基于随机游走的社区发现方法综述

高阳[1,2]，张宏莉[2]
（1. 传播内容认知全国重点实验室，人民网，北京 100733；2. 哈尔滨工业大学网络空间安全学院，黑龙江 哈尔滨 150001）

摘 要：随机游走技术可实现准确、高效的社区发现。为总结分析基于随机游走的社区发现方法，将随机游走技术细分为个性化网页排名方法、热核扩散方法和其他随机游走方法，将社区发现问题分为局部社区发现和全局社区结构识别两类任务。详细综述了不同类型的随机游走技术及其在 2 种社区发现任务中的应用方式，并分析了现有方法存在的问题，对未来研究方向进行了展望。最后，针对不同社区发现任务从相似性标准与结构性标准两方面总结了社区发现准确性的评价指标，为相关研究提供便利。

关键词：局部社区发现；全局社区结构识别；随机游走；图扩散
中图分类号：TP391
文献标志码：A
DOI: 10.11959/j.issn.1000−436x.2023108

Survey on community detection method based on random walk

GAO Yang[1,2], ZHANG Hongli[2]
1. State Key Laboratory of Communication Content Cognition, People's Daily Online, Beijing 100733, China
2. School of Cyberspace Science, Harbin Institute of Technology, Harbin 150001, China

Abstract: Random walk techniques achieve high accuracy and efficiency in community detection. To summarize and analyze community detection methods based on random walk, the random walk technique was classified into personalized PageRank, heat kernel diffusion and other random walk methods, and community detection was classified into tasks of local community detection and global community structure identification. A detailed overview of different techniques based on random walk and their application to the tasks of community detection was provided, problems in existing methods were analyzed, and future research directions were pointed out. Finally, evaluation metrics of community detection accuracy for different community detection tasks were summarized in terms of similarity and structure respectively to facilitate research in community detection.

Keywords: local community detection, global community structure identification, random walk, graph diffusion

0 引言

复杂网络是现实世界中各种系统的数学抽象，与人类的生产生活密不可分，如通信网络、交通网络、社交网络、万维网络和生物网络等。虽然这些网络千差万别且属于不同的领域，却遵守着相同的规律，如小世界效应[1]、无标度性[2]、具有社区结构等[3-12]。社区常定义为边连接紧密的节点集合，体现了复杂网络的聚簇效应，例如，社交网络中用户自行组建的各种群体；商品网络中同一产品种类的商品集合等常被定义为网络的真实社区。同一个社区内的节点多具有相同的属性特征或在网络中扮演着相似的角色。

社区发现旨在识别一个网络的社区结构，是分析复杂网络拓扑结构与演化机制的基础方法，广泛应用于社会学、计算机科学、生物学和物理学等众多学科，已成为多学科交叉领域的研究热点之一。例如，在社交网络中，社区发现可用于链路预测，

收稿日期：2023−02−28；**修回日期：**2023−05−11
基金项目：传播内容认知全国重点实验室课题基金资助项目（No.A12003）
Foundation Item: State Key Laboratory of Communication Content Cognition, People's Daily Online (No.A12003)

实现精准的社交推荐；在蛋白质网络中，社区结构可用于预测陌生蛋白质的功能。

基于不同识别任务，社区发现可分为全局社区结构识别和局部社区发现。前者也称为图划分，旨在发现一个网络中的所有社区，依据社区间是否存在共有节点，又可分为重叠社区发现和非重叠社区发现；后者也称为社区搜索，旨在发现包含指定节点或节点集的个性化社区。鉴于局部社区发现常不依托网络的整体结构信息，时间需求较低，且对网络的结构信息缺失不敏感，因此更适应大规模网络。面向社区发现问题，现已形成众多理论方法，如标签传播（LP, label propagation）方法[7,13-14]、派系过滤方法（CPM, clique percolation method）[15]、非负矩阵分解（NMF, nonnegative matrix factorization）方法[16-17]、基于图神经网络（GNN, graph neural network）的社区发现方法[18-19]、基于特定子图结构的社区搜索[20]等。基于随机游走的社区发现方法由于其优秀的可扩展性与准确性，已成为当前研究的热点方法之一。本文对现阶段基于随机游走的社区发现方法及其存在问题进行总结，并对将来的研究方向进行展望。

1　相关概念

本文使用的主要符号及其含义如表 1 所示。

表 1　主要符号及其含义

符号	含义
$G(V, E)$	包含节点集 V 和边集 E 的网络
$\boldsymbol{A}$	网络的邻接矩阵
$\boldsymbol{D}$	节点度的对角矩阵
$\boldsymbol{P}$	随机游走的转移矩阵
$\boldsymbol{s}$	图扩散的起始向量
$\boldsymbol{h}$	图扩散向量
$\tilde{\boldsymbol{h}}$	近似图扩散向量
ID(C)	社区 C 的内部密度
AD(C)	社区 C 的平均度
TPR(C)	社区 C 的三角形参与率
Ex(C)	社区 C 的延展
CR(C)	社区 C 的切率
Co(C)	社区 C 的导率
$\boldsymbol{x}(\boldsymbol{s})$	相对于起始向量 $\boldsymbol{s}$ 的个性化网页排名向量
$\tilde{\boldsymbol{x}}(\boldsymbol{s})$	相对于起始向量 $\boldsymbol{s}$ 的近似个性化网页排名向量
$\boldsymbol{h}(\boldsymbol{s})$	相对于起始向量 $\boldsymbol{s}$ 的热核向量

1.1　社区发现

给定网络 $G(V,E)$，V 表示网络的节点集，E 表示边集，n 表示节点数量，即 $n = |V|$，m 表示边的数量，即 $m = |E|$。多种社区发现任务定义如下。

定义 1（非重叠社区发现）[21]　给定网络 G，非重叠社区发现任务是将 G 分割为 k 个相互独立的社区 $C_1, \cdots, C_k$，使 $C_1 \cup \cdots \cup C_k = V$，且对于 $\forall i, j \in [1, k]$，$C_i \cap C_j = \varnothing$。

定义 2（重叠社区发现）[21]　给定网络 G，重叠社区发现任务是识别网络中的所有社区 $C_1, \cdots, C_k$，使 $C_1 \cup \cdots \cup C_k \subseteq V$，且 $\exists i, j \in [1, k]$，使 $C_i \cap C_j \neq \varnothing$。

定义 3（局部社区发现）　给定网络 G 和目标社区 C 中的节点集 S ($S \subset C$)，局部社区发现旨在识别社区 C 中的剩余节点。

在局部社区发现问题中，给定节点集 S 常被称为目标社区的种子，可包含单一节点或多个节点。

1.2　随机游走技术

本文聚焦基于随机游走的社区发现方法，下面介绍其中的一些基本概念。本文用 $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ 表示网络的邻接矩阵，即当 $(i, j) \in E$ 时，$\boldsymbol{A}_{ij} > 0$，否则 $\boldsymbol{A}_{ij} = 0$。用 d_i 表示节点 i 的度，$\boldsymbol{D} \in \mathbb{R}^{n \times n}$ 表示节点度的对角矩阵，即 $\boldsymbol{D}_{ii} = d_i$，对于 $\forall i \neq j, \boldsymbol{D}_{ij} = 0$。用 $\boldsymbol{P} \in \mathbb{R}^{n \times n}$ 表示随机游走的转移矩阵，通常被定义为 $\boldsymbol{P} = \boldsymbol{D}^{-1}\boldsymbol{A}$，也被定义为 $\boldsymbol{P} = \frac{1}{2}(\boldsymbol{I} + \boldsymbol{D}^{-1}\boldsymbol{A})$ [22]，基于 $\boldsymbol{P} = \frac{1}{2}(\boldsymbol{I} + \boldsymbol{D}^{-1}\boldsymbol{A})$的随机游走称为惰性随机游走。

