点编号分别为：36，42，80，82，90）表示通过该算法未被划分到社区的节点，而剩余的 5 个节点（节点编号分别为：58，59，63，97，110）表示社区划分错误的节点。实验的归一化互信息计算为 $NMI=0.912$。

4.5　实验结果比较

为了验证所提方法的有效性，本文进行了以下两组对比实验。

（1）基于单层节点和多层节点相似度的社区划分实验，实验细节如图 4－图 9 所示，实验结果如表 1 所列。

表 1　基于单层节点和多层节点相似度的社区划分实验结果

Table 1　Community detection results based on single-layer and multi-layer nodes similarity

网络	参数		划分错误节点数（基于单层节点）	划分错误节点数（基于多层节点）
	T	k		
Karate 网络	5	2	1	0
Dolphins 网络	2	2	2	0
Football 网络	5	11	12	10

针对 3 个真实网络的社区划分结果显示，本文提出的基于多层节点相似度的社区划分方法比基于单层节点相似度的社区划分方法有更少的划分错误节点。该组实验结果表明，本文提出的方法在划分结果上优于基于单层节点相似度的社区划分方法。

（2）基于多层节点相似度的社区划分方法与其他典型的社区划分方法的比较实验。

该组实验将本文提出的方法与其他常见的典型社区发现方法进行了比较。参与对比的算法有 GN 算法[13]、Fast Newman 算法[24] 和改进的标签传播算法[28]（LPALS）。GN 算法采用逐次移除介数中心性最大边的方法得出社区划分，是分裂型算法的典型代表；Fast Newman 算法是一种基于模块度优化的算法[27]，是在使得模块度不断增加的基础上进行的，即每次合并沿着使模块度增多到最大或减少到最小的方向进行，是一种基于贪婪法思想[11] 的凝聚算法；LPALS 是一种改进的标签传播算法，有效地继承了 LPA 算法的精度，同时提高了 LPA 算法的精度。评价指标采用了归一化互信息。4 种算法在不同网络上的实验结果如表 2 所列。

表 2　4 种方法的实验结果

Table 2　Experimental results of four methods

网络	GN 算法	Fast Newman	LPALS	本文算法（最大模块度）	本文算法（NMI 最优）
Karate 网络	0.836	0.693	0.870	0.826	1.000
Dolphins 网络	0.781	0.573	—	0.490	1.000
Football 网络	0.868	0.762	0.860	0.886	0.912

表 2 对本文的社区发现方法列出了两种评价结果：1）社区划分后模块度最大的社区划分结果；2）社区划分后 NMI 值最大的社区划分结果。

模块度最大的社区划分结果比最优结果的 NMI 值小的原因在于两者对应的社区划分数不同。模块度最大的社区划分结果将 Karate 跆拳道俱乐部网络划分为 3 个社区，将 Dolphins 网络划分为 13 个群体，将 Football 网络划分为 10 个社区。而 NMI 值最大的社区划分结果在 3 个网络上的划分社区数分别为 2，2 和 11。显然，通过表 2 的两组数据可以看出，本文提出的方法在社区划分时若能够预先确定与真实社区划分相近的社区数 k，则会表现出较好的划分结果；反之若通过模块度来选择最优的社区划分结果，则可能得到相对较差的社区划分。

该实验结果表明，与 GN 算法[13]、Fast Newman 算法[24] 和改进的标签传播算法[28] 相比，本文提出的方法对已知社区数的小规模网络的社区划分准确率优于这 3 种算法。

4.6　基于论文作者合作网络数据集的社区划分结果分析

本文方法在两种合作网络数据集上分别进行了实验。

其一是 Chris McCarty 于 2008 年在 INSNA 会议上准备的一个数据集，它记录了社会网络杂志的合著关系，共包括 475 名作者参与的 295 篇文章的合著关系。实验利用模块度来评价本文模型的社区划分结果，表 3 给出了模块度最大为 0.92 时的实验结果，共将 475 名作者分成 110 个子社区。

表 3 结果显示了合作网络中存在很多零散的小社区，节点数在 5 以下的子社区数量占所有子社区数量的 81%；同时，实验显示 350 号节点代表和 162 号节点代表的作者影响力较大，合著范围广。

表 3　基于 McCarty 科学合作网络数据集的社区划分结果

Table 3　Community detection result on McCarty collaboration network

社区中的节点数目	社区数目
33	1
27	1
⋮	⋮
3	30
2	51
总计	110

其二是 Mark Newman 于 2006 年 5 月编写的 netScience 数据集，共有 1589 名科学家和 2742 条边。本文提出的方法在该数据集上实验，当子社区合并到 396 个时停止合并，此时各个子社区之间没有连边，模块度达到 0.81，最大的子社区包含的节点数为 379，表 4 显示了该实验的社区划分结果。

表 4　基于 netScience 数据集的社区划分结果

Table 4　Community detection result on netScience data

社区中的节点数目	社区数目
379	1
57，31，28，21，13，12，11	1
14，10	2
9	4
8	9
7	7
6	10
5	15
4	38
…	…
总计	396

表 4 的结果显示了本文方法能有效地划分出社区中由 379 位科学家组成的最大子社区。同时，针对该 netScience 合

作网络的社区划分结果与表 3 的结果相似，也存在很多零散的小社区，节点数在 5 以下的子社区数量占所有子社区数量的85.6%。该实验表明，本文提出的方法在中等规模的论文作者合作网络数据集上会取得模块度较大的社区划分结果。

针对不同类型数据集上的实验结果表明，相较于基于单层节点相似度的社区划分方法，本文提出的基于多层节点相似度的社区划分方法有更高的划分准则率。同时，在划分前确定了与真实社区划分相近的社区数 k，相比于其他典型的社区发现方法，该方法会表现出更好的划分效果。

结束语 本文针对现有节点相似度计算方法中存在的不足，提出了一种基于多层节点的节点相似度计算方法，并进一步基于这种改进的节点相似度计算方法和团体之间的连接紧密度度量准则构建了社区发现模型。分别在 4 种真实网络上进行了实验，验证了本文所提方法相比基于单层节点相似度的社区发现方法和其他典型社区方法存在的优点。

参 考 文 献

[1] PORTER M A, ONNELA J P, MUCHA P J. Communities in networks[J]. Notices American Mathematical Society, 2009, 56(9): 1082-1097.

