基于矢量影响力聚类系数的高效有向网络社团划分算法

邓小龙*[①] 翟佳羽[②] 尹栾玉[③]

[①](北京邮电大学网络空间安全学院可信分布式计算与服务教育部重点实验室 北京 100876)

[②](北京邮电大学国际学院 北京 100876)

[③](北京师范大学中国社会管理研究院 北京 100875)

摘 要：社团结构划分对于分析复杂网络的统计特性非常重要，以往研究往往侧重对无向网络的社团结构挖掘，对新兴的微信朋友圈网络、微博关注网络等涉及较少，并且缺乏高效的划分工具。为解决传统社团划分算法在大规模有向社交网络上无精确划分模拟模型，算法运行效率低，精度偏差大的问题。该文从构成社团结构最基础的三角形极大团展开数学推导，对网络节点的局部信息传递过程进行建模，并引入概率图有向矢量计算理论，对有向社交网络中具有较大信息传递增益的节点从数学基础创造性地构建了有向传递增益系数(Information Transfer Gain, ITG)。该文以此构建了新的有向社团结构划分效果的目标函数，提出了新型有向网络社团划分算法 ITG，通过在模拟网络数据集和真实网络数据集上进行实验，验证了所提算法的精确性和新颖性，并优于 FastGN, OSLOM 和 Infomap 等经典算法。

关键词：有向社团划分；信息传递增益；目标函数优化；算法可扩展性

中图分类号：TP393; TP391 **文献标识码**：A **文章编号**：1009-5896(2017)09-2071-10

DOI: 10.11999/JEIT170102

Vector Influence Clustering Coefficient Based Efficient Directed Community Detection Algorithm

DENG Xiaolong[①] ZHAI Jiayu[②] YIN Luanyu[③]

[①](*Key Laboratory of Trustworthy Distributed Computing and Service of Education Ministry*, *Beijing University of Posts and Telecommunications*, *Beijing* 100876, *China*)

[②](*International School*, *Beijing University of Posts and Telecommunications*, *Beijing* 100876, *China*)

[③](*China Academy of Social Management*, *Beijing Normal University*, *Beijing* 100875, *China*)

Abstract: Community detection method is significant to character statistics of complex network. Community detection in directed structured network is an attractive research problem while most previous approaches attempt to divide undirected networks into communities while there has appeared many large scale directed social network such as WeChat circle of friends and Sina Micro-Blog. To solve the problem that low quality of model, low efficiency of execution and high deviation of precision from the conventional community detection algorithm on large-scale social network and directed network, this paper provides an approach that starts with the triangle structure of community basis and models the local information transfer to detect community in large-scale directed social network. Basing on the directed vector theory in probability graph and the high information transfer gain of vertex in directed network, this paper constructs the Information Transfer Gain (ITG) method and the corresponding target functions for evaluating the quality of a specific partition in community detection algorithm. Then the combine of ITG with the target function to compose the new community detection algorithm for directed network. Extensive experiments in synthetic signed network and real-life large networks derived from online social media, it is proved that the proposed method is more accurate and faster than several traditional community detection methods such as FastGN, OSLOM and Infomap.

Key words: Community detection in directed network; Information Transfer Gain (ITG); Target function optimization; Scalability

1 引言

目前小规模的社会网络研究已经比较成熟，而互联网中已形成了海量规模的社交网络。截止到2015 年 12 月 31 日，Twitter 全球用户数已超过 5 亿，其中活跃用户数超 2 亿；微信用户数则已超过 6 亿，其中活跃用户数超 4 亿。这些大规模社会网络中的社区发现对于进一步研究网络的拓扑结构和层次结构至关重要。因为大规模社交网络结构复杂，包含海量的结点和边，一般图分析方法很难对其进行系统的研究。但社区发现可以得到网络层次结构关系，并根据需要在不同层次上进行网络分析，以发现其中蕴藏的知识。但目前的社区发现算法多半无法适应大规模有向社交网络中社团发现的工作，如新浪微博的粉丝 Follow 关系、微信朋友圈中的双向朋友圈查看权限等，都是重要的有向社交网络关系。因此需要研究效率高、算法结果精确的社团划分算法，从庞大复杂网络中发现潜在的社区结构，满足研究的需要。

为解决传统社团划分算法在大规模网络和有向社交网络无精确模拟模型和算法运行效率低，精度偏差大的问题，本文从构成社团结构基础的三角形极大团入手，对网络中节点的局部信息传递进行建模，引入概率图有向矢量计算理论，对有向社交网络中具有较大信息传递增益的节点从数学基础上构建了有向传递增益系数，并构建了对应的衡量社团结构划分效果的目标函数，构建了新型的可并行大规模社交网络社团划分算法，并通过在模拟网络数据集和真实网络数据集上进行实验，充分验证了本文所提算法的精确性和新颖性。

本文引言阐述了当前社团发现算法的类型和关键算法的优劣情况以及其在有向社交网络中实现的可行性。第 2 节详细介绍了基于信息传递增益的有向网络社团发现算法。第 3 节将本文所提算法与其他主流算法在各种数据集上的社团发现情况进行了比较。

网络中的“社区”是根据某种或某几种性质对网络进行的划分，由结点和边组成，也被称为群组(group)、聚类(cluster)或模块(module)。社团结构(community structure)是一组同社团内节点相互连接密集、异社团间节点相互连接稀疏特点的节点，真实社团结构对于信息传播和社交网络中信息影响力模型的构建，具有重要意义。当前社团划分算法，依据分析对象的不同，可分为以下 4 类：层次聚类方法、矩阵谱分析方法、基于边图思想的方法和基于极大团思想的方法[1]。而针对大规模网络中的社团划分，当前的划分算法主要有基于模块度 Q 值优化方法、基于随机游走的方法和基于重叠社团划分的方法 3 大类：

(1)基于模块度 Q 值优化的方法：此类算法尝试将小型网络中基于模块度优化的思想带入到大规模网络的社团划分中，通过模块度优化来获得较好社团划分结果，模块度 Q 值函数 $Q=\sum_r\left(e_{rr}-\alpha_r^2\right)$ [2] 由 Newman 提出。若将社区看作一个子图，如果社区内部的边数越多则社区结构越好，对应的模块度函数值越大。此类算法大多思路为大规模网络社团划分中的模块度函数值优化，代表性算法有：2004 年 Newman 等人[2]提出的基于模块度优化的优化算法 FastGN(简称 FN)，利用每次在不同社团之间交换边产生的 Q 值增益来寻找模块度函数值优化的方向，但这种算法在网络规模超过 104 时性能下降较严重。后来，Clauset 等人[3]利用堆结构对 FN 算法进行改进，提出了 CNM 算法，算法复杂度已经接近线性，并且可以适用于大规模网络。2008 年，Blondel 等人[4]提出了一种快速算法 Louvain，该算法也是利用模块度优化进行网络社团划分，最后当模块度收敛于某一最大值时，社团划分停止，这种算法适合于大规模复杂网络。但是 2014 年 WWW2014 发表的文章中，文献[5]认为该算法在网络规模增大时，性能下降较快，需要进行更深入的探究。此外是基于标签传播[6](Label Propagation Algorithm, LPA)系列的大规模社团发现算法，较之其他复杂机器学习算法，LPA 执行复杂度低且分类效果好，时间复杂度为 $O(n^2)$ (n 为网络节点个数)。2007 年，Raghavan 等人[7]将 LPA 算法进行改进，提出了社团划分操作与网络规模成正比的近似线性增长 RAK 算法，通过预先定义目标函数简化 LPA 迭代复杂度，利用网络结构作为指导来探测社区结构。RAK 在空手道俱乐部网和美国大学橄榄球网数据集上社区检测效果良好，但在 LFR benchmark 网络上实验结果存在一定缺陷。模块度类算法的解决思路有限制，当图规模很大时，不能发现小的并且符合定义的社团。

