计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210200009

基于节点相似性和网络嵌入的复杂网络社区发现算法

杨旭华　王　磊　叶　蕾　张　端　周艳波　龙海霞
浙江工业大学计算机科学与技术学院　杭州 310023
(xhyang@zjut.edu.cn)

摘　要　社区发现算法对分析复杂网络的拓扑和层次结构、预测复杂网络的演化趋势等具有十分重要的意义。传统的社区发现算法划分精度不高,忽略了网络嵌入的重要性。针对这样的问题,提出了基于节点相似性和网络嵌入 Node2Vec 方法的无参数社区发现算法。首先,使用网络嵌入 Node2Vec 方法将网络节点映射成欧氏空间中低维向量表示的数据点,计算低维向量表示的数据点之间的余弦相似性,根据相应节点间的最大相似性构建偏好网络,得到初始社区划分,把每个初始社区的最大度节点作为备选节点;然后根据网络平均度和平均最短路径找出备选节点中的中心节点;最后将中心节点对应的数据点及其数量作为初始质心和聚类数,用 K-Means 算法对低维向量表示的数据点进行聚类,从而对相应的网络节点完成社区划分。该算法为无参数社区划分方法,可以自主地从网络中提取参数,无须根据网络的不同设定不同的超参数,从而可以自动地快速识别复杂网络的社区结构。在 8 个真实网络和人工网络上,将其与其他 5 个知名社区发现算法相比较,数值仿真实验表明所提算法具有很好的社区发现效果。

关键词:无参数社区发现;节点相似性;偏好网络;网络嵌入;K-Means 聚类
中图法分类号　TP391

Complex Network Community Detection Algorithm Based on Node Similarity and Network Embedding

YANG Xu-hua, WANG Lei, YE Lei, ZHANG Duan, ZHOU Yan-bo and LONG Hai-xia
College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China

Abstract　The community detection algorithm is very important for analyzing the topology and hierarchical structure of complex networks and predicting the evolution trend of complex networks. Traditional community detection algorithm does not have high accuracy and ignores the importance of network embedding. Aiming at such problems, a parameter-free community detection algorithm based on node similarity and network embedding Node2Vec method is proposed. First, we use the network embedding Node2Vec method to map network nodes into data points represented by low-dimensional vectors in Euclidean space, calculate the cosine similarity between the data points represented by the low-dimensional vector, construct a preference network according to the maximum similarity between the corresponding nodes, obtain the initial community detection, and use the maximum degree node of each initial community as a candidate node. Then we find the central node among the candidate nodes according to the average degree of the network and the average shortest path. Finally, the data points and their numbers corresponding to the central node are used as the initial centroid and cluster number, and the data represented by the low-dimensional vector are calculated by K-Means algorithm. The points are clustered, and the corresponding network nodes are divided into communities. This algorithm is a method of community division without parameters, which can independently extract parameters from the network without setting different hyper-parameters according to different networks, so that it can automatically and quickly identify the community structure of complex networks. In 8 real networks and artificial networks above, by comparing with other 5 well-known community discovery algorithms, numerical simulation experiments show that the proposed algorithm has good community discovery effect.

Keywords　Parameter-free community detection, Node similarity, Preference network, Network embedding, K-Means clustering

到稿日期:2021-02-01　返修日期:2021-05-11　　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61773348)
This work was supported by the National Natural Science Foundation of China(61773348).
通信作者:龙海霞(longhaixia@zjut.edu.cn)

1 引言

社区结构是复杂网络的重要结构特征之一，很多现实网络具有明显的社区结构，也就是说一个网络可以看作由若干个社区组成，社区内部节点间的连接比较稠密，而不同社区之间的节点连接比较稀疏。例如，社交网络[1]中的社区往往代表着一群拥有共同爱好或者共同行为习惯的用户，在该网络中进行社区发现可以分析人际关系等；在生物网络中[2]中，同一社区中的节点可能表示功能相近的基因组织；在 world wide web[3]中，社区代表了一系列具有相同或者相似内容主题的网页的集合，这些网页之间存在大量的超链接。

社区发现是网络分析中的基本任务之一，例如，在社交网络[1]中，通过挖掘潜在的用户社区结构，可以捕捉具有相似兴趣的用户小组，以便为他们推荐更精准的内容；在生物网络[2]中，通过某些生物组织的社区结构和演化过程，可以更好地为医疗发展、基因治疗等提供依据；在科学家合著网络[4]中，随着时间的变化观察其社区结构的演化，不仅可以发现新的研究热点，还可以预测学科的发展趋势。因此，研究高效、准确的社区发现算法具有重要意义。近年来，社区发现引起了不同领域许多研究者的关注，涌现出了多种社区发现方法。GN[4]算法根据网络的边介数中心性(Edge Betweenness Centrality)值，通过迭代移除网络的边来发现存在于网络中的社区，其中，边介数中心性最大的边最先被移除。Louvain 算法[5]是一个基于模块度最优化的算法，通过计算节点合并之后的模块度增益大小来选择合并的节点，并且将合并之后的社区作为新的节点重复合并，直至模块度增益不再变化。标签传播算法[6](LPA)对每个节点在初始状态时打上唯一的标签，标签传播过程中，节点选择其邻居节点中数量最多的标签作为新的标签，具有相同标签的节点形成社区，它是一种近似线性时间的方法，但结果并不总是稳定的。Walktrap[7]是一种基于随机游走的算法，该方法针对于无向网络图，采用图随机游走的方式，获得网络中的社区。Infomap[8]算法是一种基于信息论中的方法，将社区发现和信息编码联系到一起，通过量化编码长度，找到使得长度最短的社区划分。