定义 4（图扩散）[23]　给定网络的转移矩阵 $\boldsymbol{P}$，图扩散可表示为

$$\boldsymbol{h} = \sum_{k=0}^{\infty} \alpha_k \boldsymbol{s}\boldsymbol{P}^k \tag{1}$$

其中，$\sum_k \alpha_k = 1$，$\boldsymbol{s} \in \mathbb{R}^n$ 是一个随机向量（即向量中的所有元素非负，且和为 1）。

可以看出，图扩散可有效捕获节点信息在网络中的传播情况。其中，系数 α_k 通常单调递减，以保证矩阵级数收敛。若应用在局部社区发现任务中，起始向量 $\boldsymbol{s}$ 常依据种子集 S 设定，如若 $i \in S$，则 $\boldsymbol{s}_i = \frac{1}{|S|}$，否则 $\boldsymbol{s}_i = 0$。显而易见，面向大规模网络的图扩散向量难以精确计算，而近似的图扩散向量也常可有效捕获目标社区的结构。给定从种子集 S 出发的近似图扩散向量 $\tilde{\boldsymbol{h}}$，局部社区发现任务可通过以下步骤完成：

1) 对在向量 $\tilde{\boldsymbol{h}}$ 中取值非零的节点 i 按照 $\frac{\tilde{\boldsymbol{h}}_i}{d_i}$ 降序排

列，用 $\boldsymbol{q}$ 表示得到的节点序列；2) 对于任意 k，计算 $\boldsymbol{q}$ 中前 k 个节点的社区质量函数，函数最优值对应的前 k 个节点即发现的目标社区。以上流程被称为清扫技术[22]。对于全局社区结构识别问题，首先基于网络结构选择每个社区的种子，然后通过以上局部社区发现任务的步骤将每个种子扩展为一个社区，即可得到网络的全局社区结构。可以看出，基于图扩散技术的全局社区结构识别可形成自然的重叠社区结构。

1.3 社区指标

现有众多社区指标用于度量社区质量，为方便读者使用，本文对常见的社区指标进行介绍。

1) 内部密度

社区内部密度（ID, internal density）[24-25]指社区内边的真实数量占社区可容纳边数量的比例，社区 C 的内部密度表示为

$$\mathrm{ID}(C)=\frac{e(C)}{\frac{|C|(|C|-1)}{2}} \tag{2}$$

其中，$e(C)$表示社区内边的数量，$|C|$表示社区内节点的数量。

2) 平均度

平均度（AD, average degree）[24-25]指社区中节点的平均内部度，这里节点内部度表示与该节点相连的社区内部节点的数量。社区 C 的平均度表示为

$$\mathrm{AD}(C)=\frac{2e(C)}{|C|} \tag{3}$$

3) 三角形参与率

三角形参与率（TPR, triangle participation ratio）[25]指社区内节点对社区内三角形结构的从属比例，这里三角形结构表示由 3 个节点和 3 条边构成的网络子图。社区 C 的三角形参与率表示为

$$\mathrm{TPR}(C)=|x:x\in C,\{y,z\in C,(x,y)\in E,(x,z)\in E,(y,z)\in E\}|\cdot\frac{1}{|C|} \tag{4}$$

4) 延展

社区延展（Ex, expansion）[24-25]指社区内节点的平均外部度，这里节点外部度指与该节点相连的社区外部节点的数量。社区 C 的延展表示为

$$\mathrm{Ex}(C)=\frac{\mathrm{vol}(C)-2e(C)}{|C|} \tag{5}$$

其中，vol(C)表示社区 C 内节点度的总和，常被称为 C 的体积。

5) 切率

社区切率（CR, cut ratio）[25]指社区与外部连边数量占所有可能连边数量的比例，社区 C 的切率表示为

$$\mathrm{CR}(C)=\frac{\mathrm{vol}(C)-2e(C)}{|C|(n-|C|)} \tag{6}$$

6) 导率

社区导率（Co, conductance）指社区与外部连边数量和社区体积与社区外节点集体积中较小值的比例。社区 C 的导率表示为

$$\mathrm{Co}(C)=\frac{\mathrm{vol}(C)-2e(C)}{\min(\mathrm{vol}(C),2m-\mathrm{vol}(C))} \tag{7}$$

在上述社区指标中，前 3 个侧重社区内部节点间的关联关系，较大的指标值对应较高的社区质量；后 3 个侧重社区与外部结构间的关联关系，较大的指标值表示较低的社区质量。在图 1 所示的网络中，左侧社区 C_1 的 6 种指标值分别为 $\mathrm{ID}(C_1)=\frac{5}{6}$，$\mathrm{AD}(C_1)=\frac{5}{2}$，$\mathrm{TPR}(C_1)=1$，$\mathrm{Ex}(C_1)=\frac{1}{4}$，$\mathrm{CR}(C_1)=\frac{1}{16}$，$\mathrm{Co}(C_1)=\frac{1}{9}$；右侧社区 C_2 的 6 种指标值分别为 $\mathrm{ID}(C_2)=\frac{2}{3}$，$\mathrm{AD}(C_2)=2$，$\mathrm{TPR}(C_2)=0$，$\mathrm{Ex}(C_2)=\frac{1}{4}$，$\mathrm{CR}(C_2)=\frac{1}{16}$，$\mathrm{Co}(C_2)=\frac{1}{9}$。

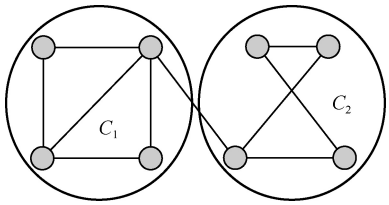


图 1 包含 2 个社区的简单网络

1.4 常见的随机游走方法

在基于随机游走的社区发现框架下，个性化网页排名（PPR, personalized PageRank）[26-27]和热核扩散[23,28]是最常用的方法，其中，PPR 定义如下。

定义 5（个性化网页排名） 给定转移矩阵 $\boldsymbol{P}$，随机向量 $\boldsymbol{s}$ 和参数 $\alpha\in(0,1]$，以下线性系统被称为个性化网页排名问题，线性系统的解 $\boldsymbol{x}(\boldsymbol{s})\in\mathbb{R}^n$ 被称为个性化网页排名向量。

$$\boldsymbol{x}(\boldsymbol{s})=\alpha\boldsymbol{s}+(1-\alpha)\boldsymbol{x}(\boldsymbol{s})\boldsymbol{P} \tag{8}$$

其中，α为转移概率，$\boldsymbol{s}$ 为起始向量，在社区发现问题中，$\boldsymbol{s}$ 依据种子设定。该方程的解可准确地表示

为 $\boldsymbol{x}(\boldsymbol{s})=\alpha\sum_{k=0}^{\infty}(1-\alpha)^k\boldsymbol{s}\boldsymbol{P}^k$，为指数系数的图扩散向量，即$\alpha_k=\alpha(1-\alpha)^k$，该表示形式蕴含了个性化网页排名向量基于随机游走的定义。方便起见，假设种子集 S 包含单一节点 s，定义$(\boldsymbol{l}_0, \boldsymbol{l}_1, \cdots, \boldsymbol{l}_T)$为起始于 s 的随机游走，即 $\boldsymbol{l}_0=s$，随机游走长度服从关于α的几何分布，即 $P(T=t)=(1-\alpha)^t\alpha$。该随机游走每步以$\alpha$的概率停止在当前节点，以 $1-\alpha$的概率游走到当前节点的一个随机邻节点。若给定网络为无权重网络，每个邻节点被选择的概率相同，即 $\frac{1}{d_u}$，其中，u 为随机游走当前所在节点；若给定网络为有权重网络，则依据边的权重选择邻节点。个性化网页排名向量 $\boldsymbol{x}$ 在任何节点 u 处的值即该随机游走最终停留在节点 u 的概率，即 $\boldsymbol{x}(\boldsymbol{s})_u=P(\boldsymbol{l}_T=u)$。