(2)基于随机游走的方法：随机游走方法本质上

收稿日期：2017-01-25；改回日期：2017-08-16；网络出版：2017-08-18
*通信作者：邓小龙 shannondeng@bupt.edu.cn
基金项目：国家 973 计划项目(2013CB 329600)，教育部哲学社会科学重大攻关项目(15JZD027)，“十二五”国家科技支撑计划国家文化科技创新工程 2013 年备选项目(2013BAH43F01)
Foundation Items: The National 973 Project of China (2013CB 329600), The Philosophy and Social Science Project of Education Ministry (15JZD027), The National Culture Support Foundation Project of China (2013BAH43F01)

具有信息更容易在高密度社团内部更多“游走”的本质。不同于模块度的优化，此类算法侧重于从信息的传播或者某些物理物质的渗流过程，快速而且较好地获得社团结构，此类代表性算法有：2006 年，Pons 等人[8]提出的针对大规模网络的基于节点相似性的随机游走社团划分算法 Walktrap，利用欧氏距离定义了不同社团之间的距离，该算法具有较好的时间复杂度。2008 年，Rosvall 等人[9]对基于随机游走的社团划分算法进行了详细的综述性介绍，并通过将不同节点间信息流动的概率通过信息论中的信息熵函数进行建模，提出了 Infomap 算法，经过在多个大规模科学家合作网和 LFR[10]标准测试网络上的比较性实验，该算法性能优越，优于基于重叠社团划分的相关算法。

(3)基于重叠社团划分的方法：该类社团划分方法的构建思路不同于前面所提的算法，代表性算法有：2005 年，Palla 等人[11]提出了派系逾渗算法 CPM (Clique Percolation Method)，基于社团内部的边往往具有较紧密的连接以形成一个派系组成一个社团，相反，具有通用性的一些边并不一定是来自一个派系。CPM 算法对派系彼此互通的限制条件过于苛刻，使该算法的缺陷导致搜寻的时间复杂度较高。2010 年，Ahn 等人[12]通过对边建模而不是对点建模提出了 LCA 算法(Link Clustering Algorithm)，在计算过程中采用 Jaccard coefficient 聚集系数计算相连边之间的相似性，使得重叠社团的出现变得很自然。2011 年，Filippo 等人[13]提出了类似于 Q 值的 Significance 函数作为划分适合度的度量函数，并基于此提出了适合大规模网络的 OSLOM(Order Statistics Local Optimization Method) 算法，OSLOM 算法是首个可针对有向有权边的网络进行社团划分的新颖算法。2013 年，Yang 等人[14]对 LPA 算法进行改进，在非负矩阵上通过优化群组节点对于社团附属关系的目标函数值，提出了 BigClam 算法，该算法通过对不同社团间重叠社团重新进行定义，在大规模网络上获得了较好的实验结果。

但是以上社团划分算法并未针对大规模有向社交网络进行有针对性的社团划分，算法运行时间较长且精度欠佳，因此需构建效率更高算法提升运行效率。

2　基于矢量影响力聚类的高效有向网络社团划分算法

真实社交网络朋友关系中，某人的两个朋友也很可能彼此是朋友，这种属性称为网络的聚类特性[15]。聚类特性反映了网络的宏观聚类特性，是网络中相邻节点之间联系紧密程度的重要衡量指标，其定义为：在具有 N 个节点的网络中一个节点 i 有 k_i 条边将它和其他节点相连，即节点 i 存在 k_i 个邻居节点，假设这 k_i 个节点间实际存在的边数为 E_i，则节点 i 的聚类系数定义为

$$C_i = \frac{2E_i}{k_i(k_i-1)} \tag{1}$$

从几何意义上看，式(1)中 C_i 等价定义为

$$C_i = \frac{\text{与节点}i\text{相连的三角形的数量}}{\text{与节点}i\text{相连的三元组的数量}} \tag{2}$$

其中，与节点 i 相连的三元组是指包括节点 i 的 3 个节点，并且至少存在从节点 i 到其他两个节点 j 和 k 的两条边的结构，如图 1 为三元组的两种可能结构。

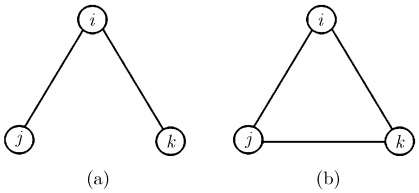

图 1　以点 i 为一个顶点的三元组的可能两组形式

同时，网络的平均聚类系数定义为

$$C_{G=(V,\varepsilon)} \equiv \frac{1}{N}\sum_{v_i\in V, i=1}^{N}\frac{2E_i}{k_i(k_i-1)} \tag{3}$$

$C_{G=(V,\varepsilon)}$ 可用来测量网络中三角结构连接的密度，三角结构所占比重越大，则这些三角结构所属的社团结构内部连接更紧密，网络的平均聚类系数也越大[15]。整个网络的平均聚类系数 $0 \le C_{G=(V,\varepsilon)} \le 1$，事实上，在很多类型社交网络中，某用户 U 的朋友的朋友 U'' 同时也是 U 的朋友的概率随着网络规模 N 增加时，并不是趋近于最大值 1，而是当 $N \to \infty$ 时，$C_{G=(V,\varepsilon)}=O(1)$ 趋近于某个非零常数，体现了网络中“物以类聚，人以群分”的特性。

2.1　矢量影响力聚类系数模型的增益系数

传统社交网络划分算法建模时，网络边通常作为无权无向边进行处理，忽略了边的有向性，但社交网络中边的方向往往承载了重要的信息传递方向信息和社交网络以重要节点(如意见领袖)作为信息传播源头的特性。本文在图灵奖获得者 Pearl 的经典概率图 PGM(Probabilistic Graphical Model)理论基础上，将不同社交网络节点之间的信息传递边的方向进行有向矢量形式的抽象，提出了兼具信息传播方向和信息传播概率的矢量影响力聚类系数模型。