基于节点相似性的社区发现算法一般采用节点相似性指标计算节点之间的相似性，将相似性高的节点划分到同一个社区中，相似性低的节点则划分到不同的社区。这类算法依赖节点相似性指标，不同的节点相似性指标划分的社区结果可能存在不同。常见的节点相似性指标有 Salton，Jaccard，AA，RA 等[9]。Tasgin 等[10]使用共同邻居 CN 指标计算节点对之间的相似性值，将相似性最大的节点划分到同一个社区中。Liu 等[11]使用 AA 指标计算网络中互为邻居的节点对之间的相似性，进行初步的社区划分。这些算法普遍存在仅考虑网络的局部结构信息的问题，对网络的全局信息考虑不足，划分的结果中往往存在一些规模过小的社区。

网络嵌入算法可以将网络数据映射为低维向量，能够很好地解决网络数据难以高效输入机器学习算法的问题，目前已成为研究热点。在复杂网络中研究者们提出了许多融合网络嵌入算法的改进算法。Zhao 等[12]提出了一种结合粗糙粒化的网络嵌入社区发现方法，通过网络嵌入获得融合拓扑信息和属性信息的节点向量表示，并将相似的节点映射到距离相近的低维连续的向量空间。Ye 等[13]提出了一种基于邻居节点和关系模型优化的网络表示学习算法。首先，该算法采用节点的邻居节点优化网络表示模型，使上下文窗口中节点的位置信息被嵌入到网络表示中；然后，引入知识表示学习中的关系模型建模节点之间的结构特征，使得节点之间的文本内容以关系约束的形式嵌入到向量中。Zhao 等[14]提出了将社区嵌入和改进的节点嵌入相结合的方法，从而获得融合结构信息和属性信息的节点表示，节点嵌入将节点表示为低维向量，类似地，社区嵌入把社区表示为低维空间中的高斯分布，从而获得更为准确的社区发现结果。Xie 等[15]提出了一种基于深度稀疏滤波的网络社区发现方法，以寻找网络的有效表示。Hu 等[16]提出了一种结合 Node2Vec 算法和谱聚类算法的社区发现算法，用于在复杂网络中学习社区划分，通过 Node2Vec 算法得到网络中节点的嵌入向量，计算相似性矩阵，将相似性矩阵作为谱聚类算法的相似性矩阵对数据点进行聚类操作，得到社区划分结果。

现有社区发现算法大多属于有参数算法，对于不同类型的网络需要设定不同的超参数，这些参数往往没有确定性的获取方法，只能通过大量的迭代实验来获取，对算法的性能影响巨大。本文结合节点相似性和网络嵌入 Node2Vec 方法提出了一种无参数复杂网络社区发现算法，该算法可以自主地从网络中提取参数，不用根据网络的结构不同来预设不同的超参数，能够实现自动和快速的社区划分。理论分析和数值仿真表明，该算法具有良好的社区划分效果和较低的时间复杂度。

本文第 2 节介绍了相关工作；第 3 节详细描述了本文算法；第 4 节给出了数值仿真和结果分析；最后总结全文。

2 相关工作

目前已有大量复杂网络社区发现算法，本文提出的算法与已有的算法存在明显的不同。我们利用网络嵌入方法得到网络节点的低维向量表示，根据余弦相似性构建偏好网络得到初始社区划分，根据网络的平均度和平均最短路径确定中心节点，最后使用 K-Means 算法进行聚类操作得到社区划分。下面介绍相关的研究基础。

2.1 网络嵌入

网络嵌入又称图嵌入，是网络表征学习的一种方法，用低维、稠密的向量空间表示高维、稀疏的向量空间，所学习到的特征可以用于分类、回归、聚类等机器学习任务。Word2vec 算法[17]采用 skip-gram 模型，将单词转换为低维嵌入向量，相似的单词应该有相似的向量表示。DeepWalk 算法[18]通过将随机游走和 skip-gram 模型相结合来学习网络节点的表示向量。Node2Vec 算法[14]通过最大化随机游走采样序列中的节点出现概率，来保持节点之间的高阶邻近性，生成比 DeepWalk[18]质量更高和信息量更大的网络嵌入向量。SDNE 算法[19]设计了一个由多层非线形函数组成的深度模型，通过同时联合优化一阶相似度和二阶相似度来学习高度非线性的网络结构，学习得到的向量表示能够保留局部和全局结构，并且对稀疏网络具有鲁棒性。图卷积神经网络算法[20](GCN)是

一种基于卷积定理的图卷积神经网络，是一种半监督的节点分类算法，通过傅里叶变换使得在非规则的图数据上可以使用卷积操作提取图上的特征对节点进行分类。

2.2 节点相似性

节点相似性可以衡量两个节点之间的接近程度，同一社区内部的节点之间的相似性较高，而不同社区的节点之间的相似性较低。许多节点相似性指标被应用到社区发现算法中，这些指标大致分为以下几类：基于共同邻居的节点相似性指标、基于路径的相似性指标、基于随机游走的相似性指标。共同邻居指标指两个节点的公共邻居的个数，数量越多，说明两个节点的联系越紧密，相似性越高。Jaccard[21]相似性指标考虑到仅仅以公共邻居作为相似性度量方法不够完善，提出两个节点共同邻居数量在所有邻居节点中的占比越大，则二者具有更大的相似性。AA[22]和 RA[21]指标不仅考虑了节点之间的共同邻居个数，还考虑了每一个共同邻居的度数，提出共同邻居的个数越多，则公共节点度越小，节点之间的相似性越高。HP 度量指标[23]为两个节点的共同邻居个数和两个节点度的最小值之比。LHN 相似性算法[24]认为，如果一个节点的邻居与另一个节点具有相似性，则这两个节点也相似，即节点的相似性是可以传递的。余弦相似性指标[21]计算向量之间的相似性，可以用来衡量网络嵌入之后低维向量之间的相似性。