幂迭代[26]是计算近似个性化网页排名向量的传统方法，该方法将网页排名向量初始化为一个随机向量，如 $\boldsymbol{x}(\boldsymbol{s})^0=\left(\frac{1}{n},\cdots,\frac{1}{n}\right)$，并反复迭代式(8)，即 $\boldsymbol{x}(\boldsymbol{s})^{t+1}=\alpha\boldsymbol{s}+(1-\alpha)\boldsymbol{x}(\boldsymbol{s})^t\boldsymbol{P}$，结果将收敛于准确解，可选择 $\left\|\boldsymbol{x}(\boldsymbol{s})^{t+1}-\boldsymbol{x}(\boldsymbol{s})^t\right\|_\infty$ 小于设定阈值作为迭代的终止条件。幂迭代方法的时间需求与网络中边的数量呈线性关系，难以在大规模网络中多次执行，尤其在基于个性化网页排名的全局社区结构识别任务中，对每个种子执行幂迭代将产生过高的硬件需求。

为此，文献[22]提出一种不依托网络全局社区结构信息的局部方法 APPR （approximate personalized PageRank）计算近似的个性化网页排名向量。该方法维护一对向量 $\left(\tilde{\boldsymbol{x}}(\boldsymbol{s}),\boldsymbol{r}\right)$，使 $\tilde{\boldsymbol{x}}(\boldsymbol{s})=\boldsymbol{x}(\boldsymbol{s}-\boldsymbol{r})$，其中 $\tilde{\boldsymbol{x}}(\boldsymbol{s})\in\mathbb{R}^n$ 为相对于起始向量 $\boldsymbol{s}$ 的近似 PPR 向量，$\boldsymbol{r}\in\mathbb{R}^n$ 为残留向量。APPR 将 $\left(\tilde{\boldsymbol{x}}(\boldsymbol{s}),\boldsymbol{r}\right)$ 初始化为$(\boldsymbol{0},\boldsymbol{s})$，其中，$\boldsymbol{0}\in\mathbb{R}^n$ 为零向量，并通过一系列 push 操作不断缩小残留向量 $\boldsymbol{r}$，$\tilde{\boldsymbol{x}}(\boldsymbol{s})$ 将最终收敛于 $\boldsymbol{x}(\boldsymbol{s})$。该方法的时间需求与网络规模无关，与准确性成反比。

鉴于现有 PPR 向量计算方法多面向单机，在分布式框架下效率较低，文献[29]提出了 PPR 的大规模并行算法，设计了并行 push 算法，并与蒙特卡罗法结合，有效提高了并行计算效率。

定义 6（热核扩散）　给定转移矩阵 $\boldsymbol{P}$ 和参数 t，热核向量定义为

$$\boldsymbol{h}(\boldsymbol{s})=\mathrm{e}^{-t}\sum_{k=0}^{\infty}\frac{t^k}{k!}\boldsymbol{s}\boldsymbol{P}^k \tag{9}$$

相对于 PPR，热核扩散使用热核系数 $\frac{t^k}{k!}$ 代替了指数系数，对不同长度的随机游走设定了不同的权重。显而易见，热核系数衰减更快，因此为短步长的随机游走设置了更高的权重。实验结果[23]表明热核扩散可更好地捕获大规模网络上的个性化社区。

2　基于随机游走的局部社区发现

局部社区发现旨在识别网络指定区域，包含指定节点/节点集的个性化社区，相关方法常具有良好的可扩展性，适应真实网络的节点海量性，且可辅助完成众多下游任务。本文将局部社区发现分为基于网络原始拓扑的局部社区发现和高阶局部社区发现。

2.1　基于网络原始拓扑的局部社区发现

传统的局部社区发现方法依托网络的原始拓扑结构抽取目标社区，基于随机游走技术的相关研究主要采用个性化网页排名和热核扩散等图扩散方法。

1) 基于个性化网页排名的局部社区发现

该类方法借助网络中节点的 PPR 分数描述节点与相应种子的密切程度，从而高效捕获目标社区。文献[22]提出的 APPR 面向单一种子的时间需求与网络规模无关，与向量准确性成反比。该方法使用度标准化的方式对 APPR 向量中的节点排序，采用一种清扫技术对节点序列进行切割，使切割点前的节点集具有最优导率，该节点集即发现的目标社区。

文献[30]将 PPR 与邻节点计数[31]、贪婪结构优化[32-33]等其他局部社区发现方法进行了比较分析，采用幂迭代方法计算近似 PPR 向量，基于 PPR 分数对节点排序，并在序列中抽取特定数量节点作为发现社区。实验结果表明 PPR 显著优于其他方法，相对于度标准化的排序方式，使用 PPR 向量直接排序可得到更好的社区结构。

上述基于标准 PPR 的局部社区发现方法具有良好的可扩展性，然而单一随机游走捕获的社区常以种子为中心，若给定种子不在目标社区中心位置，识别结果可能出现较大误差；且随机游走可自由跨越社区边界，缺乏限制机制。面对以上问题，文献[34]提出一种多重随机游走模型 MWC（multi-walker chain），基于随机游走的高频访问节点不断改变式(8)中的起始向量，使随机游走过程难以逃离目标社区。MWC

采用幂迭代方法更新随机游走向量，最终采用清扫技术抽取目标社区。文献[35]基于 PageRank 算法识别种子所在社区的中心节点，并构建社区的核心区域，最终对核心区域扩展获取目标社区。

文献[36]提出了基于二阶随机游走的个性化网页排名方法 PP^2（second-order personalized PageRank）。在该方法中，随机游走下一步的移动方向由当前所在节点与上一步所在节点共同决定，从而限制随机游走逃离目标社区。原始的 PPR 采用节点到节点的转移矩阵，PP^2 则定义了边到边的转移矩阵，进而给出了 PP^2 的形式化定义，并设计了近似计算 PP^2 向量的幂迭代方法与蒙特卡罗方法。

传统的局部社区发现问题给定单个社区中的单一节点或多个节点，以发现社区中的剩余节点为目标。而实际问题常给定多个种子节点，这些节点可能属于同一真实社区，也可能分布在多个未知社区内，传统方法无法处理。为解决以上问题，文献[37]提出了一种基于记忆的随机游走（MRW, memory-based random walk）模型，为每个给定节点分配一个随机游走，并记录每个随机游走的节点访问历史。对于任一随机游走，MRW 基于节点访问历史不断更新式(8)中的起始向量，使随机游走下一步的移动方向不仅由当前所在节点决定，而由被访问过的节点共同判定。对于不同随机游走，MRW 基于访问节点的相似情况判定 2 个初始节点是否在相同真实社区内，并对相似随机游走进行相互限制，从而提高局部社区发现的鲁棒性。

文献[38]对相似问题进行了研究，但假定已知给定节点是否属于相同社区，提出了着色随机游走（CRW, colored random walk）模型。CRW 为分布在不同社区内的给定节点分配不同颜色，对每种颜色分别执行随机游走，通过对式(8)中的转移矩阵不断更新，使随机游走更倾向于向相同颜色节点移动，并阻碍其向不同颜色节点移动，从而借助不同颜色随机游走的相互限制，准确获取不同社区。