概率图模型利用图结构表示变量的联合概率分布，近年来已成为解决问题中不确定性推理的研究

热点，PGM 通过图来表示随机变量间的依赖关系，为多变量统计建模提供了有力的表示框架。PGM 的分类可分两种情况：(1)根据边有无方向性分类；(2)根据表示的抽象级别不同分类。根据边有无方向性，PGM 可分为 3 类：(1)有向图模型[16]，也称为贝叶斯网(Bayesian Network, BN)，其网络结构使用有向无环图；(2)无向图模型[16]，也称为马尔可夫网(Markov Network, MN)，其网络结构为无向图；(3)局部有向模型，即同时存在有向边和无向边的模型，包括条件随机场(Conditional Random Field, CRF)和链图(Chain Graph, CG)。而根据表示的抽象级别不同，PGM 可分两类：(1)基于随机变量的概率图模型，如贝叶斯网、马尔可夫网、条件随机场和链图；(2)基于模板的概率图模型。经分析比较，可确定本文研究的社交网络节点中因有向边关系形成的信息传播概率与有向边的方向以及具体的传播模式(对应不同的随机传播概率)具有紧密关系，其模型属于基于边随机变量概率的贝叶斯信息传播网络。首先，从图 1 所示的传统的无向聚类系数的示意图的图 1(a)和图 1(b)出发，我们可推导出图 2 中相关边的信息传递概率及方向，实际上在有向社交网络中的矢量影响力聚类系数应该从图 2 进行分别推导。

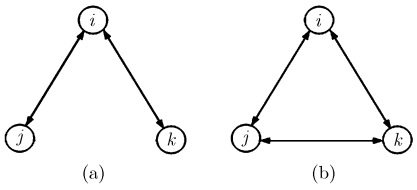

图 2 以点 i 为一个顶点的有向三元组两组基础形式

我们假设，当图 2(a)和图 2(b)中所有的边都是双向时，可将图 2(a)等同于图 1(a)，将图 2(b)等同于图 1(b)，即所有的边均为双向时，所有边上的信息传递增益路径 ITGP(Information Transfer Gain Path)等同于无向网络，回溯到式(1)，可以对 k_i 个节点间实际存在的边数 E_i 中的每条边上的信息传递增益进行拆分。对应于图 2(a)，由于节点 j 和 k 之间没有边相连接，情况比较简单，假设节点 i 到 j 和 k 的信息传递增益 ITG(Information Transfer Gain)定义为

$$\mathrm{ITG}_i=\alpha\times\mathrm{ITG}_{i\leftrightarrow j}+\beta\times\mathrm{ITG}_{i\leftrightarrow k} \tag{4}$$

$$\mathrm{ITG}_{i\leftrightarrow j}=\begin{cases}\delta_{i\to j}\times\mathrm{ITG}_{i\to j}+\delta_{i\leftarrow j}\times\mathrm{ITG}_{i\leftarrow j}, & \delta_{i\to j}\neq 0\\ \theta_{i\to j}\times\delta_{i\leftarrow j}\times\mathrm{ITG}_{i\leftarrow j}, & \delta_{i\to j}=0\end{cases} \tag{4a}$$

$$\mathrm{ITG}_{i\leftrightarrow k}=\begin{cases}\delta_{i\to k}\times\mathrm{ITG}_{i\to k}+\delta_{i\leftarrow k}\times\mathrm{ITG}_{i\leftarrow k}, & \delta_{i\to k}\neq 0\\ \theta_{i\to k}\times\delta_{i\leftarrow k}\times\mathrm{ITG}_{i\leftarrow k}, & \delta_{i\to k}=0\end{cases} \tag{4b}$$

其中，α 是 $\mathrm{ITG}_{i\leftrightarrow j}$ 的概率系数，β 是 $\mathrm{ITG}_{i\leftrightarrow k}$ 的概率系数，对于 ITG_i 而言，$\mathrm{ITG}_{i\leftrightarrow j}$ 和 $\mathrm{ITG}_{i\leftrightarrow k}$ 为分叉累加关系，故 α 和 β 的默认值均为 1。$\delta_{i\to j}$，$\delta_{i\leftarrow j}$，$\delta_{i\to k}$，$\delta_{i\leftarrow k}$ 分别为 $i\to j$，$i\leftarrow j$，$i\to k$，$i\leftarrow k$ 方向的信息传递概率，默认值均为 0.5。$\theta_{i\to j}$ 和 $\theta_{i\to k}$ 为逆向信息传递增益概率，由于图 2 是以节点 i 为信息源点计算聚类系数，类似于真实社交网络中的互为关注或互为粉丝关系，即对于 $\mathrm{ITG}_{i\leftrightarrow j}$ 和 $\mathrm{ITG}_{i\leftrightarrow k}$ 而言，当式(4a)中无 $i\to j$ 方向的边时，$\delta_{i\to j}$ 数值为 0，这时 j 不是 i 的粉丝(或者 j 没有查看 i 的朋友圈信息的权限)，信息无法沿 $i\to j$ 方向进行传播，但是有可能有 $i\leftarrow j$ 的逆向信息，但是我们是以节点 i 为信息源点计算聚类系数，相对 $i\to j$ 方向而言，该信息为逆向信息传递增益，因此我们定义了 $\theta_{i\to j}=0.5$。

对于图 2(b)，节点 i 到节点 j 的信息传递增益路径 ITGP 定义为信息从节点 i 出发，经过路径 $i-j$ 和 $i-k-j$ 给节点 j 带来的信息增益。依据概率图理论模型计算贝叶斯网络概率时，将一个有向无环图抽象为贝叶斯网络，其中的节点代表了随机变量，边代表了随机变量之间的概率关系，其联合概率分布可以用贝叶斯链式法则来表示：

$$P(X_1,X_2,\cdots,X_n)=\prod_i P\left(X_i\left|\mathrm{Par}_G(X_i)\right.\right) \tag{5}$$

其中，$\mathrm{Par}_G(X_i)$ 表示节点 X_i 在图 G 中的父节点对应的随机变量，本文所研究的有向社交网络中因为信息有向性传递带来的从源节点到其他节点的信息影响，也会影响对应社团的结构的形成，因为实际社交网络中往往有粉丝节点聚集在信息传播源头大 V 节点周围，形成社团结构，极大影响了整个社交网络的信息传递模型，因此对应到图 2(a)中的节点 i，可假定 i 相对节点 j 和 k，是“信息源点”。结合概率图理论，从源点到其他节点的概率影响的流动性反映信息传递的流动性。首先依据前面给出的信息增益路径的定义，可得图 2(b)中信息增益路径的公式：

$$\begin{aligned}\sum\mathrm{ITG}_{i\leftrightarrow j}&=\mathrm{ITG}_{i\leftrightarrow j}+\mathrm{ITG}_{i\leftrightarrow k\leftrightarrow j}\\&=\mathrm{ITG}_{i\leftrightarrow j}+\mathrm{ITG}_{i\leftrightarrow k}\times\mathrm{ITG}_{k\leftrightarrow j}\end{aligned} \tag{6}$$