3 基于节点相似性和网络嵌入方法的复杂网络社区发现

基于节点相似性和网络嵌入方法的复杂网络社区发现(Similarity Node2Vec Community Detection，SNV)算法，首先使用网络嵌入 Node2Vec 算法获取网络中节点的低维向量表示，根据向量计算相应节点之间的余弦相似度，构建偏好网络，得到初始社区划分，通过网络的平均度和最短路径长度获取网络中的中心节点；然后将中心节点及其数量作为质心和聚类数目，使用 K-Means 算法对节点的向量表征进行聚类操作，从而获取相应网络节点的社区划分结果。

3.1 问题描述

针对一个无权无向的复杂网络 $G(V,E)$，节点集合 V 表示网络中的节点，连边集合 E 表示节点之间的连边关系，N 表示网络中节点总数。$\boldsymbol{A}=(A_{i,j})_{N\times N}$ 表示网络的邻接矩阵，其中 $A_{i,j}=1$ 表示节点 i 和节点 j 之间存在一条连边，反之表示两者之间没有边。

本文的目标是通过相似性构建偏好网络，从原始网络中提取中心节点，将中心节点及其数量作为质心和聚类数目，用 K-Means 算法完成社区划分。该算法可以自主地提取参数，无须针对不同的网络提前人为设置任何参数，是一种无参数的社区发现算法。

3.2 SNV 算法框架

SNV 算法的整体框架如图 1 所示。我们首先使用 Node2Vec 方法得到网络中每一个节点的低维向量表示，计算网络中邻居节点相应的低维向量之间的余弦相似性。对于每一个节点，保留与其相似性最大的节点之间的连边，删除其他连边，构建偏好网络。偏好网络分裂成若干个彼此不连通的子网，每个子网对应一个初始社区。选取初始社区中最大的度节点作为备选节点。根据网络的平均节点度和平均最短路径从备选节点中筛选出中心节点。最后，利用 K-Means 算法对低维向量表示的数据点集合进行聚类，完成相应网络节点的社区划分。

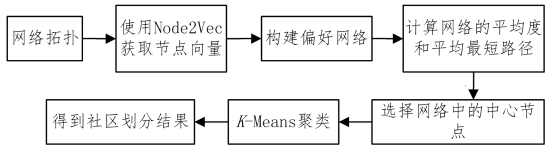


图 1 SNV 算法框架

Fig. 1 SNV algorithm framework

3.3 偏好网络

在社区发现中，有连边的邻居节点之间的影响力比不相连的节点对大，即一个节点与它的邻居节点之间的相似性应该大于它与非邻居节点之间的相似性。本文使用余弦相似性[21]计算每个节点与其邻居节点之间的相似性，而不是任意两个节点之间的相似性。余弦相似性指标如式(1)所示，其中 $\boldsymbol{M}_i$ 表示节点 i 的向量表示。

$$Sim(i,j)=\frac{\boldsymbol{M}_i\cdot\boldsymbol{M}_j}{\|\boldsymbol{M}_i\|\cdot\|\boldsymbol{M}_j\|} \tag{1}$$

如果仅连接网络中每一个节点和它相似性最高的节点，我们可以得到偏好网络 $G^p(V,E^p)$，其中节点集为 V，连边集为 $E^p=\{e_{i,j}\mid i,j\in V\}$，$e_{i,j}=1$ 表示节点 i 的最大相似性节点为 j。以 Karate 网络为例，每个节点仅连接最相似的节点，删掉其余的所有连边，得到 Karate 网络的偏好网络如图 2 所示。

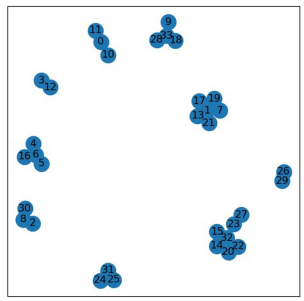


图 2 Karate 网络的偏好网络图

Fig. 2 Preference network of Karate

在偏好网络中，网络不再有整体的连通性，我们把网络中的连通子图作为网络的初始社区划分。比如图 2 的偏好网络存在 9 个连通子图，即初始社区个数为 9。

3.4 中心节点选择策略

根据偏好网络可以得到网络初始社区，但这类社区不能作为最终的社区划分结果，我们需要根据选择策略从初始社区集合中筛选出中心节点，将中心节点的表征向量作为 K-Means 算法的质心，将中心节点的个数作为 K-Means 算法的聚类数对数据进行聚类操作。

为了选择网络中的中心节点，我们提出了选择策略，通过计算网络平均度和最短路径长度选择中心节点。首先，根据偏好网络的连通性获得初始社区集合 $C=\{c_1,c_2,c_3\cdots c_h\}$，每个初始社区的最大度节点作为备选节点集 V_{p0}。计算每个备选节点在原始网络中的度值并降序排列，得到节点-度的二元组集合 $X=\{(i,degree(i))\mid i\in V_{p0}\}$，其中 $degree(i)$ 表示