鉴于单层网络常存在噪声与信息缺失，文献[39]提出了面向多层网络的随机游走（RWM, random walk in multiple network）模型。RWM 对每层网络分别执行随机游走，定义了网络间的相关矩阵并融入转移矩阵中，使不同随机游走形成关联，从而提高局部社区发现的准确性。RWM 采用幂迭代方法计算随机游走向量，并给出 2 种模型的加速方案。

现有局部社区发现方法多假定给定种子节点属于单一真实社区，并以搜索该社区为目标。而真实网络中的社区间存在大量重叠，给定节点常位于多个社区的重叠区域，因此同时属于多个社区。为此，文献[40]提出了一种多重局部社区发现方法 HqsMLCD（high-quality seed based multiple local community detection），面向单一给定节点，可识别多个相关社区。HqsMLCD 首先基于图表示学习获取与给定节点相关的多个种子，然后借助 PPR 技术对每个种子进行扩展，从而获取多个目标社区。

2) 基于热核扩散的局部社区发现

该类方法采用始于种子的热核向量描述网络中的节点与种子的密切程度，基于热核分数对节点排序，采用清扫技术抽取目标社区，方法的核心在于热核向量的近似计算。热核扩散同样基于随机游走技术，与个性化网页排名技术相比，热核扩散更加注重短步长的随机游走，采用热核系数替代 PPR 中的指数系数。文献[23]是采用热核扩散识别局部社区的代表性工作，基于高斯−赛德尔迭代法的思想提出了一种近似计算热核向量的松弛法 hk-relax，并采用导率作为目标函数抽取社区结果。相对于 PPR，热核扩散在小规模网络上运行更快，在大规模网络上时间需求更高；在真实网络中，热核扩散得到的社区包含相对较少的节点，准确性较高，但是导率结果较差。

计算热核向量可直接采用蒙特卡罗方法，即模拟大量始于种子节点且长度服从泊松分布的随机游走，终止于某一节点的随机游走在总量中所占比例即热核向量在该节点的近似值。该方法的精度与随机游走的数量正相关，满足需求的准确性常对应过高的时间需求。为提高蒙特卡罗方法的效率，文献[28]设计了一种确定性图遍历方法 HK-Push，与蒙特卡罗方法结合可降低所需随机游走的数量。首先利用 HK-Push 方法计算热核向量的粗略近似值，然后基于蒙特卡罗随机游走构建了 *k*-RandomWalk 算法对结果进一步优化，进而提出了 2 种基于热核扩散的局部社区发现方法 TEA（two-phase heat kernel approximation）和 TEA+。相对于 hk-relax，TEA+在运行效率方面提升 4 倍以上。

文献[41]提出了一种基于子图抽样的热核向

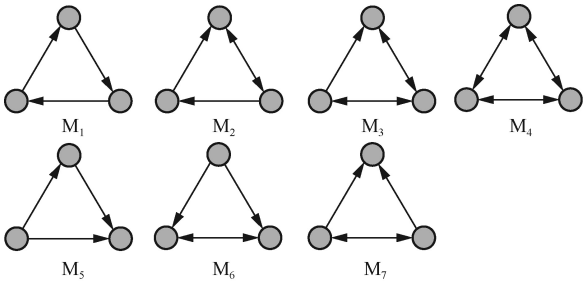
量计算方法。给定种子节点 s，该方法采用以下贪婪算法抽取网络子图：首先将子图初始化为节点 s 及其全部邻节点组成的集合；然后逐渐向子图中加入节点，试图让子图导率值减少最多或增加最少，直到子图大小达到预期值为止；最后基于抽取子图上的随机游走计算热核向量的近似值，并将导率作为目标函数获取社区结果。在人工模拟网络上的实验结果表明，该方法识别社区的导率优于 TEA+。

2.2　高阶局部社区发现

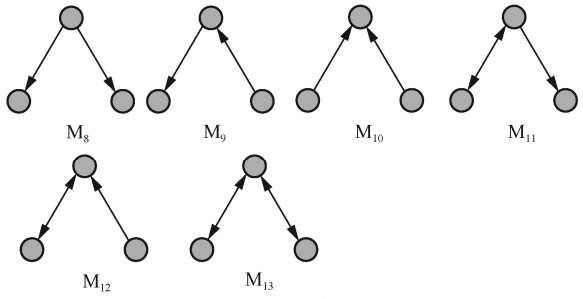
本节从网络高阶结构的定义和高阶局部社区发现方法等两方面进行总结分析。

1) 网络高阶结构的定义

真实网络中常存在大量特定结构的子图，如三角形、正方形等，这些子图被称为网络模体[42-43]。相对于随机网络，相同规模的真实网络往往包含更多这类子图[44]，因此网络模体被认为是复杂网络的基本构成要素。例如，三角形模体（M_1～M_7）对社交网络十分关键，两跳路径（M_8～M_{13}）是理解空中交通模式的重要结构，如图 2[43]所示。



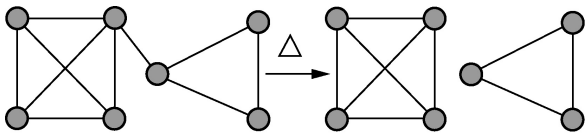
(a) 三角形模体



(b) 两跳路径

图 2　有向图中的三节点模体

在社区发现过程中，有效利用网络模体捕获的结构信息可提高识别准确性。现有方法常基于模体构建网络超图，直接使用传统社区发现方法或扩展传统方法在超图中识别社区。针对高阶社区发现中存在的问题，现有多种超图的构建方式。文献[42]将原始网络的所有节点作为超图的节点集，若 2 个节点分布在同一个模体中，则认为超图中存在一条边，并将同时包含 2 个节点的模体数量作为边的权重。选择三角形作为网络模体，图 3 给出了一个简单网络的拓扑结构和对应的超图结构。其中，△表示选择三角形作为网络模体。从图 3 可以看到，超图能够展示网络更加清晰的社区结构。



(a) 简单网络的拓扑结构　　(b) 超图结构

图 3　一个简单网络的拓扑结构和对应的超图结构

然而真实网络常具有稀疏的网络结构，上述超图将移除大量网络原始边，从而出现超图碎片化[45]问题。为此，文献[45]提出了一种基于边强化的超图构建方式 EdMot（edge enhancement approach for motif-aware community detection）。EdMot 首先基于网络模体构建初始超图，利用一种全局社区发现方法将超图中较大的连通子图分割为若干模块；然后对每个模块进行边填充，使其成为完全子图；最后将新加入的边与网络原始边集合并得到最终超图的边集，超图节点集与原始网络相同。

以上超图保留了原始网络拓扑，有效解决了超图碎片化问题，然而添加大量额外边可能破坏原始网络结构。为此，文献[44]提出了一种微型单元连接网络（MCN, micro-unit connection network），将网络模体或不在任何模体内的边定义为网络微型单元，基于杰卡德相似度定义微型单元之间的关联关系，从而建立以微型单元为节点，以微型单元间的连接关系为边的超图。MCN 充分利用了网络的原始结构，并引入了高阶网络信息，可有效提高社区发现的准确性。然而在稠密的网络中，微型单元数量将远大于节点数量，严重影响社区发现效率。因此，该方法的效率提升可作为将来的研究内容之一。

2) 高阶局部社区发现方法

面向社区发现问题定义的网络超图大多是一个有权重或无权重的简单图，因此大多传统局部社区发现方法均可（直接或经适当调整）从超图中抽取目标社区。本文主要介绍一些基于随机游走的代表