依据信息传递增益的有向性和式(4a)，可以对 $\mathrm{ITG}_{i\leftrightarrow j}$ 进行拆分。同样，依据对称性质，信息传递增益 $\sum\mathrm{ITG}_{i\leftrightarrow k}$ 为

$$\begin{aligned}\sum\mathrm{ITG}_{i\leftrightarrow k}&=\mathrm{ITG}_{i\leftrightarrow k}+\mathrm{ITG}_{i\leftrightarrow j\leftrightarrow k}\\&=\mathrm{ITG}_{i\leftrightarrow k}+\mathrm{ITG}_{i\leftrightarrow j}\times\mathrm{ITG}_{j\leftrightarrow k}\end{aligned} \tag{7}$$

同样，依据对称性质可将式(7)的$\mathrm{ITG}_{i\leftrightarrow k}$拆分为$\delta_{i\to k}\times\mathrm{ITG}_{i\to k}+\delta_{i\leftarrow k}\times\mathrm{ITG}_{i\leftarrow k}$。由于不同节点间信息传递的对称性，假设$\mathrm{ITG}_{i\to k}$和$\mathrm{ITG}_{i\leftarrow k}$数值为默认单位量“1”，$\delta_{i\to k},\delta_{i\leftarrow k}$数值为默认值 0.5，这样节点$i$和$k$间存在双向信息增益传递边时，$\mathrm{ITG}_{i\leftrightarrow k}$值为$0.5\times1+0.5\times1=1$，依据对称性，当节点$i$和$j$间存在双向信息增益传递边时，$\mathrm{ITG}_{i\leftrightarrow j}$值也为$0.5\times1+0.5\times1=1$。针对图 2(a)中的情况进行信息增益路径的具体建模，可得不同的 9 种和 27 种不同的信息传递增益图，其对应的子图分别在图 3、图 4 和图 5 中体现。

首先，依据式(4)、式(4a)和式(4b)我们可以得出如表 1 所示的信息传递增益计算情况。

其中，每一种图对应的ITG_i均可以代表该类型三元组的权重，统计网络图中所有三元组类型的个数并加权于该权重，然后求和，可以得到有向图三元组加权个数。类似地，延伸到对三角形的信息传递增益的计算中，可以得到 27 种信息传递增益图。以图 4(a)为例，节点i,j和k之间的边均为双向连通边，而双向连通边在之前的介绍中系数定义为 1。针对节点i和节点j之间的信息传递增益和由节点i和节点j的单信息传递增益和节点j通过节点k的中继的信息传递增益组成，中继信息传递增益由节点k和节点i与节点j和节点k相加得到。由此可得

$$\begin{aligned}\sum\mathrm{ITG}_{i\leftrightarrow j}&=\mathrm{ITG}_{i\leftrightarrow j}+\mathrm{ITG}_{i\leftrightarrow k\leftrightarrow j}\\&=\mathrm{ITG}_{i\leftrightarrow j}+\mathrm{ITG}_{i\leftrightarrow k}+\mathrm{ITG}_{k\leftrightarrow j}\end{aligned}\tag{8}$$

同理可得针对节点i和节点k之间的信息传递增益和：

$$\sum\mathrm{ITG}_{i\leftrightarrow k}=\mathrm{ITG}_{i\leftrightarrow k}+\mathrm{ITG}_{i\leftrightarrow j}+\mathrm{ITG}_{j\leftrightarrow k}\tag{9}$$

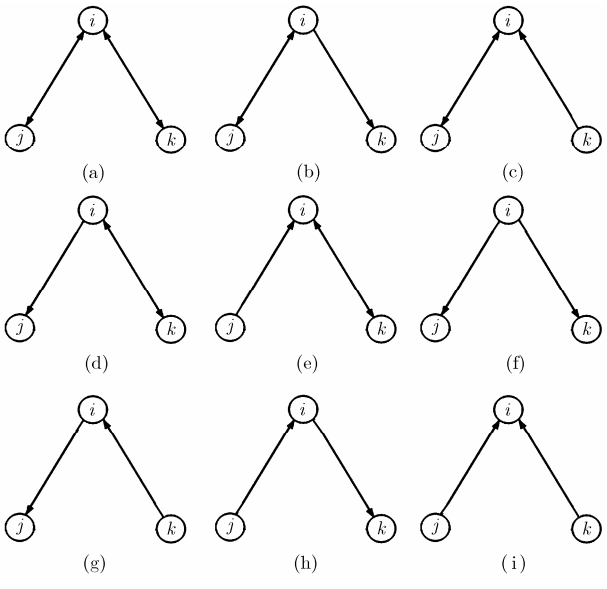

图 3 以节点i为顶点的有向三元组(图 2(a)所有子图)