节点 i 的度；计算网络 G 的平均节点度 AD，从节点-度集合 X 中去除备选节点集中度小于 AD 的节点，更新备选节点集合 V_{p0} 为 V_{p1}。

其次，计算备选节点之间在原始网络中的最短路径长度，得到节点-节点-最短路径长度的三元组集合 $Y=\{(i,j,distance(i,j))|i,j\in V_{p1}\}$，其中 $distance(i,j)$ 表示节点 i 和节点 j 之间的最短路径长度。计算网络 G 的平均最短路径 AWN，遍历备选节点集 V_{p1}，计算每一个备选节点 i 和其他备选节点之间的最短路径的平均值 ASP_i。遍历备选节点集，根据式(2)计算 V_{p1} 中每一个备选节点 i 的最短路径的阈值 $ASP_{i\text{-threshold}}$，去除最短路径 $distance(i,j)$ 小于阈值 $ASP_{i\text{-threshold}}$ 的节点 j，更新备选节点集 V_{p1} 为中心节点集 V_{p2}。

$$ASP_{i\text{-threshold}}=\min(AWN,ASP_i) \tag{2}$$

将 V_{p2} 中的中心节点对应的低维向量表示的数据点及其数量作为初始质心和聚类数目，使用 K-Means 算法进行聚类，得到向量表示的数据点聚类结果，从而对相应的网络节点完成社区划分。

以 Karate 网络为例，偏好网络得到的初始社区的基本信息如表 1 所列，得到节点-度的二元组集合 $X=\{(33,17),(0,16),(32,12),(1,9),(3,6),(31,6),(8,5),(6,4),(26,2)\}$，根据节点的度的大小得到备选节点集 $V_{p0}=\{33,0,32,1,3,31,8,6,26\}$；然后，计算可得 Karate 网络的平均度 $AD=4$，去除备选节点集中度小于网络平均度的备选节点，根据表 1 节点的度去除节点 6 和节点 26，更新备选节点集为 $V_{p1}=\{33,0,32,1,3,31,8\}$；计算备选节点之间的最短路径长度，得到节点-节点-最短路径的三元组 $Y=\{(33,0,2),(33,32,1),(33,1,2),(33,3,2),(33,31,1),(33,8,1),(33,6,3),(0,32,2),(0,1,1),(0,3,1),(0,31,1),(0,8,1),(32,1,2),(32,3,2),(32,31,1),(32,8,1),(1,3,1),(1,31,2)(1,8,2),(3,31,2),(3,8,2),(31,8,2)\}$，Karate 网络的平均最短路径 $AWN=2$，遍历备选节点集 V_{p1}，根据式(2)计算每个备选节点的最短路径的阈值 $ASP_{i\text{-threshold}}$，根据阈值去除最短路径小于平均阈值的备选节点。例如计算节点 33 的最短路径阈值为 1.625，节点 33 和节点 32 的最短路径为 1，小于 1.625，去除节点 32；节点 33、节点 31 以及节点 8 的最短路径均为 1，小于 1.625，去除节点 31 和 8。迭代计算备选节点集中节点的最短路径阈值，去除最短路径低于阈值的其他备选节点，直至不存在最短路径低于阈值的节点，更新备选节点集 V_{p1} 为中心节点集 V_{p2}。

表 1 偏好网络初始社区信息和备选节点信息

Table 1 Preference network initial community information and candidate node information

Community	Member	Candidate node	Degree of candidate node
$c0$	{9,18,28,33}	33	17
$c1$	{0,10,11}	0	16
$c2$	{32,14,15,20,22,23,27}	32	12
$c3$	{1,7,13,17,19,21}	1	9
$c4$	{3,12}	3	6
$c5$	{24,25,31}	31	6
$c6$	{2,8,30}	8	5
$c7$	{4,5,6,16}	6	4
$c8$	{26,29}	26	2

最后，将 V_{p2} 中的中心节点对应的低维向量及其数量作为初始质心和聚类数目，使用 K-Means 算法对向量表示的数据点进行聚类，从而对相应的网络节点完成社区划分，得到的社区划分结果如图 3 所示。

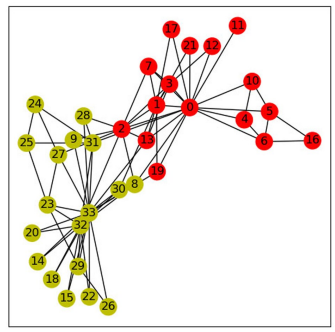


图 3 Karate 网络社区发现图

Fig. 3 Community detection of Karate

3.5 算法步骤

SNV 算法的整体步骤如下。

(1)采用 Node2Vec 网络嵌入方法将具有 N 个节点的网络 G 转化为欧氏空间的 N 个 d 维向量表征的数据点。

(2)根据式(1)和节点的表征向量计算互为邻居的节点之间的余弦相似度，连接网络 G 中的每个节点及其最大相似性节点，构成偏好网络。

(3)偏好网络中的每个连通子图作为一个初始社区，得到初始社区集合，每个初始社区的最大度节点作为备选节点集。

(4)计算网络的平均度和备选节点集中关于节点-度的二元组集合，根据节点-度二元组集合，将备选节点集中度小于网络平均度的节点移除。

(5)遍历备选节点集，计算节点集中每个节点到其他节点的最短路径距离，得到关于节点-节点-最短路径的三元组集合。

(6)计算网络 G 的平均最短路径长度，根据式(2)计算备选节点集中每个节点的最短路径阈值，根据节点-节点-最短路径三元组集合移除最短路径小于阈值的节点，筛选得到中心节点集。

(7)将中心节点的表征向量作为 K-Means 算法的质心、中心节点的个数作为 K-Means 算法的聚类数对数据进行聚类操作。

(8)聚类之后的向量对应原始网络的节点，得到社区发现结果。

3.6 时间复杂度分析

该算法大致分成 3 个阶段。在第 1 个阶段，我们利用 Node2Vec 算法得到网络中每个节点的低维向量表示，第 1 阶段的时间复杂度为 $O(N)$；在第 2 个阶段，我们计算每个有连边的节点对之间的余弦相似度并构建偏好网络，第 2 阶段的时间复杂度为 $O(E)$；在第 3 阶段，我们使用 K-Means 算法对数据进行聚类操作，所以第 3 阶段的时间复杂度为 $O(N)$。因为 $E\gg N$，所以整个算法的时间复杂度为 $O(E)$。

