doi:10.3969/j.issn.1672-1292.2022.01.001

一种基于非对称三角形割的重叠社区发现算法

郑文萍[1,2,3],毕欣琦[1],杨　贵[1]

(1.山西大学计算机与信息技术学院,山西 太原 030006)
(2.山西大学计算智能与中文信息处理教育部重点实验室,山西 太原 030006)
(3.山西大学智能信息处理研究所,山西 太原 030006)

[摘要]　发现由相似功能的个体所形成的社区结构是复杂网络分析的重要任务之一. 提出一种基于非对称三角形割的重叠社区发现算法,首先根据社区内三角形连接情况对社区质量进行评价,并根据节点与社区的三角形连接定义了节点对社区的归属度和连接强度. 考虑到网络不同部分连接密度的差异,在将节点从社区中移除或加入社区的过程中,为每个节点分别设置了不同的移除阈值和扩展阈值,以提高社区发现质量. 将每个节点与其邻居节点组成初始社区,将归属度低于移除阈值的边缘节点从社区中移除,将连接强度高于扩展阈值的外围节点加入社区,社区节点移除和扩展阶段迭代进行直至社区结构趋于稳定,最后去掉重叠率过高的社区得到最终结果. 在 7 个带社区标签的网络上将所提算法与其他 7 个经典重叠社区检测算法进行比较,通过重叠标准互信息和 F_1 指标进行评价,结果表明所提算法可以较好地发现不同规模网络中的社区结构.

[关键词]　复杂网络,社区发现,重叠社区发现算法,非对称三角形割,社区适应度
[中图分类号]TP39　[文献标志码]A　[文章编号]1672-1292(2022)01-0001-08

An Overlapping Community Detection Algorithm Based on Asymmetric Triangle Cuts

Zheng Wenping[1,2,3],Bi Xinqi[1],Yang Gui[1]

(1.School of Computer and Information Technology,Shanxi University,Taiyuan 030006,China)
(2.Key Laboratory of Computation Intelligence and Chinese Information Processing of Ministry of Education,
Shanxi University,Taiyuan 030006,China)
(3.Institute of Intelligent Information Processing,Shanxi University,Taiyuan 030006,China)

Abstract:Overlapping community detection has attracted more and more attention in the research of complex networks. An overlapping community detection algorithm has been presented based on asymmetric triangle cuts(ATCO). The fitness of a community is defined as the ratio of the triangles within the community and the asymmetric triangle cuts. Furthermore, the membership and connection strength of a node to a community is defined according to the triangle connection between the node and the community. Considering that difference parts of the complex network usually have different link density, we compute specific removal threshold and extension threshold to each node in community reduction and expansion process. ATCO algorithm consists of three main processes: community initialization, node removal and extension, and high overlapping community removing. An initial community consists of a node and its neighbors, and the neighbors are the periphery of the initial community. A peripheral node will be removed from the community, if its membership to the community is lower than the predefined removal threshold. An external node will be added to a community, if its connection strength to the community is higher than the predefined extension threshold. The removal and extension process will be performed iteratively until a stable result is obtained. Finally, communities with high overlap will be postprocessed. Compared with other 7 classical overlapping community detection algorithms on 7 networks with ground-truth, the proposed algorithm ATCO shows good performance in overlapping standard mutual information and F_1 index.

Key words:complex network, community detection, overlapping community detection algorithm, asymmetric triangle cut, community fitness

收稿日期:2021-08-31.
基金项目:国家自然科学基金项目(62072292、62006145)、山西省自然科学基金项目(201801D121123).
通讯作者:郑文萍,博士,教授,研究方向:复杂网络分析、图聚类算法. E-mail:wpzheng@ sxu.edu.cn

社区发现是复杂网络分析的关键任务之一[1]. 研究社区结构有助于人们探索系统实体间的互作用模式,了解系统运行机制[2-4]. 重叠社区结构是社区发现研究的重点之一.

早期,研究者提出了利用网络中的特殊结构发现重叠社区结构,如派系过滤算法 CPM[5]、局部谱聚类算法[6]、基于非负矩阵分解的重叠社区发现算法 BigClam[7],这些算法通常需要预先给定社区数目且对大规模网络计算代价较高. 改进标签传播算法如 COPRA[8]、SLPA[9]等根据邻居节点标签更新节点的社区标签,由于允许社区重叠,算法难以较快得到稳定的社区发现结果.

Baumes 等[10]提出基于中心社区扩展策略的重叠社区发现算法,仅计算重要节点对社区连边密度比率的影响,确定其重叠社区归属. Clauset[11]计算边界节点与社区内连边的密度,得到社区的局部模块度,通过优化局部模块度函数发现社区. Lancichinetti 等[12]提出 LFM 算法,以连边密度作为社区适应度指标进行社区发现. 2015 年,Bandyopadhyay[13]等提出 FOCS 算法,将节点对社区的归属度定义为节点在社区的邻居节点数目占社区总节点数的比值,根据节点归属度对社区进行扩展. Yadav[14]根据节点与社区中所有节点的 Pagerank 分数得到节点对社区的归属度以扩展社区.

节点对社区的归属度定义和社区适应性度量是社区检测算法的核心. 目前大多数算法通过节点与在某社区中邻居节点的数量来确定节点对社区的归属度,也即,节点在某社区中邻居节点越多则归属度越高. 实际上,节点对社区的归属度不仅与社区内邻居节点数有关,还与这些邻居节点间的连接方式有关.