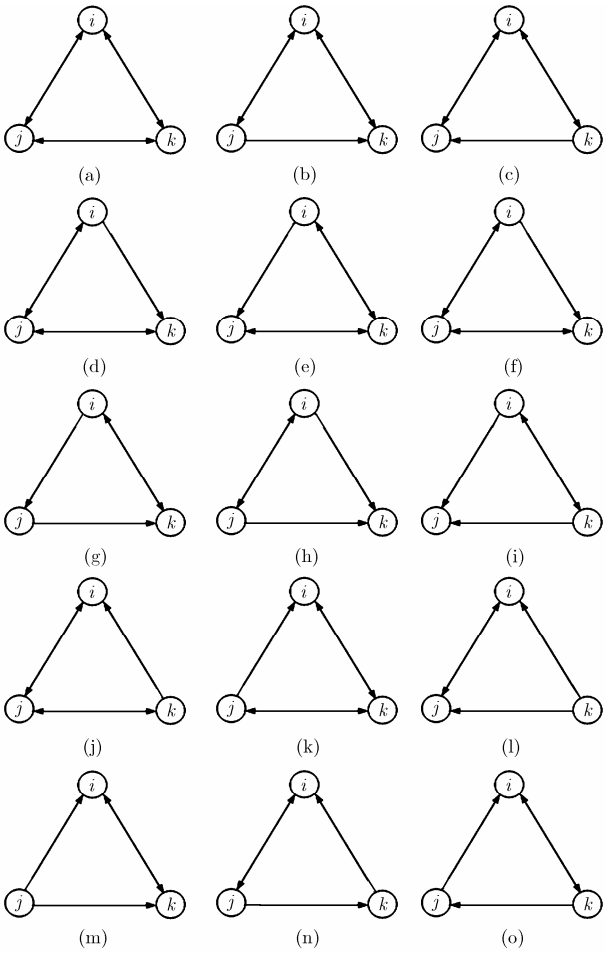

图 4 以节点i为顶点的有向三元组的前 15 个子图

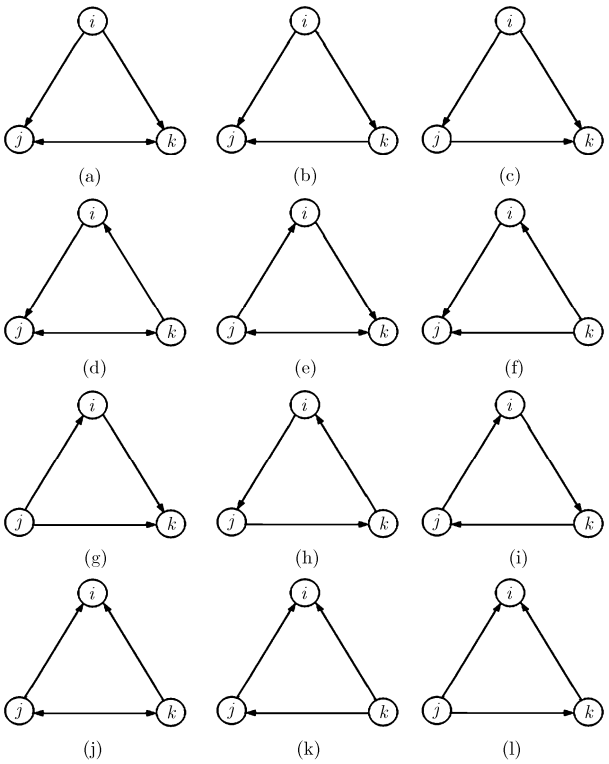

图 5 以节点i为顶点的有向三元组后 12 个子图

表 1 图 3 中各子图源节点的信息传递增益

编号	图序号	ICP_i	$\mathrm{ICP}_{i\leftrightarrow j}$	$\mathrm{ICP}_{i\leftrightarrow k}$
1	图3(a)	2.00	1.00	1.00
2	图3(b)	1.50	1.00	0.50
3	图3(c)	1.25	1.00	0.25
4	图3(d)	1.50	0.50	1.00
5	图3(e)	1.25	0.25	1.00
6	图3(f)	1.00	0.50	0.50
7	图3(g)	0.75	0.50	0.25
8	图3(h)	0.75	0.25	0.50
9	图3(i)	0.50	0.25	0.25

此时，节点 i 的信息传递增益即为节点 i 与两节点的信息传递增益的和，由此整理所有 27 种三角形信息传递增益图，可得表 2。

由于每条边都有 3 种定义，其中三元组的节点 i 的相邻边分别有互相关注、关注、粉丝 3 种定义，节点 i 的对边定义有节点 j 与节点 k 相互关注、左单向关注、右单向关注 3 种定义，因此共有 $3\times3\times3=27$ 种有向三元组表示。将每种边的定义用 0,1 和 2 代替可得如图 6 所示的 27 种排列组合结果。

在以上排列组合结果中，由于是以节点 i 为统计对象，三元组中一些情况是关于节点 i 的对称结果，例如图 4(b)和图 4(c)就是一组对称结果。同样，在表 3 中可以发现色块相同的就是一组对称结果，将所有具有对称结果的情况合并，可以得到 15 种完全独立的情况。据此可以得到有向图中节点 i 信息传递增益系数的定义为

表 2 图 4 和图 5 各子图源节点的信息传递增益

编号	图序号	ITG_i	$\sum\mathrm{ITG}_{i\leftrightarrow j}=\mathrm{ITG}_{i\leftrightarrow j}+\mathrm{ITG}_{i\leftrightarrow k\leftrightarrow j}$ $=\mathrm{ITG}_{i\leftrightarrow j}+\mathrm{ITG}_{i\leftrightarrow k}+\mathrm{ITG}_{k\leftrightarrow j}$	$\sum\mathrm{ITG}_{i\leftrightarrow k}=\mathrm{ITG}_{i\leftrightarrow k}+\mathrm{ITG}_{i\leftrightarrow j}+\mathrm{ITG}_{j\leftrightarrow k}$
1	图4(a)	3.000	0.5×1+0.5×(1+1)=1.500	0.5×1+0.5×(1+1)=1.500
2	图4(b)	2.375	0.5×1+0.5×(1+(0.5×0.5))=1.125	0.5×1+0.5×(1+0.5)=1.250
3	图4(c)	2.375	0.5×1+0.5×(1+0.5)=1.250	0.5×1+0.5×(1+(0.5×0.5))=1.125
4	图4(d)	2.500	0.5×1+0.5×(0.5+1)=1.250	0.5×0.5+0.5×(1+1)=1.250
5	图4(e)	2.500	0.5×0.5+0.5×(1+1)=1.250	0.5×1+0.5×(0.5+1)=1.250
6	图4(f)	1.875	0.5×1+0.5×(0.5+0.5)=1.000	0.5×0.5+0.5×(1+(0.5×0.5))=0.875
7	图4(g)	1.875	0.5×0.5+0.5×(1+(0.5×0.5))=0.875	0.5×1+0.5×(0.5+0.5)=1.000
8	图4(h)	1.875	0.5×1+0.5×(0.5+(0.5×0.5))=0.875	0.5×0.5+0.5×(1+0.5)=1.000
9	图4(i)	1.875	0.5×0.5+0.5×(1+0.5)=1.000	0.5×1+0.5×(0.5+(0.5×0.5))=0.875
10	图4(j)	2.250	0.5×1+0.5×((0.5×0.5)+1)=1.125	0.5×(0.5×0.5)+0.5×(1+1)=1.125
11	图4(k)	2.375	0.5×0.5+0.5×(1+1)=1.250	0.5×1+0.5×((0.5×0.5)+1)=1.125
12	图4(l)	1.750	0.5×1+0.5×((0.5×0.5)+0.5)=0.875	0.5×0.5+0.5×(1+(0.5×0.5))=0.875
13	图4(m)	1.750	0.5×0.5+0.5×(1+(0.5×0.5))=0.875	0.5×1+0.5×((0.5×0.5)+0.5)=0.875
14	图4(n)	1.750	0.5×1+0.5×((0.5×0.5)+(0.5×0.5))=0.750	0.5×0.5+0.5×(1+0.5)=1.000
15	图4(o)	1.625	0.5×(0.5×0.5)+0.5×(1+0.5)=0.875	0.5×1+0.5×((0.5×0.5)+(0.5×0.5))=0.750
16	图5(a)	2.000	0.5×0.5+0.5×(0.5+1)=1.000	0.5×0.5+0.5×(0.5+1)=1.000
17	图5(b)	1.375	0.5×0.5+0.5×(0.5+0.5)=0.750	0.5×0.5+0.5×(0.5+(0.5×0.5))=0.625
18	图5(c)	1.375	0.5×0.5+0.5×(0.5+(0.5×0.5))=0.625	0.5×0.5+0.5×(0.5+0.5)=0.750
19	图5(d)	1.750	0.5×0.5+0.5×((0.5×0.5)+1)=0.875	0.5×(0.5×0.5)+0.5×(0.5+1)=0.875
20	图5(e)	1.750	0.5×(0.5×0.5)+0.5×(0.5+1)=0.875	0.5×0.5+0.5×((0.5×0.5)+1)=0.875
21	图5(f)	1.125	0.5×0.5+0.5×((0.5×0.5)+0.5)=0.625	0.5×(0.5×0.5)+0.5×(0.5+0.5×0.5)=0.500
22	图5(g)	1.125	0.5×(0.5×0.5)+0.5×(0.5+0.5×0.5)=0.500	0.5×0.5+0.5×((0.5×0.5)+0.5)=0.625
23	图5(h)	1.125	0.5×0.5+0.5×((0.5×0.5)+(0.5×0.5))=0.500	0.5×(0.5×0.5)+0.5×(0.5+0.5)=0.625
24	图5(i)	1.125	0.5×(0.5×0.5)+0.5×(0.5+0.5)=0.625	0.5×0.5+0.5×((0.5×0.5)+(0.5×0.5))=0.500
25	图5(j)	1.500	0.5×(0.5×0.5)+0.5×((0.5×0.5)+1)=0.750	0.5×(0.5×0.5)+0.5×((0.5×0.5)+1)=0.750
26	图5(k)	0.875	0.5×(0.5×0.5)+0.5×((0.5×0.5)+0.5)=0.500	0.5×(0.5×0.5)+0.5×((0.5×0.5)+(0.5×0.5))=0.375
27	图5(l)	0.875	0.5×(0.5×0.5)+0.5×((0.5×0.5)+(0.5×0.5))=0.375	0.5×(0.5×0.5)+0.5×((0.5×0.5)+0.5)=0.500