[2] HUANG F L. Studies on community detection and its application in information network[J]. Complex Systems and Complexity Science, 2010, 7(1): 64-74.

[3] LUCE R D, PERRY A D. A method of matrix analysis of group structure[J]. Psychometrika, 1949, 14(2): 95-116.

[4] ALBA R D. A graph-theoretic definition of a sociometric clique [J]. Journal of Mathematical Sociology, 1973, 3(1): 113-126.

[5] LUCE R D. Connectivity and generalized cliques in sociometric group structure[J]. Psychometrika, 1950, 15(2): 169-190.

[6] MOKKEN R J. Cliques, clubs and clans[J]. Quality and Quantity, 1979, 13(2): 161-173.

[7] SEIDMAN S B, FOSTER B L. A graph-theoretic generalization of the clique concept[J]. Journal of Mathematical Sociology, 1978, 6(1): 139-154.

[8] SEIDMAN S B. Network structure and minimum degree[J]. Social networks, 1983, 5(3): 269-287.

[9] LUCCIO F, SAMI M. On the decomposition of networks into minimally interconnected networks[J]. IEEE Transactions Circuit Theory, 1969, 16(2): 184-188.

[10] 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.

[11] GUIMERA R, SALES-PARDO M, AMARAL L A N. Modularity from fluctuations in random graphs and complex networks [J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2004, 70(2): 025101.

[12] POTHEN A, SIMON H D, LION K P. Partitioning sparse matrices with eigenvectors of graphs[J]. SIAM Journal on Matrix Analysis and Applications, 1990, 11(3): 430-452.

[13] GIRVAN M, NEWMAN M E J. Community structure in social and biological networks[J]. Proceedings of the national academy of sciences, 2002, 99(12): 7821-7826.

[14] REICHARDT J, BORNHOLDT S. Detecting fuzzy community structures in complex networks with a Potts model[J]. Physical Review Letters, 2004, 93(21): 218701.

[15] REICHARDT J, BORNHOLDT S. Statistical mechanics of community detection[J]. Physical Review E, 2006, 74(1): 016110.

[16] WU F, HUBERMAN B A. Finding communities in linear time: a physics approach[J]. The European Physical Journal B-Condensed Matter and Complex Systems, 2004, 38(2): 331-338.

[17] NEWMAN M E J. Fast algorithm for detecting community structure in networks[J]. Physical Review E, 2004, 69(6): 066133.

[18] BOETTCHER S, PERCUS A G. Optimization with extremal dynamics[J]. complexity, 2002, 8(2): 57-62.

[19] ZHOU T, BAI W J, CHENG L J, et al. Continuous extremal optimization for Lennard-Jones clusters[J]. Physical Review E, 2005, 72(1): 016702.

[20] SCOTT J. Social Network Analysis: A Handbook(2nd ed)[M]. London: Sage Publications, 2002.

[21] QIU L Q, CHEN Z Y. Community discovery algorithm based on common neighbor similarity[J]. Information System Engineering, 2014(5): 140-141. (in Chinese)
仇丽青，陈卓艳. 基于共同邻居相似度的社区发现算法[J]. 信息系统工程，2014(5)：140-141.

[22] XU W, LIN B G, LIN S J, et al. Research on the discovery method of social network based on user interaction behavior and similarity[J]. Information network security, 2015(7): 77-83. (in Chinese)
许为，林柏钢，林思娟，等. 一种基于用户交互行为和相似度的社交网络社区发现方法研究[J]. 信息网络安全，2015(7)：77-83.

[23] WEI Q J, LI J T, WANG Y. Micro-blog network community discovery algorithm based on user density[J]. Computer Applications and Software, 2016, 33(9): 254-258. (in Chinese)
韦庆杰，李京腾，汪雨. 基于用户紧密度的微博网络社区发现算法[J]. 计算机应用与软件，2016，33(9)：254-258.

[24] CLAUSET A, NEWMAN M E J, MOORE C. Finding community structure in very large networks[J]. Physical review E, 2004, 70(6): 066111.

[25] FANG P, GUO Z B, LI Z T, et al. Online social network community structure detection algorithm based on number of shared friends[J]. Journal of Frontiers of Computer Science & Technology, 2012, 6(5): 456-464.

[26] DANON L, DIAZ-GUILERA A, DUCH J, et al. Comparing community structure identification[J]. Journal of Statistical Mechanics: Theory and Experiment, 2005, 2005(9): 09008.

[27] NEWMAN M E J, GIRVAN M. Finding and evaluating community structure in networks[J]. Physical Review E, 2004, 69(2): 026113.

[28] KANG X B, JIA C Y. An improved fast community detection algorithm based on label propagation[J]. Journal of Hefei University of Technology, 2013, 36(1): 43-47. (in Chinese)
康旭彬，贾彩燕. 一种改进的标签传播快速社区发现方法[J]. 合肥工业大学学报，2013，36(1)：43-47.