4 数值仿真

为了验证 SNV 算法的有效性，我们在真实网络和人工网络上进行了一系列的数值仿真实验，将本文算法与 5 个知名算法进行了比较，对比算法包括 K-means 聚类算法[25]、谱聚类算法[16]、标签传播算法[6]（LPA）、Fastgreedy 算法[7] 和 Eigenvector 算法[26]。

4.1 评价指标

（1）模块度 Q[27]：该指标用于评价网络社区划分结果的优劣，模块度值越大，表示社区结构越合理，反之模块度越小，表明社区划分结果越差。模块度的计算方法如下：

$$Q=\sum_{s=1}^{k}\left[\frac{l_s}{l}-\left(\frac{d_s}{2l}\right)^2\right] \tag{3}$$

（2）标准化互信息[28]（NMI）和调整互信息[28]（AMI）是目前广泛使用的用于测量两种数据分布之间吻合程度的一种方法，通常作为评价聚类算法的指标，也可以用来计算社区划分的准确性。在信息论中熵被定义为数据分布中包含的信息，其公式如下：

$$MI(U,V)=\sum_{i=1}^{|U|}\sum_{j=1}^{|V|}P(i,j)\log\left(\frac{P(i,j)}{P(i)P'(j)}\right)$$
$$P(i,j)=|U_i\cap U_j|/N \tag{4}$$

其中，U 和 V 是对应于熵 $H(U)$ 和 $H(V)$ 的标签。NMI[28] 和 AMI[28] 的公式分别如下：

$$NMI(U,V)=\frac{MI(U,V)}{\sqrt{H(U)H(V)}} \tag{5}$$

$$AMI(U,V)=\frac{MI-E[MI]}{\max(H(U),H(V))-E[MI]} \tag{6}$$

其中，$E[MI]$表示 MI 的期望值，NMI 和 AMI 的取值范围分别是[0,1]和[−1,1]。

（3）调整兰德系数[28]（ARI）：ARI[29] 的取值范围为[−1,1]，其绝对值越大，聚类结果越好。ARI 的公式如下：

$$ARI=\frac{RI-E[RI]}{\max(RI)-E[RI]} \tag{7}$$

其中，RI 表示兰德系数，其取值范围为[0,1]，该值越大，说明聚类效果越好。

4.2 真实网络

本文使用 4 个带有真实划分的无向无权网络和 4 个没有真实划分的无向无权网络来测试 SNV 算法的效率和准确率，数据集类型包括人类社交网络、通信网络、基础设施网络、词汇网络等，数据的信息如表 2 和表 3 所列。其中，Network 表示网络名称，Node 表示网络的节点数，Edge 表示网络中节点连边数。数据集包括空手道俱乐部网络[29]（Karate）、海豚网络[30]（Dolphins）、美国大学生橄榄球网络[31]（Football）、政治图书网络[32]（Polbooks）、《悲惨世界》小说人物网络[33]（Lesmis）、爵士音乐人合作网络[34]（Jazz）、《大卫》书籍词汇网络[26]（David）以及美国大学生邮件网络[35]（Email）。

表 2 带有真实标签的网络统计信息

Table 2 Network statistics with true grouth

Network	Node	Edge
Karate[29]	34	78
Dolphins[30]	62	159
Football[31]	115	613
Polbooks[32]	105	441

表 3 不带有真实标签的网络统计信息

Table 3 Network statistics without true grouth

Network	Node	Edge
Lesmis[33]	77	254
Jazz[34]	198	2700
David[26]	112	613
Email[35]	1133	5451

表 4 列出了 SNV 算法和其他 5 种常见算法在具有真实划分的网络上进行社区划分任务的 NMI 值和模块度值。可以看出，在 4 个网络中，SNV 都得到了最高的 NMI 值和两个最高的 Q 值，在所有算法中总体表现最佳。在 Karate 网络上，SNV 算法取得了最高的 NMI 值（为 1），表示 SNV 算法划分的社区完全正确；SNV 算法取得了最高的模块度值（为 0.37），与 Fastgreedy 算法和 Eigenvector 算法取得的模块度值相同。在 Dolphins 网络上，SNV 取得了最高的 NMI 值 0.78，模块度值为 0.36，LPA 算法取得了最高的模块度值 0.49。在 Football 网络上，SNV 算法取得了最高的 NMI 值 0.92，Eigenvector 算法取得的 NMI 值最低，为 0.70，SNV 算法同样取得了最高的模块度值 0.60，Eigenvector 算法取得的模块度值最低，为 0.49。在 Polbooks 网络上，SNV 算法取得了最高的 NMI 值 0.63，其他算法的 NMI 值普遍较低，Spectral-Clustering 算法取得了最高的模块度值 0.50，略高于 SNV 的模块度值 0.45。

表 4 各类算法在具有真实划分的网络上的 NMI 和模块度 Q 比较

Table 4 Comparison of NMI and modularity Q of various algorithms on a network with true grouth

Networks	SNV		Fastgreedy		Spectral-Clustering		K-Means		LPA		Eigenvector	
	NMI	Q	NMI	Q	NMI	Q	NMI	Q	NMI	Q	NMI	Q
Karate	**1.0**	**0.37**	0.70	0.37	0.54	0.31	0.61	0.24	0.44	0.32	0.71	0.37
Dolphins	**0.78**	0.36	0.60	0.49	0.46	0.47	0.28	0.36	0.57	**0.49**	0.49	0.49
Football	**0.92**	**0.60**	0.70	0.54	0.82	0.59	0.82	0.58	0.86	0.58	0.70	0.49
Polbooks	**0.63**	0.45	0.53	0.40	0.48	**0.50**	0.42	0.33	0.53	0.48	0.52	0.46