000	001	002	010	011	012	020	021	022
100	101	102	110	111	112	120	121	122
200	201	202	210	211	212	220	221	222

图 6　27 种子图的排列组合

表 3　DWCC_I 估算统计值的含义变化

参数	原统计意义	有向化统计意义
r	社团内节点个数	社团内节点个数
δ	社团边密度	社团加权边密度
d_{in}	节点的社团内邻居个数	节点的社团内加权邻居个数
d_{out}	节点的社团外邻居个数	节点的社团外加权邻居个数
b	社团的边界边个数	社团的加权边个数
ω	网络图平均聚类系数	有向图平均信息传递增益系数

$$\mathrm{ITGC}_i=\frac{\sum_{t=1}^{15}\mathrm{ITG}_{i\_\mathrm{triangle}}(t)\times\mathrm{Number}(t)}{\sum_{t=1}^{6}\mathrm{ITG}_{i\_\mathrm{triple}}(t)\times\mathrm{Number}(t)} \tag{10}$$

其中，ITGC_i 为节点 i 的信息传递增益系数，式(10)的分子为以节点 i 为顶点的三角形加权个数，每种三角形的权值为 15 种不同类型三角形对于节点 i 的信息增益 $\mathrm{ITG}_{i\_\mathrm{triangle}}(t)$ 中的一种；式(10)的分母为以节点 i 为顶点的三元组加权个数，每种三元组的权值为 6 种不同类型三元组对于节点 i 的信息增益 $\mathrm{ITG}_{i\_\mathrm{triple}}(t)$ 中的一种。类似无向图聚类系数，ITGC_i 也反映了节点及其邻节点所形成社团的紧密程度。

2.2 目标函数及其有向化改进

在构建矢量影响力聚类模型目标函数时，基于 Arnau 提出的社团凝聚指数 WCC(Weighted Community Clustering)[5]，我们针对有向图性质进行了有向化改进，将新目标函数定义为有向社团凝聚指数 DWCC(Directed WCC)。在定义 DWCC 中节点在社团中的指数时，$Vt(x,C)$ 表示节点 x 与社团 C 中节点能够形成三角形的加权节点个数。定义为由统计该节点与节点 x 所形成的边的信息传递增益系数时，所统计到的节点个数。加权三角形个数 $wt(x,C)$ 是以节点 x 为源顶点所形成的信息传递增益加权三角形个数。

根据 WCC 的 4 个支持条件使其成为社团发现算法目标函数的基本性质[17]，可延伸其到矢量影响力社团发现算法中，该 4 个支持条件可支撑 DWCC 成为有向网络社团发现算法的目标函数，因此可采用 DWCC 的目标函数迭代优化过程作为矢量影响力社团发现算法的优化部分。

基于 Arnau 提出的目标函数优化过程[5]，关于精细划分的部分涉及到 3 种对社团凝聚指数 WCC 的可能数值增大的函数以及其统一性，即 3 种方式均可转化为 WCC_I 的计算。在有向性改进中，我们同样涉及到了对本文定义的式(11)中 DWCC_I 的估算，以降低计算 DWCC 值的时间复杂度。同时我们对相关的参数统计值进行了有向化改进，具体请参见表 3。

$$\begin{aligned}&\mathrm{DWCC}(P')-\mathrm{DWCC}(P)\\&=\mathrm{DWCC}_I'(v,C)\\&=\frac{1}{V}\cdot\left(d_{\text{in}}\cdot\Theta_1+(r-d_{\text{in}})\cdot\Theta_2+\Theta_3\right)\end{aligned} \tag{11}$$

其中，$q=(b-d_{\text{in}})/r$。

$$\begin{aligned}\Theta_1=\{&((r-1)\delta+1+q)(d_{\text{in}}-1)\delta\}/\{(r+q)((r-1)\\&\cdot(r-2)\delta^3+(d_{\text{in}}-1)\delta+q(q-1)\delta\omega\\&+q(q+1)\omega+d_{\text{out}}\omega)\}\end{aligned} \tag{11a}$$

$$\begin{aligned}\Theta_2=&-\frac{(r-1)(r-2)\delta^3}{(r-1)(r-2)\delta^3+q(q-1)\omega+q(r-1)\delta\omega}\\&\cdot\frac{(r+1)\delta+q}{(r+q)(r-1+q)}\end{aligned} \tag{11b}$$

$$\begin{aligned}\Theta_3=&\frac{d_{\text{in}}(d_{\text{in}}-1)\delta}{d_{\text{in}}(d_{\text{in}}-1)\delta+d_{\text{out}}(d_{\text{out}}-1)\omega+d_{\text{out}}d_{\text{in}}\omega}\\&\cdot\frac{d_{\text{in}}+d_{\text{out}}}{r+d_{\text{out}}}\end{aligned} \tag{11c}$$

在所有统计值给定情况下，式(11)具有常数的时间复杂度，并且节点的统计值更新仅在社团结构发生改变时，对于全图来说时间复杂度为 $O(m)$，m 为社团结构发生改变的次数。

2.3 迭代过程改进和迭代终止条件

在完成上述有向化改进之后，依据有向社团凝聚指数进行计算的目标函数优化过程，能够通过 Arnau[5]的可扩展社团划分算法 SCD 的计算过程进行矢量影响力社团划分算法的模型构建。

初始划分中，我们采用了信息传递增益系数 ITGC 对每个节点进行计算，并作为排序过程参考值。而在精细划分过程中，经过实验过程的检验，发现了算法的贪婪性问题较大，以原有算法结构很难达到算法的有效收敛。因此我们对算法迭代过程的结构进行了改进，加入了标志位 flag 表示当前迭代是否对最大 DWCC 值有更新。并通过得到迭代过程中最大 DWCC 值及其划分方式的方法，来模拟达到近似最优解的过程。经过实验，我们将终止条件