性工作。

文献[42]提出了基于个性化网页排名的高阶局部社区发现方法 MAPPR（motif-based approximate personalized PageRank）。给定种子，MAPPR 采用以下流程识别目标社区：首先基于网络模体定义有权重超图，利用同时包含 2 个节点模体的数量定义节点间的连接强度；然后将 PPR 向量的局部计算方法 APPR 扩展到有权重网络中，以计算面向超图的 MAPPR 向量，时间需求与 APPR 方法类似；最终同样采用清扫技术识别目标社区，在此过程中，MAPPR 在传统社区导率中引入了网络高阶信息，定义了模体导率，并以此为目标函数进行社区抽取。在该方法中，构建超图依托网络的全局结构信息，时间复杂度与网络规模和所选择的模体规模相关，因此在大规模网络中时间需求较高。然而对于同一网络，相应超图仅需构建一次，即可反复执行后续步骤来识别多个目标社区。相对于 APPR，MAPPR 在社区发现的准确性方面有显著提升。

文献[46]提出了高阶局部社区发现方法 HOSPLOC（high-order structure-preserving local cut），旨在识别包含丰富高阶结构的网络子图，且从网络中分离该子图不破坏大量高阶结构。HOSPLOC 的核心在于近似计算基于高阶网络结构的高阶随机游走分布，同时定义了高阶导率来提取目标社区。该方法基于图数据的张量表示进行计算，用户可指定目标社区需维护的高阶结构类型，可应用于多类型网络中，如符号网络、二分网络、多方网络等。

鉴于多数高阶局部社区发现方法面向静态网络，文献[47]提出了用于动态网络的局部社区发现方法 L-MEGA（local motif clustering on time-evolving graphs），根据网络的变化情况跟踪高阶随机游走的稳态分布，基于模体导率识别目标社区。L-MEGA 主要包含以下步骤：首先使用转移张量定义高阶随机游走过程的转移概率，并基于网络结构变化不断更新；然后获取随机游走在每个时间点的稳态概率分布；最终设计增量方法在稳态分布中提取目标社区。L-MEGA 虽然设计了多种加速方案，但仍有较高的时间需求，因此如何进一步提高方法的可扩展性可作为将来的研究内容之一。本文将基于随机游走的局部社区发现代表性方法总结为表 2。

表 2 基于随机游走的局部社区发现代表性方法

分类	方法	核心思想
基于 PPR	APPR[22]	PPR 向量的局部计算，清扫技术
	MWC[34]	动态起始向量，幂迭代方法
	PP²[36]	二阶个性化网页排名方法
	MRW[37]	面向多种子问题，动态起始向量
	CRW[38]	面向多种子问题，动态转移矩阵
	RWM[39]	面向多层网络的局部社区发现问题
基于热核扩散	hk-relax[23]	松弛法
	TEA[28]	确定性图遍历，蒙特卡罗随机游走
	文献[41]	子图抽样
高阶方法	MAPPR[42]	基于模体的个性化网页排名方法
	HOSPLOC[46]	近似计算高阶随机游走分布
	L-MEGA[47]	跟踪高阶随机游走的稳态分布

3 基于随机游走的全局社区结构识别

基于随机游走的全局社区结构识别多需借助局部社区发现技术，主要通过以下步骤完成：1) 根据网络拓扑结构选择每个社区的种子，每个种子可包含单一节点或多个关联紧密的节点；2) 采用局部社区发现方法将每个种子扩展成一个社区，形成网络的完整社区结构。采用以上步骤识别全局社区结构的方法常被称为局部扩展法，在步骤 2)中，每个社区的扩展过程一般相互独立，因此可形成自然的重叠社区结构，符合真实网络的结构特性。本文首先从种子识别与社区扩展两方面对基于局部扩展的全局社区结构识别进行总结分析，并列举基于随机游走的其他全局社区结构识别方法。

3.1 种子识别

在社区发现任务中，网络的真实社区结构未知，因此种子选择方法常出现众多问题，如选取的种子数量与真实社区数量差距过大，从而发现过多或过少的社区；同一个种子内的节点分布在多个真实社区中，导致一个发现社区包含多个真实社区；2 个种子属于同一真实社区，出现重复或高度相似的社区结果等。另外，现有研究结果表明局部社区发现的准确性与种子的选取关系密切，在社区中心位置的节点常可扩展为高质量社区，其他节点则常使社区扩展偏离目标社区[30,34]。因此，大多种子识别方法侧重获取社区的中心节点，下面介绍相关代表性方法。

1) 单节点种子

文献[48]提出了一种简单高效的种子选取方

法，每次选择不属于任何已发现社区的一个随机节点作为种子，并扩展得到相应社区，直到网络没有孤立节点（这里孤立节点指没有社区隶属关系的节点）为止。该方法时间需求极低，可保证不同种子扩展为不同社区，然而没有考虑种子的中心性特征，选取的种子常出现在真实社区的边界位置。

文献[49]将度较大的节点定义为网络的核心节点，每次选择度最大的节点作为一个种子，并将该节点及其所有邻节点移出网络，直到网络的节点集为空。文献[50]基于子图密度将无向无权重网络转化为有权重网络，并在有权重网络中定义了节点的权重，依据节点权重顺次选取种子。首先将种子集初始化为网络的节点集；然后选择其中权重最大的节点作为种子进行社区扩展；最后将该种子及扩展成的社区移出种子集，并重复上一步骤，直到种子集为空。文献[51]基于节点间的相似度指标定义了加权网络，进而定义了节点的加权度，依据节点的加权度选择社区的种子节点。文献[52]将给定网络进行剪枝并转化为边图，基于 PageRank 算法对边图中的节点排序，依据排序结果选择社区的种子节点。

上述方法依据节点的属性信息选取种子，可有效促使种子集分布在社区的中心位置。然而文献[16]指出社区间的重叠区域具有较高的密度，因此，其中包含的节点可能具有显著的中心特征，如较大的节点度，相对于随机选择种子，上述方法同样可出现负面效果。因此，如何识别社区中心位置的节点可作为未来研究内容之一。

2) 多节点种子

文献[21]提出了 2 种种子识别方法 Graclus Centers 和 Spread Hubs。Graclus Centers 采用一种非重叠社区发现方法将网络划分为若干节点集，在每个集合中选择最中心的节点作为一个种子。Spread Hubs 简单依据网络中节点度的大小选择种子，且保证种子间不相邻。2 种方法均需事先指定种子数量。在社区扩展过程中，该模型将每个种子与其全部邻节点组成的集合作为随机游走的起点。因此，2 种方法实际上得到的是每个种子的中心节点，相应的节点集为真实种子。

在上述方法中，每个种子包含多个节点，这些节点可能分布在不同真实社区内，严重影响社区发现的准确性。鉴于社区中常存在稠密区域，文献[53]基于子图边密度对种子密度进行了限制。文献[54]基于边密度与三角形密度定义了子图的增强密度，并为该密度设定单调递增的动态阈值，使较大的种子倾向于拥有较高的增强密度。文献[55]定义了一种强连接三角形 *k*-triangle, 通过在网络中搜索 *k*-triangle 获取种子集。

3.2　社区扩展

给定网络的种子集合，可通过对每个种子的扩展获取网络的社区结构。现有的局部社区发现方法大多可直接应用于社区扩展过程，针对全局社区发现问题的特点，也出现了众多不同技术。本文将社区扩展方法分为基于网络原始拓扑的社区扩展方法和基于高阶网络信息的社区扩展方法，并重点对基于随机游走的代表性工作进行总结分析。