表 5 列出了 SNV 算法和其他 5 种常见算法在具有真实划分的网络上进行社区划分任务的 AMI 值和 ARI。SNV 算法在 4 个真实网络上的 AMI 值和 ARI 值都取得了最高值，在所有算法中总体表现最佳。K-Means 算法和 Spectral Clustering 算法由于没有进行参数调优，在真实网络上的表现一般。在 Karate 网络上，SNV 算法取得了最高的 AMI 值和 ARI 值，均为 1，表示 SNV 算法划分的社区完全正确。在 Dolphins 网络上，SNV 算法取得了最高的

AMI 值 0.77，并取得了最高的 *ARI* 值 0.85。在 Football 网络上，SNV 算法取得了最高的 *AMI* 值 0.90，并取得了最高的 *ARI* 值 0.86。在 Polbooks 网络上，SNV 算法取得了最高的 *AMI* 值 0.63，并取得了最高的 *ARI* 值 0.70。实验表明，SNV 算法在真实网络上的效果优于其他算法。

表 5 各类算法在具有真实划分的网络上的 *AMI* 和 *ARI* 比较

Table 5 Comparison of *AMI* and *ARI* of various algorithms on a network with true grouth

Networks	SNV		Fastgreedy		Spectral-clustering		*K*-Means		LPA		Eigenvector	
	AMI	*ARI*	*AMI*	*ARI*	*AMI*	*ARI*	*AMI*	*ARI*	*AMI*	*ARI*	*AMI*	*ARI*
Karate	**1.0**	**1.0**	0.68	0.68	0.42	0.23	0.56	0.40	0.42	0.47	0.60	0.51
Dolphins	**0.77**	**0.85**	0.56	0.45	0.37	0.20	0.24	0.16	0.50	0.36	0.43	0.28
Football	**0.90**	**0.86**	0.65	0.47	0.78	0.65	0.79	0.61	0.83	0.75	0.63	0.46
Polbooks	**0.63**	**0.70**	0.52	0.64	0.43	0.40	0.38	0.28	0.52	0.59	0.51	0.54

表 6 列出了 SNV 算法和其他 5 种常用算法在不具有真实划分的数据集上的模块度值表现。在大部分数据集上，SNV 算法都取得了良好的效果。在 Lesmis 网络上，Spectral Clustering 算法取得了最大的模块度值 0.54，*K*-Means 算法取得了最低的模块度值 0.32，SNV 取得的模块度值为 0.41。在 Jazz 网络上，SNV 取得了最高的模块度值 0.43，高于其他算法，LPA 算法取得了最低的模块度值 0.28。在 David 网络上，Fastgreedy 取得了最高的模块度值 0.29，SNV 算法取得的模块度值为 0.16，略低于 Fastgreedy 算法。在 Email 网络上，SNV 取得了最高的模块度值 0.56，*K*-Means 算法取得了最低的模块度值 0.32。

表 6 各类算法在不具有真实划分的网络上模块度比较

Table 6 Comparison of modularity of various algorithms on a network with true grouth

Networks	SNV	Fastgreedy	Spectral-clustering	*K*-Means	LPA	Eigenvector
	Q	*Q*	*Q*	*Q*	*Q*	*Q*
Lesmis	0.41	0.50	**0.54**	0.32	0.52	0.53
Jazz	**0.43**	0.42	0.38	0.31	0.28	0.39
David	0.16	**0.29**	0.23	0.20	0.15	0.24
Email	**0.56**	0.50	0.49	0.32	0.46	0.48

总体而言，在 8 个真实的网络数据集上，与其他 5 种算法相比，SNV 算法取得了绝对的优势，尤其是在 *NMI* 值的表现上，整体性能最佳。

4.3 人工网络

我们使用的人工网络是 LFR 基准网络[36]，这些网络具有节点度和社区规模的幂律分布，是真实网络的特征。因此，它始终被视为具有社区结构的真实网络的替代品，并用于评估社区发现算法的性能。LFR 网络的参数定义如表 7 所列。

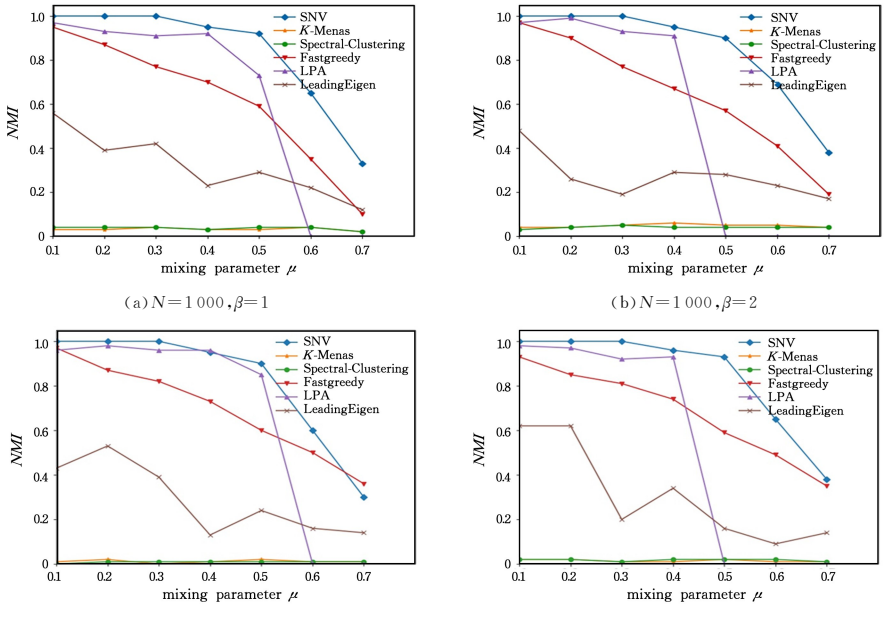
表 7 LFR 网络的参数

Table 7 LFR network parameters

Parameters	Description
N	number of nodes
$\langle K\rangle$	average degree
max k	maximum degree
α	power-law exponent for the degree sequence
β	power-law exponent for the community size distribution
min c	minimum for the community sizes
max c	maximum for the community sizes
μ	mixing parameter