设置为满足以下两点中的一点，即逻辑“或”关系：

(1)对最大 DWCC 值有更新，且更新比例小于阈值比例 t；

(2)未对最大 DWCC 值有更新，且此迭代情况持续了 20 次以上。

图 7 给出了精细划分的 ITG 具体算法细节。首先通过将初步划分的结果作为精细划分的输入，得到初始迭代划分 P，然后开始迭代。每步迭代通过对 P 的 DWCC 值进行计算，并以节点为对象，计算每个节点的最佳移动策略，并将所有最佳移动策略应用于当前划分 P 上。通过计算新的划分 new P 的 DWCC 值并依据上述改进的算法结构，记录最大 DWCC 值所对应的划分 P。最后判断是否达到迭代终止条件来决定是否继续迭代。迭代完成后返回迭代过程中最大的 DWCC 所对应的划分 P 作为算法的输出结果。

Input: Graph $G(V,E)$ and Partition P
Output: newPartition P'
new P=P;
new DWCC=computeDWCC(P);
max DWCC=newDWCC;
preMaxDWCC=0;
flag=FALSE;
counter=0;
repeat
  flag=FALSE;
  DWCC′=newDWCC;
  P'=new$P$;
  M=$\varnothing$;
  foreach v in V **do**
    M.add(BestMovement(v,P'));
  end
  new$P$=applyMovements(M,P');
  newDWCC=computeWCC(new$P$);
  if maxDWCC<newDWCC **then**
    flag=TRUE; preMaxDWCC =maxDWCC; count=0;
    maxDWCC=newDWCC;
  end
  count++;
until NOT(flag AND(maxDWCC −preMaxDWCC)/preMaxDWCC<t) OR counter<20;
return P'.

图 7 ITG 算法精细划分过程

经过实验结果分析可知，该迭代终止条件能够一定程度上消除陷入“局部最优解”的情况，并且不会因过长的迭代次数导致执行效率过低。

3 实验结果

3.1 实验环境

单机处理器 Intel(R) Xeon(R) CPU E5-2440 v2@1.90 GHz，内存 16G，硬盘 4 TB，操作系统为 Windows 7，编程语言为 Matlab 7.0 和 JDK 1.8.0。在并行化计算环节，我们使用了 8 个 Slave 节点，其中 Slave 节点上 CPU 也是 Intel(R) Xeon(R) CPU E5-2440 v2@1.90 GHz，并行化框架采用 Spark 2.1.0 和 Hadoop 2.7.3。

3.2 实验数据集

本文采用了 4 个有代表性的数据集进行实验，既有人工模拟数据集，也有相关经典数据集，还包括了一个真实的国内某城市部分联系紧密人群的移动电话呼叫数据集，表 4 列出了数据集的统计特征以及密度特征：

表 4 数据集基本属性

数据集	节点数	边数	图密度
人工测试集	30	275	9.1667
OSLOM数据集	301	6234	20.7110
学科引用数据集	40	306	7.6500
电话呼叫数据集	284	3934	13.8521

数据集 1：人工数据集，随机生成 3 个社团组成的图。其中，随机生成条件为：社团内部形成边的概率为 0.50，社团之间形成边的概率为 0.25。

数据集 2：OSLOM 数据集，由 OSLOM 开源算法提供的例子数据集。数据集来源 OSLOM 算法源码提供的测试例子数据集 example.dat[13]。

数据集 3：学科引用数据集，由 Infomap 算法提出者提供的学科引用数据集。数据来自 Infomap 算法源代码提供的测试数据集 MultiphysChemBio Eco40W_weighted_dir.net[9]。

数据集 4：移动电话呼叫数据集。数据来源于国内某城市部分联系紧密人群的移动电话呼叫数据集。

实验中将本文所提 ITG 算法与以下 3 种经典算法进行了比较：

FastGN 算法，由 Newman 等人[2]提出，作为快速社团发现典型算法。该算法基于模块度优化，将模块度定义为有向模块度后能够在有向网络中进行实验。

OSLOM 算法：由 Lancichinetti 等人[13]提出，该算法适用于有向网络。

Infomap 算法：由 Rosvall 等人[9]提出，该算法适用于有向网络。

3.3 实验结果

经过对已有研究工作的分析，我们采用以下 4 个评估社团发现结果的经典指标作为实验的比较参数：

(1)社团个数：经过社团发现后，得到的社团个数。

(2)有向模块度：有向模块度用于有向图中模块度的计算[18]，公式为 $Q=\frac{1}{m}\sum_{i,j}\left[\boldsymbol{A}_{ij}-\frac{k_i^{\text{out}}k_j^{\text{in}}}{m}\right]\cdot\delta\left(c_i,c_j\right)$，$m$ 为划分的社团个数，$\boldsymbol{A}_{ij}$ 是网络连接矩阵，k_i^{out} 是节点 i 的出度，k_j^{in} 是节点 j 的入度，$\delta\left(c_i,c_j\right)$ 是克罗内克函数值。

(3)Jaccard 系数：计算发现结果的精确度，计算公式为权威划分结果与算法划分结果的交集与并集的比[19]。

(4)F-measure 系数：衡量发现算法的结果，本实验中采用 F-1 标准[19]。

在相同实验条件下进行实验，得到的实验结果如表 5 至表 9 所示。

实验结果表明，在几个测试数据集上，本文 ITG 算法的社团划分结果与其他算法的划分结果，在精度上达到能够接受的精度标准，在某些情况下甚至明显优于传统社团划分算法。在算法可扩展性上，由于计算每个节点的最佳移动策略中是节点独立(vertex-independent)的计算模式，能够进行并行化操作，因此并行化实验中，ITG 算法在并行环境下具有更高的效率。

表 5 社团个数

数据集	FastGN	OSLOM	Infomap	本文 ITG
人工测试集	5	无法形成社团	1	4
OSLOM数据集	9	9	3	10
学科引用数据集	4	4	4	3
电话呼叫数据集	8	14	22	18

表 6 有向模块度

数据集	FastGN	OSLOM	Infomap	本文 ITG
人工测试集	0.3322	无法形成社团	1.0000	0.3588
OSLOM数据集	0.8651	0.8794	0.9963	0.6910
学科引用数据集	0.8007	0.8072	0.8007	0.8235
电话呼叫数据集	0.5486	0.4822	0.3915	0.4080

表 7　Jaccard 系数

数据集	FastGN	OSLOM	Infomap	本文 ITG
人工测试集	0.2448	无法形成社团	0.3333	0.2660
OSLOM数据集	0.9592	1.0000	0.3851	0.5183
学科引用数据集	1.0000	0.9070	0.9310	0.5974
电话呼叫数据集	1.0000	0.2864	0.2448	0.1121