1) 基于网络原始拓扑的社区扩展方法

文献[21]提出了基于个性化网页排名技术的社区发现方法 NISE （neighborhood-inflated seed expansion），采用以下流程识别网络的社区结构：首先将网络划分为一个重连通核心和多个小分支；然后在重连通核心上选择种子，并执行 APPR 对每个种子进行扩展；最后将划分的小分支加入已发现的社区中。相对于 bigclam[16]和 DEMON[13]等其他类别的全局社区发现方法，NISE 具有较好的运行效率与准确性。

文献[56]基于同样的扩展方法提出了社区发现方法 LECM （local expansion and conductance minimizing），并引入了以下社区优化机制：节点移动，通过改变社区内部节点和与社区直接相连节点的社区隶属关系提高发现社区质量；社区合并，鉴于社区扩展可能得到相同或高度相似的社区，LECM 对相似度过高的社区进行合并，在合并过程中综合考虑社区间的重叠程度与社区导率的变化，使合并后的社区结构倾向于拥有更高质量；为孤立节点选择社区，由于在节点移动过程中可形成较多孤立节点，LECM 最后将孤立节点加入已发现社区中。实验结果表明，上述社区优化机制可有效提高发现社区结构与真实社区的相似性，并降低社区结构的平均导率。

在基于局部扩展的全局社区结构识别方法中，每个种子的扩展存在先后顺序，先形成的社区结构可为后续的社区扩展提供丰富的网络结构信息。为此，文献[57]提出 CPPR（constrained PPR）方法在 PPR 方程的转移矩阵中引入了强化矩阵，阻碍随机

游走向已发现社区内部移动，鼓励随机游走走出已发现社区，从而避免扩展成相同或高度相似的社区，并允许社区间发生重叠。

2) 基于高阶网络信息的社区扩展方法

高阶网络信息在全局社区结构识别问题中同样产生了较好效果。文献[54]提出了基于个性化网页排名技术的社区发现方法 BTLCD（biased triangle enhanced local community detection），借助社区结构定义了多种三角形，以刻画社区内部的强连接关系、社区和外部结构间的强弱连接关系，进而提出了基于三角形结构的社区导率，用以提取目标社区。与 APPR 方法不同，BTLCD 没有采用清扫技术，而使用一种贪婪方法获取目标社区。实验结果表明，高阶信息的引入有效提高了社区识别的准确性。

基于模体的网络超图常存在结构信息缺失问题，在稀疏网络中表现尤为明显。在局部社区发现任务中，给定种子可能出现在网络的稠密区域，该问题将不影响社区识别的准确性。然而全局社区结构识别任务面向整个网络，结构信息缺失将不可避免地降低发现社区结构的质量。为此，文献[55]定义了基于三角形结构的混合超图，并将个性化网页排名技术扩展到混合超图中，提出了社区发现方法 BPPR （biased personalized PageRank）。BPPR 同时对社区的导率指标进行了扩展，定义了混合导率来抽取每个社区。该方法有效解决了传统超图结构信息缺失问题。

3.3 基于随机游走的其他方法

1) 基于网络拓扑的全局社区发现

传统聚类方法多基于一跳邻节点定义节点间的相似性，因此对网络结构变化十分敏感，网络拓扑的微小变化可导致社区发现结果的剧烈变化。针对该问题，文献[58]借助个性化网页排名技术定义了网络中 2 个节点间的距离和 2 个社区间的距离，不仅利用了节点间的直接关联，且可有效捕获长距离相似性，进而设计了基于聚类的全局社区发现方法 C_PPR。文献[59]采用 SimRank[60]计算节点间的距离，利用 PageRank 计算节点的重要度，进而构造树图识别网络的层次化社区结构。

起始于社区内某个节点的随机游走在初始阶段倾向于在社区内部移动，因此从同一社区 2 个不同节点出发的随机游走常经过相似的节点集。基于上述观点，文献[61]提出了一种基于受限随机游走相似性（RR, restrained random-walk similarity）的社区发现方法，将访问相似节点集随机游走的起始节点划分到同一社区中。RR 设计了 2 种强化机制。首先，有些随机游走很快离开起始节点所在社区，这种随机游走被称为反常随机游走，RR 将起始于每个节点的随机游走执行多次，被访问次数较少的节点将从可被访问节点集中移除。其次，RR 将随机游走过程分为 3 个阶段：第一阶段，随机游走常频繁访问陌生节点；第二阶段，随机游走重复访问已经访问过的节点，在前 2 个阶段，随机游走多出现在初始节点所在社区；第三阶段，随机游走将逃离该社区。显而易见，第二阶段是停止随机游走的最佳时机，为此，RR 在随机游走访问节点集增长缓慢时停止随机游走。

2) 基于图神经网络的全局社区发现

真实网络除了具有拓扑结构外，节点中还常包含丰富的属性信息，如在社交网络中，用户可具有性别、年龄、爱好等众多标签，这类网络常被称为属性网络。面向属性网络的图分割任务一般被称为节点分类，本文也将其称为全局社区发现问题。图神经网络（GNN, graph neural network）可综合网络拓扑与节点属性实现属性网络的社区发现。近年来，随机游走技术在其中得到了有效利用，现对其中的代表性工作进行介绍。

鉴于图卷积网络（GCN, graph convolutional network）的信息传播范围与网络层数相关，而过多层数将导致过平滑问题。为此，文献[62]将个性化网页排名技术引入图神经网络中，提出了 PPNP（personalized propagation of neural prediction）模型。首先，PPNP 模型基于节点的自身特征生成每个节点的标签预测，该步骤使用神经网络完成；然后，借助 PPR 技术对标签进行传播得到最终预测。PPNP 模型的主要优势在于将神经网络与传播方案分离，在不改变神经网络的前提下实现大范围的预测传播。然而 PPNP 在训练过程中使用幂迭代方法计算 PPR 矩阵，时间需求过高，难以应用在大规模网络中。为此，文献[63]使用 push 操作[22]计算相对于每个节点的 PPR 向量，并仅保留向量中最大的 k 个值，使 PPR 矩阵稀疏化。相对于 PPNP，该方法的主要优势是可在训练前预先计算稀疏化的 PPR 矩阵，使训练和推理在 $O(k)$的时间复杂度上完成。本文将基于随机游走的全局社区结构识别代表性方法总结为表 3。

表 3　基于随机游走的全局社区结构识别代表性方法

分类	方法	核心思想
局部扩展	NISE[21]	网络修剪，计算近似 PPR 向量
	LECM[56]	计算近似 PPR 向量，社区优化
	CPPR[57]	在转移矩阵中引入强化矩阵
	BTLCD[54]	定义多类型高阶单元
	BPPR[55]	定义混合超图
其他方法	C_PPR[58]	捕获节点间的长距离相似性
	RR[61]	利用受限随机游走相似性
	PPNP[62]	将神经网络与传播方案分离

4　社区发现评价标准

发现社区的质量常作为评价社区发现方法好坏的基本标准，可通过发现社区与真实社区间的相似性评价社区质量，也可通过社区结果自身的结构特性量化社区质量。

4.1　相似性标准

若目标网络提供真实社区结构，可通过对比发现社区（对应局部社区发现）/社区结构（对应全局社区结构识别）与对应真实社区/社区结构间的相似程度衡量社区发现质量。

文献[64]提出利用标准互信息（NMI, normalized mutual information）度量发现社区与真实社区结构间的相关性。鉴于真实网络中社区常发生重叠，文献[48]提出了面向重叠社区结构的标准互信息（ONMI, overlapping normalized mutual information）。文献[65]针对 ONMI 的非直观性表现，对该方法做了进一步改进。

基于标准互信息的相关方法多面向全局社区结构识别问题，文献[16]基于准确率和召回率的调和平均数定义了发现社区集与真实社区集间的平均 F1 值。该指标可应用于度量 2 个社区间的相似性，因此可应用于局部社区发现任务[66]。