图 4 给出了 SNV 算法和其他 5 种算法在 LFR 人工网络上的 *NMI* 对比。图 4 表明，随着网络规模和混合参数的增大，算法的 *NMI* 值不断降低。这是因为混合参数越多意味着网络结构越模糊复杂，因此，随着混合参数的增多，准确地检测出社区变得越来越困难。图 4(a)和图 4(b)给出了节点个数为 1000、社区大小分布的幂律分布指数分别是 1，2 的 LFR 网络的 *NMI* 指标结果。其中 LFR 网络的参数为：节点个数为 $N=1000$，网络的平均度大小为 20，网络中的节点最大度为 100，度分布幂律指数为 2，社区大小分布的幂律指数分别设置为 1 和 2，最小的社区规模为 20，最大的社区规模为 100，混合参数设置为 0.1～0.7。*K*-Means 算法和 Spectral-Clustering 算法在 LFR 网络的表现较差，这是由于本文并没有对 *K*-Means 算法和 Spectral-Clustering 算法进行调优。当 $\mu\leqslant 0.3$ 时，Fastgreedy 算法和 Eigenvector 算法的 *NMI* 值都出现了明显的降低，LPA 算法的 *NMI* 值随着混合参数的增大出现了略微的降低，其中 Eigenvector 算法的表现最差，SNV 算法的 *NMI* 值并没有下降，接近 1.0。当 $\mu\leqslant 0.5$ 时，LPA 算法、Fastgreedy 算法和 Eigenvector 算法的 *NMI* 值出现了明显的下降，LPA，SNV 算法的 *NMI* 值出现了略微的下降，但是 SNV 算法的 *NMI* 值大于其他的算法。当 $\mu\leqslant 0.7$ 时，所有算法的 *NMI* 值都出现了下降，除了 SNV 算法，其余算法在 $\mu=0.7$ 时基本失效了。

图 4(c)和图 4(d)展现了节点个数为 3000、社区大小分布的幂律分布指数分别是 1，2 的 LFR 网络的 *NMI* 指标结果。其中 LFR 网络的参数为：节点个数为 $N=3000$，网络的平均度大小为 50，网络中的节点最大度为 300，度分布幂律指数为 2，社区大小分布的幂律指数分别设置为 1 和 2，最小的社区规模等于 50，最大的社区规模等于 300，混合参数设置为 0.1～0.7。虽然网络规模与社区的规模与之前的网络有很大区别，但是从图中可以看出，SNV 和其他 5 种算法在 3000 个节点的 LFR 网络上的表现与在 1000 个节点规模的 LFR 网络上的表现基本相同。



(a) $N=1\,000$, $\beta=1$　(b) $N=1\,000$, $\beta=2$

(c) $N=3\,000$, $\beta=1$　(d) $N=3\,000$, $\beta=2$

图 4　SNV 和其他算法在 LFR 人工网络上的 *NMI* 值比较

Fig. 4　Comparison of *NMI* value of SNV and other algorithms on LFR artificial network

综上所述,SNV 算法取了最高的 *NMI* 值,在所有算法中表现最佳。

结束语　本文提出的 SNV 算法将传统复杂网络领域的节点相似性和目前新兴的网络嵌入算法相结合,完成社区划分任务。SNV 算法的特点是可以自动地从原始网络中利用节点相似性、最短路径等属性提取网络中的中心节点,自动设定 *K*-Means 算法中的聚类数目和质心,无须针对不同规模的网络人工设定不同的参数,可以自动地快速识别复杂网络的社区结构。本文基于网络拓扑结构挖掘社区,而现实中网络节点具有丰富的属性信息,比如电商网络中的商品不仅与用户有购买的关系,本身还具有产地、厂家等属性信息,因此未来我们考虑对属性网络进行社区发现和数据挖掘研究,其在社交网络和电商网络社区划分领域具有广阔的应用前景。

参 考 文 献

[1] SAGANOWSKI S. Predicting Community Evolution in Social Networks[C]// the 2015 IEEE/ACM International Conference. ACM, 2015: 3053-3096.

[2] ACMAN M, DORP L V, SANTINI J M, et al. Large-scale network analysis captures biological features of bacterial plasmids [J]. Nature Communications, 2020, 11(1): 46.

[3] CHOUCHANI N, ABED M. Online social network analysis: detection of communities of interest[J]. Journal of Intelligent Information Systems, 2020, 54(2): 1-17.

[4] NEWMAN M, GIRVAN M. Finding and evaluating community structure in networks[J]. Physica A, 2004, 69(2): 26113.

[5] KRAMER J, BOONE L, CLIFFORD T, et al. Analysis of Medical Data Using Community Detection on Inferred Networks[J]. IEEE Journal of Biomedical and Health Informatics, 2020, 24(11): 3136-3143.

[6] KOUNI I B E, KAROUI W, ROMDHANEL B. Node Importance based Label Propagation Algorithm for overlapping community detection in networks[J]. Expert Systems with Applications, 2019, 162: 113020.