表 8　F-measure 系数

数据集	FastGN	OSLOM	Infomap	本文 ITG
人工测试集	0.3830	无法形成社团	0.5000	0.4075
OSLOM数据集	0.9700	1.0000	0.3851	0.6386
学科引用数据集	1.0000	0.9070	0.9536	0.7108
电话呼叫数据集	1.0000	0.4149	0.3611	0.1945

表 9 大规模有向网络数据集

数据集	节点数	边数	点边比率
Call L-1	13310	34591	0.3847
Call L-2	29624	55423	0.5345
Call L-3	61510	65202	0.9433
Call L-4	512024	1021861	0.5011

表 9 中为来自中国某城市的连续 4 个月的电话呼叫数据，具有一定的规模。在表 10 中，我们通过并行化计算环境比较了本文所提基于矢量影响力聚类系数的 DITG(Distributed ITG)算法和其他算法的计算性能。通过分析表 10 的实验结果可知，ITG 算法在并行化以后，计算所需的时间，明显低于 FastGN, OSLOM 和 Infomap 算法。

表 10 并行化算法运行实验结果(s)

数据集	FastGN	OSLOM	Infomap	DITG
Call L-1	2906.124	3002.762	3230.453	950.125
Call L-2	2867.634	2994.986	2898.872	901.651
Call L-3	4676.767	5877.877	6030.331	1500.765
Call L-4	46778.771	49001.225	50020.222	8228.472

4　结束语

本文首先结合经典的概率图和聚类系数方法，提出了新的衡量有向图节点的增益系数 ITG，并结合定义的目标函数 DWCC 对社团划分结果进行迭代优化过程。在迭代过程中，由于对节点统计值的计算是相互独立的，因此可以在此处最耗时的步骤进行并行化操作。经过模拟数据集和真实数据集的实验，ITG 算法在精确度上达到了能够接受的精度标准，并在时间复杂度上具有明显优势。后续研究将如下展开：(1)进一步优化算法中的迭代中止条件，使算法的贪婪性得到弥补。(2)在更大规模和更

多类型有向社交网络数据集上验证并行化算法正确性和运行性能，提升算法效果。

参 考 文 献

[1] XIE J, KELLEY S, and SZYMANSKI B K. Overlapping community detection in networks: The state-of-the-art and comparative study[J]. *ACM Computing Surveys* (*CSUR*), 2013, 45(4): 2–35. doi: 10.1145/2501654.2501657.

[2] NEWMAN M E J. Fast algorithm for detecting community structure in networks[J]. *Physical Review E*, 2004, 69(6): 066133. doi: 10.1103/PhysRevE.69.066113.

[3] CLAUSET A, NEWMAN M E J, and MOORE C. Finding community structure in very large networks[J]. *Physical Review E*, 2004, 70(6): 066111. doi: 10.1103/PhysRevE.70.066111.

[4] BLONDEL V D, GUILLAUME J L, LAMBIOTTE R, *et al*. Fast unfolding of communities in large networks[J]. *Journal of Statistical Mechanics*: *Theory and Experiment*, 2008, 2008(10): 1–12. doi: 10.1088/1742-5468/2008/10/P10008.

[5] PRAT-PÉREZ A, DOMINGUEZ-SAL D, and LARRIBA-PEY J L. High quality, scalable and parallel community detection for large real graphs[C]. The 23rd International Conference on World Wide Web, ACM, Seoul, Korea 2014: 225–236. doi: 10.1145/2566486.2568010.

[6] ZHU X, GHAHRAMANI Z, and LAFFERTY J. Semi-supervised learning using Gaussian fields and harmonic functions[C]. International Conference on Machine Learning, Washington D.C., US, 2003, 3: 912–919.

[7] RAGHAVAN U N, ALBERT R, and KUMARA S. Near linear time algorithm to detect community structures in large-scale networks[J]. *Physical Review E*, 2007, 76(3): 036106. doi: 10.1103/PhysRevE.76.036106.

[8] PONS P and LATAPY M. Computing communities in large networks using random walks[C]. International Symposium on Computer and Information Sciences. Springer Berlin Heidelberg, Krakow, Poland, 2005: 284–293. doi: 10.1007/11569596_31.

[9] ROSVALL M and BERGSTROM C T. Maps of random walks on complex networks reveal community structure[J]. *Proceedings of the National Academy of Sciences*, 2008, 105(4): 1118–1123. doi: 10.1073/pnas.0706851105.

[10] LANCICHINETTI A and FORTUNATO S. Community detection algorithms: A comparative analysis[J]. *Physical Review E*, 2009, 80(5): 056117. doi: 10.1103/PhysRevE.80.056117.

[11] PALLA G, DERÉNYI I, FARKAS I, *et al*. Uncovering the overlapping community structure of complex networks in nature and society[J]. *Nature*, 2005, 435(7043): 814–818. doi: 10.1038/nature03607.

[12] AHN Y Y, BAGROW J P, and LEHMANN S. Link communities reveal multi scale complexity in networks[J]. *Nature*, 2010, 466(7307): 761–764. doi: 10.1038/nature09182.

[13] LANCICHINETTI A, RADICCHI F, RAMASCO J J, *et al*. Finding statistically significant communities in networks[J]. *PloS One*, 2011, 6(4): e18961. doi: 10.1371/journal.pone.0018961.

[14] YANG J and LESKOVEC J. Overlapping community detection at scale: A nonnegative matrix factorization approach[C]. The Sixth ACM International Conference on Web Search and Data Mining. ACM, Rome, Italy, 2013: 587–596. doi: 10.1145/2433396.2433471.

[15] NEWMAN M E J and CLAUSET A. Structure and inference in annotated networks[J]. *Nature Communications*, 2016,7: 11863. doi: 10.1038/ncomms11863.

[16] KOLLER D and FRIEDMAN N. Probabilistic Graphical Models: Principles and Techniques[M]. Massachusetts USA, MIT Press, 2009: 1–5.

[17] PRAT-PÉREZ A, DOMINGUEZ-SAL D, BRUNAT J M, *et al*. Shaping communities out of triangles[C]. The 21st ACM International Conference on Information and Knowledge Management, ACM, 2012: 1677–1681. doi: 10.1145/2396761.2398496.

[18] LEVORATO V and PETERMANN C. Detection of communities in directed networks based on strongly p-connected components[C]. IEEE 2011 International Conference on Computational Aspects of Social Networks (CASoN), Salamanca, Spain, 2011: 211–216. doi: 10.1109/CASON.2011.6085946.

[19] ARENAS A, DUCH J, FERNÁNDEZ A, *et al*. Size reduction of complex networks preserving modularity[J]. *New Journal of Physics*, 2007, 9(6): 1–14. doi: 10.1088/1367-2630/9/6/176.

邓小龙： 男，1977 年生，博士，讲师，研究方向为社交网络、数据挖掘.

翟佳羽： 男，1995 年生，硕士生，研究方向为社交网络社团发现.

尹栾玉： 女，1974 年生，教授，研究方向为公共服务、社会治理.