4.2　结构性标准

很多真实网络没有真实社区结构或未提供真实社区结构，此时可通过发现社区的自身结构指标评价其质量。常见的社区指标多通过社区的结构特性定义，即社区内部关联紧密，社区与外部结构间关联稀疏。社区导率与模块度 [25,67]是最常用的 2 个指标。面向局部社区发现问题，也出现了一些其他指标，如三角形密度、三阶导率等[47]。

5　研究展望

面向社区发现任务，现已形成众多随机游走模型，且取得了优秀的准确性与效率。然而，仍存在需进一步研究或探索的问题。

1) 社区的中心节点识别。基于随机游走的社区发现方法常具有良好的扩展性，然而得到的目标社区常以种子为中心[34]。因此，在全局社区结构识别任务中，准确识别每个真实社区的中心节点是提高社区发现准确性的关键因素。然而，社区间的重叠区域同样关联紧密[16]，其中包含的节点与社区的中心节点难以区分。因此，如何准确捕获社区的中心节点可作为未来的研究方向之一。一方面，可基于网络结构进一步探索中心节点所具有的独有属性，如文献[68]利用节点的拓扑优势识别社区的核心节点；另一方面，鉴于真实网络中的节点常具有大量的标签信息，可借助该信息判定节点的核心性特征。

2) 全局社区发现中的热核扩散技术。热核扩散已应用于局部社区发现任务，且取得了较好的准确性，然而尚未用于识别网络的全局社区结构。因此，将热核扩散技术移植到基于局部扩展的全局社区发现方法中将是一项有意义的研究内容。例如，可针对全局社区发现任务的特征对热核向量的定义进行变换，更加准确地捕获网络的社区结构；也可将热核扩散技术引入网络的超图结构，构建基于热核扩散的高阶社区发现方法。

6　结束语

随着真实网络规模的不断增长，基于随机游走的社区发现得到了广泛的关注，正逐渐成为大规模网络社区结构识别的基本方法。本文从局部社区发现和全局社区结构识别 2 个方面对随机游走相关技术的基本思想、流程、创新之处和存在问题进行全面总结分析，并对社区发现评价标准进行了介绍，拟为相关领域的研究人员提供参考。

参考文献：

[1] WATTS D J, STROGATZ S H. Collective dynamics of small-world networks[J]. Nature, 1998, 393(6684): 440-442.

[2] BARABASI A L, ALBERT R. Emergence of scaling in random networks[J]. Science, 1999, 286(5439): 509-512.

[3] JIN D, YU Z Z, JIAO P F, et al. A survey of community detection approaches: from statistical modeling to deep learning[J]. IEEE

Transactions on Knowledge and Data Engineering, 2023, 35(2): 1149-1170.

[4] SU X, XUE S, LIU F Z, et al. A comprehensive survey on community detection with deep learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, doi.org: 10.1109/TNNLS.2021.3137396.

[5] HE D X, LIU H X, FENG Z Y, et al. A joint community detection model: integrating directed and undirected probabilistic graphical models via factor graph with attention mechanism[J]. IEEE Transactions on Big Data, 2022, 8(4): 994-1006.

[6] MAGNANI M, HANTEER O, INTERDONATO R, et al. Community detection in multiplex networks[J]. ACM Computing Surveys, 2021, 54(3): 1-35.

[7] ZAREZADEH M, NOURANI E, BOUYER A. DPNLP: distance based peripheral nodes label propagation algorithm for community detection in social networks[J]. World Wide Web, 2022, 25(1): 73-98.

[8] LIU A, MOITRA A. Minimax rates for robust community detection[C]//Proceedings of the 63rd Annual Symposium on Foundations of Computer Science (FOCS). Piscataway: IEEE Press, 2022: 823-831.

[9] STEPHAN L, ZHU Y Z. Sparse random hypergraphs: non-backtracking spectra and community detection[C]//Proceedings of the 63rd Annual Symposium on Foundations of Computer Science (FOCS). Piscataway: IEEE Press, 2022: 567-575.

[10] WU X X, XIONG Y, ZHANG Y, et al. CLARE: a semi-supervised community detection algorithm[C]//Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2022: 2059-2069.

[11] 史艳翠, 王嫄, 赵青, 等. 基于局部扩展的社区发现研究现状[J]. 通信学报, 2019, 40(1): 149-162.
SHI Y C, WANG Y, ZHAO Q, et al. Research status of community detection based on local expansion[J]. Journal on Communications, 2019, 40(1): 149-162.

[12] 刘强, 贾焰, 方滨兴, 等. 并行社区发现算法的可扩展性研究[J]. 通信学报, 2018, 39(4): 13-20.
LIU Q, JIA Y, FANG B X, et al. Research on the scalability of parallel community detection algorithms[J]. Journal on Communications, 2018, 39(4): 13-20.

[13] COSCIA M, ROSSETTI G, GIANNOTTI F, et al. DEMON: a local-first discovery method for overlapping communities[C]//Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2012: 615-623.

[14] LU M L, ZHANG Z L, QU Z H, et al. LPANNI: overlapping community detection using label propagation in large-scale complex networks[J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 31(9): 1736-1749.

[15] 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.

[16] 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 Press, 2013: 587-596.

[17] SU S X, GUAN J W, CHEN B L, et al. Nonnegative matrix factorization based on node centrality for community detection[J]. ACM Transactions on Knowledge Discovery from Data, 2022, 17: 1-21.

[18] HE D X, SONG Y, JIN D, et al. Community-centric graph convolutional network for unsupervised community detection[C]//Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization, 2021: 3515-3521.

[19] JIN D, LIU Z Y, LI W H, et al. Graph convolutional networks meet Markov random fields: semi-supervised community detection in attribute networks[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 152-159.

[20] LIU Q, ZHAO M J, HUANG X, et al. Truss-based community search over large directed graphs[C]//Proceedings of the ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2020: 2183-2197.

[21] WHANG J J, GLEICH D F, DHILLON I S. Overlapping community detection using neighborhood-inflated seed expansion[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(5): 1272-1284.

[22] ANDERSEN R, CHUNG F, LANG K. Local graph partitioning using PageRank vectors[C]//Proceedings of the 47th Annual Symposium on Foundations of Computer Science (FOCS). Piscataway: IEEE Press, 2006: 475-486.

[23] KLOSTER K, GLEICH D F. Heat kernel based community detection[C]//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2014: 1386-1395.

[24] 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.

[25] CHAKRABORTY T, DALMIA A, MUKHERJEE A, et al. Metrics for community analysis[J]. ACM Computing Surveys, 2018, 50(4): 1-37.

[26] PAGE L, BRIN S, MOTWANI R, et al. The PageRank citation ranking: bringing order to the Web[R]. 1999.

[27] LOFGREN P A, GARCIA-MOLINA H, GOEL A, et al. Efficient algorithms for personalized PageRank[D]. California: Stanford University, 2015.

[28] YANG R C, XIAO X K, WEI Z W, et al. Efficient estimation of heat kernel PageRank for local clustering[C]//Proceedings of the 2019 International Conference on Management of Data. New York: ACM Press, 2019: 1339-1356.

[29] HOU G H, CHEN X G, WANG S B, et al. Massively parallel algorithms for personalized PageRank[J]. Proceedings of the VLDB Endowment, 2021, 14(9): 1668-1680.

[30] KLOUMANN I M, KLEINBERG J M. Community membership identification from small seed sets[C]//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2014: 1366-1375.

[31] MEHLER A, SKIENA S. Expanding network communities from representative examples[J]. ACM Transactions on Knowledge Discovery from Data, 2009, 3(2): 1-27.