[7] ZIVICH P N, SMITH N R, FRERICHSL M, et al. A Guide for Choosing Community Detection Algorithms in Social Network Studies: The Question Alignment Approach[J]. American Journal of Preventive Medicine, 2020, 59(4): 597-605.

[8] ROSVALL M, BERGSTROM C T. Maps of random walks on complex networks reveal community structure[J]. Proceedings of the National Academy of Sciences of the United States of America, 2008, 105(4): 1118-1127.

[9] ZHOU T, LV L Y, ZHANG Y C. Predicting missing links via local information[J]. European Physical Journal B, 2009, 71(4): 623-630.

[10] TASGIN M, BINGOL H O. Community detection using preference networks[J]. Physica A: Statistical Mechanics and its Applications, 2018, 495: 126-136.

[11] LIU Z, MA Y. A Divide and Agglomerate Algorithm for Community Detection in Social Networks[J]. Information Sciences, 2019, 482: 321-333.

[12] ZHAO X, ZHANG Z H, ZHANG C W, et al. RGNE: A coarse-grained network embedded overlapping community discovery method[J]. Computer Research and Development, 2020, 57(6): 1302-1311.

[13] YE Z L, ZHAO H X, ZHANG K, et al. Network representation learning based on neighbor node and relationship model optimization[J]. Computer Research and Development, 2019, 56(12): 2562-2577.

[14] ZHAO X, LI X, ZHANG Z H, et al. Community discovery algorithm combining community embedding and node embedding [J]. Computer Science, 2020, 47(10): 121-125.

[15] XIE Y, GONG M, WANG S, et al. Community Discovery in Networks with Deep Sparse Filtering[J]. Pattern Recognition, 2018, 81: 50-59.

[16] HU F, LIU J, LI L, et al. Community detection in complex networks using Node2vec with spectral clustering[J]. Physica A: Statistical Mechanics and its Applications, 2020, 45(2): 13247.
[17] DING Y, WEI H, PAN Z S, et al. Overview of network representation learning algorithms [J]. Computer Science, 2020, 47(9): 52-59.
[18] PEROZZI B, AL-RFOU R, SKIENA S. Deepwalk: Online learning of social representations [C] // Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2014: 701-710.
[19] WANG D, CUI P, ZHU W. Structural deep network embedding [C] // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016: 1225-1234.
[20] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[C] // 5th International Conference on Learning Representations. ICLR, 2017.
[21] YE J. Improved cosine similarity measures of simplified neutrosophic sets for medical diagnoses [J]. Artificial Intelligence in Medicine, 2015, 63(3): 171-179.
[22] ADAMIC L A, ADAR E. Friends and neighbors on the Web[J]. Social Networks, 2003, 25(3): 211-230.
[23] RAVASZ E, SOMERA A L, MONGRU D A, et al. Hierarchical Organization of Modularity in Metabolic Networks[J]. Science, 2002, 297(5586): 1551-1555.
[24] LEICHT E A, HOLME P, NEWMAN M E J. Vertex similarity in networks[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2006, 73(2): 026120.
[25] GODICHON-BAGGIONI A, MAUGIS-RABUSSEAU C, RAU A. Clustering transformed compositional data using K-means, with applications in gene expression and bicycle sharing system data[J]. Journal of Applied Statistics, 2017, 46(1): 47-65.
[26] NEWMAN M E J. Finding community structure in networks using the eigenvectors of matrices[J]. Physical Review E, 2006, 74(3): 036104.
[27] CHATTOPADHYAY S, BASU T, DAS A K, et al. A similarity based generalized modularity measure towards effective community discovery in complex networks[J]. Physica A: Statistical Mechanics and its Applications, 2019, 527: 121338.
[28] HESAMIPOUR S, BALAFAR M A. A new method for detecting communities and their centers using the Adamic/Adar Index and game theory[J]. Physica A: Statistical Mechanics and its Applications, 2019, 535: 122354.
[29] AHAJJAM S, EL HADDAD M, BADIR H. A new scalable leader-community detection approach for community detection in social networks[J]. Social Networks, 2018, 54: 41-49.
[30] MA T, LIU Q, CAO J, et al. LGIEM: Global and local node influence based community detection[J]. Future Generation Computer Systems, 2020, 105: 533-546.
[31] LIU S S, XIA Z Y. A two-stage BFS local community detection algorithm based on node transfer similarity and Local Clustering Coefficient-Science Direct[J]. Physica A: Statistical Mechanics and its Applications, 2019(537): 122717.
[32] LUO W J, LU N. Local community detection by the nearest nodes with greater centrality[J]. Information Sciences, 2020, 517: 377-392.
[33] YOU X, MA Y, LIU Z. A three-stage algorithm on community detection in social networks [J]. Knowledge-Based Systems, 2020, 187(Jan.): 104822. 1-104822. 12.
[34] CHEN D, SU H. Framework based on communicability to measure the similarity of nodes in complex networks[J]. Information Sciences, 2020, 524: 241-253.
[35] GUIMERÀ R, DANON L, DÍAZ-GUILERA A, et al. Self-similar community structure in a network of human interactions[J]. Physical Review E, 2004, 68(6 Pt 2): 065103.
[36] XU R B, CHE Y, WANG X M, et al. Stacked autoencoder-based community detection method via an ensemble clustering framework[J]. Information Sciences, 2020, 526: 151-165.

YANG Xu-hua, born in 1971, Ph.D, professor, is a senior member of China Computer Federation. His main research interests include machine learning and network science.

LONG Hai-xia, born in 1987, Ph.D, lecturer. Her main research interests include machine learning, medical image processing and complex network analysis.

(责任编辑:喻藜)