[32] CLAUSET A. Finding local community structure in networks[J]. Physical Review E, 2005, 72(2): 026132.

[33] MISLOVE A, VISWANATH B, GUMMADI K P, et al. You are who you know: inferring user profiles in online social networks[C]//Proceedings of the 3rd ACM International Conference on Web Search and Data Mining. New York: ACM Press, 2010: 251-260.

[34] BIAN Y C, NI J C, CHENG W, et al. Many heads are better than one: local community detection by the multi-walker chain[C]//Proceedings of 2017 IEEE International Conference on Data Mining (ICDM). Piscataway: IEEE Press, 2017: 21-30.

[35] JI P Y, GUO K, YU Z Y. Local community detection algorithm based on core area expansion[C]//Proceedings of Computer Supported Cooperative Work and Social Computing. Singapore: Springer Nature Singapore, 2022: 238-251.

[36] WU Y B, ZHANG X, BIAN Y C, et al. Second-order random walk-based proximity measures in graph analysis: formulations and algorithms[J]. The VLDB Journal, 2018, 27(1): 127-152.

[37] BIAN Y C, YAN Y W, CHENG W, et al. On multi-query local community detection[C]//Proceedings of IEEE International Conference on Data Mining (ICDM). Piscataway: IEEE Press, 2018: 9-18.

[38] YAN Y W, BIAN Y C, LUO D S, et al. Constrained local graph clustering by colored random walk[C]//Proceedings of The World Wide Web Conference. New York: ACM Press, 2019: 2137-2146.

[39] LUO D S, BIAN Y C, YAN Y W, et al. Local community detection in multiple networks[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM Press, 2020: 266-274.

[40] LIU J X, SHAO Y X, SU S. Multiple local community detection via high-quality seed identification over both static and dynamic networks[J]. Data Science and Engineering, 2021, 6(3): 249-264.

[41] LU Z Q, WAHLSTRÖM J, NEHORAI A. Local clustering via approximate heat kernel PageRank with subgraph sampling[J]. Scientific Reports, 2021, 11: 15786.

[42] YIN H, BENSON A R, LESKOVEC J, et al. Local higher-order graph clustering[C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2017: 555-564.

[43] BENSON A R, GLEICH D F, LESKOVEC J. Higher-order organization of complex networks[J]. arXiv Preprint, arXiv: 1612.08447, 2016.

[44] HUANG L, CHAO H Y, XIE Q. MuMod: a micro-unit connection approach for hybrid-order community detection[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(1): 107-114.

[45] LI P Z, HUANG L, WANG C D, et al. EdMot: an edge enhancement approach for motif-aware community detection[C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM Press, 2019: 479-487.

[46] ZHOU D W, ZHANG S, YILDIRIM M Y, et al. A local algorithm for structure-preserving graph cut[C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2017: 655-664.

[47] FU D Q, ZHOU D W, HE J R. Local motif clustering on time-evolving graphs[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM Press, 2020: 390-400.

[48] LANCICHINETTI A, FORTUNATO S, KERTÉSZ J. Detecting the overlapping and hierarchical community structure in complex networks[J]. New Journal of Physics, 2009: doi.org/ 10.1088/1367-2630/11/3/033015.

[49] 陈俊宇, 周刚, 南煜, 等. 一种半监督的局部扩展式重叠社区发现方法[J]. 计算机研究与发展, 2016, 53(6): 1376-1388.
CHEN J Y, ZHOU G, NAN Y, et al. Semi-supervised local expansion method for overlapping community detection[J]. Journal of Computer Research and Development, 2016, 53(6): 1376-1388.

[50] YANG J X, ZHANG X D. Finding overlapping communities using seed set[J]. Physica A: Statistical Mechanics and Its Applications, 2017, 467: 96-106.

[51] 杨贵, 郑文萍, 王文剑, 等. 一种加权稠密子图社区发现算法[J]. 软件学报, 2017, 28(11): 3103-3114.
YANG G, ZHENG W P, WANG W J, et al. Community detection algorithm based on weighted dense subgraphs[J]. Journal of Software, 2017, 28(11): 3103-3114.

[52] LIU Z H, WANG H M, WANG G S, et al. Link community detection combined with network pruning and local community expansion[J]. Modern Physics Letters B, 2021: doi.org/10.1142/S0217984921500986.

[53] GAO Y, YU X Z, ZHANG H L. Uncovering overlapping community structure in static and dynamic networks[J]. Knowledge-Based Systems, 2020: doi.org/10.1016/j.knosys.2020.106060.

[54] GAO Y, YU X Z, ZHANG H L. Graph clustering using triangle-aware measures in large networks[J]. Information Sciences, 2022, 584: 618-632.

[55] GAO Y, ZHANG H L, YU X Z. Higher-order community detection: on information degeneration and its elimination[J]. IEEE/ACM Transactions on Networking, 2023, 31(2): 891-903.

[56] GAO Y, ZHANG H L, YU X Z. Overlapping community detection based on conductance optimization in large-scale networks[J]. Physica A: Statistical Mechanics and Its Applications, 2019, 522: 69-79.

[57] GAO Y, YU X Z, ZHANG H L. Overlapping community detection by

constrained personalized PageRank[J]. Expert Systems with Applications, 2021, 173: 114682.

[58] ZHANG Y L, XIA X W, XU X, et al. Robust hierarchical overlapping community detection with personalized PageRank[J]. IEEE Access, 2020, 8: 102867-102882.

[59] FU S, WANG G Y, XU J, et al. IbLT: an effective granular computing framework for hierarchical community detection[J]. Journal of Intelligent Information Systems, 2022, 58(1): 175-196.

[60] JEH G, WIDOM J. SimRank: a measure of structural-context similarity[C]//Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2002: 538-543.

[61] OKUDA M, SATOH S, SATO Y, et al. Community detection using restrained random-walk similarity[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(1): 89-103.

[62] GASTEIGER J, BOJCHEVSKI A, GÜNNEMANN S. Predict then propagate: graph neural networks meet personalized PageRank[J]. arXiv Preprint, arXiv: 1810.05997, 2018.

[63] BOJCHEVSKI A, GASTEIGER J, PEROZZI B, et al. Scaling graph neural networks with approximate PageRank[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM Press, 2020: 2464-2473.

[64] DANON L, DÍAZ-GUILERA A, DUCH J, et al. Comparing community structure identification[J]. Journal of Statistical Mechanics: Theory and Experiment, 2005: doi.org/10.1088/1742-5468/2005/ 09/P09008.

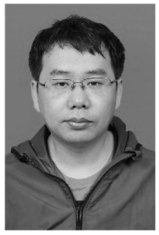
[65] MCDAID A F, GREENE D, HURLEY N. Normalized mutual information to evaluate overlapping community finding algorithms[J]. arXiv Preprint, arXiv: 1110.2515, 2011.

[66] LI Y X, HE K, KLOSTER K, et al. Local spectral clustering for overlapping community detection[J]. ACM Transactions on Knowledge Discovery from Data, 2018, 12(2): 1-27.

[67] NEWMAN M E J. Modularity and community structure in networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2006, 103(23): 8577-8582.

[68] BIAN Y C, HUAN J, DOU D J, et al. Rethinking local community detection: query nodes replacement[C]//Proceedings of IEEE International Conference on Data Mining (ICDM). Piscataway: IEEE Press, 2021: 930-935.

[作者简介]



高阳（1986− ），男，黑龙江哈尔滨人，博士，哈尔滨工业大学讲师，主要研究方向为数据挖掘、社交网络分析等。

张宏莉（1973− ），女，吉林榆树人，博士，哈尔滨工业大学教授、博士生导师，主要研究方向为社交网络分析、网络与信息安全等。