2015 年,Soundarajan 等[15]提出 Node-Perception 算法,首先在每个节点的邻域内进行局部社区发现以得到初始子社区,以这些子社区为节点构造新网络进行重叠社区检测. 2017 年,Epasto 等[16]提出 Ego-Splitting 算法对节点邻域进行初始社区发现并构造新网络,进而采用传统社区检测算法寻找社区. 这些方法利用现有的社区检测算法确定节点的直接邻域内节点的连接密度,当直接邻域内节点较多时,可较好地发现节点的社区归属,但计算代价较高,不适用大规模网络.

利用节点之间形成的三角形可以更好地度量节点连边的紧密程度. 2012 年,Zhang 等[17]给出了 k 核三角形子图的定义,寻找网络中所有极大 k 核三角形子图作为最终的社区检测结果. 2016 年,Benson 等[18]利用三角形割度量社区的显著性,通过最小化社区三角形割与社区内部三角形数的比值进行社区发现. 2018 年,Rezvani 等[19]将有 2 个节点位于某社区的三角形割称为该社区的非对称三角形割,提出了基于非对称三角形割的社区检测算法,以消除所发现社区的分离效应和搭便车效应.

本文提出一种基于非对称三角形割的重叠社区发现算法(the asymmetric triangle cut overlapping community detection algorithm,ATCO),利用非对称三角形割定义节点对社区的归属度和连接强度,并对社区结构显著性进行评价,给出基于非对称三角形割的社区适应度评价函数. ATCO 算法主要包括社区初始化、社区节点的移除和扩展、高重复率社区处理 3 个主要过程. 在社区节点移除阶段,根据节点对其周围社区的三角形连接情况,对不同节点设置了专门的移除阈值和扩展阈值,将归属度低于移除阈值的边缘节点从社区中移除,将连接强度高于扩展阈值的外围节点加入社区. 社区的移除和扩展阶段迭代进行直至社区结构趋于稳定,最后去掉重叠率过高的社区,得到最终结果. 在 7 个带社区标签的网络上将所提 ATCO 算法与其他 7 个经典重叠社区检测算法进行比较,通过重叠标准互信息和 F_1 指标进行评价,结果表明 ATCO 算法可较好地发现不同规模网络中的社区结构.

1 基础知识

1.1 基本概念和术语

一个复杂网络可以用图 $G=(V,E)$ 来表示,其中节点集 $V=\{v_i \mid 1\leqslant i\leqslant n\}$ 表示网络节点的集合,边集 $E=\{(v_i,v_j) \mid 1\leqslant i,j\leqslant n\}$ 代表节点间连边的集合. 除非特别声明,本文仅考虑无权无向的简单图.

节点 v 的直接邻居节点的集合称作其直接邻域,定义为 $N(v)=\{u \mid u\in V,(u,v)\in E\}$,则节点 v 的度 $d(v)=|N(v)|$. 对节点子集 $S\subseteq V$,令 $E_{\text{in}}(S)=\{(u,v) \mid u\in S,v\in S,(u,v)\in E\}$ 表示图 G 中两端点均属于子集 S 的边集,$E_{\text{out}}(S)=\{(u,v) \mid u\in S,v\notin S,(u,v)\in E\}$ 表示有且仅有 1 个端点属于子集 S 的边集. 节点集 S 的体积定义为 $\text{vol}(S)=\sum_{v\in S} d(v)$,显然 $\text{vol}(S)=2E_{\text{in}}(S)+E_{\text{out}}(S)$. 对节点子集 S 和 S',若 $S\cap S'=\varnothing$,则将连接这两个不相交节点子集的边集记作 $E(S,S')=\{(u,v) \mid u\in S,v\in S',(u,v)\in E\}$.

令 $\Omega=\{S_1,S_2,\cdots,S_K\}$,其中 $S_i\neq\varnothing$ 且 $\bigcup\limits_{1\leqslant i\leqslant K}S_i\subseteq V$,若 $S_i\cap S_j=\varnothing(1\leqslant i\neq j\leqslant K)$,则称 Ω 为图 G 的一种非重叠社区检测结果;若存在两个子集 S_i 和 $S_j(1\leqslant i\neq j\leqslant K)$ 使得 $S_i\cap S_j\neq\varnothing$,则称 Ω 为图 G 的一种重叠社区检测结果;称 S_i 为图 G 的一个社区. 对节点 $v\in V$,令 $C_v=\{S\mid v\in S,S\in\Omega\}$ 表示 Ω 中包含节点 v 的社区集合,则 $\bar{C}_v=\Omega-C_v$ 是 Ω 中不包含节点 v 的社区集合.

假设互不相同的节点 $u,v,w(\in V)$ 满足 $(u,v)\in E,(w,v)\in E$ 且 $(u,w)\in E$,则称节点 u,v 和 w 构成了图 G 中的一个三角形,记作 Δ_{uvw}. 令 Δ_G 表示图 G 中所有三角形的集合. 对节点子集 $S\subseteq V$,用 Δ_S 表示 S 中所有三角形的集合.

定义 1(三角形割和非对称三角形割)[19] 设 $\Omega=\{S_1,S_2,\cdots,S_K\}$ 是图 G 的一种社区检测结果,Δ_{uvw} 是端点分别为 u,v 和 w 的一个三角形. 对社区 $S_i\in\Omega$,若 $\{u,v,w\}\cap S_i\neq\varnothing$ 且 $\{u,v,w\}-S_i\neq\varnothing$,即社区 S_i 包含且仅包含三角形 Δ_{uvw} 的部分端点,则称 Δ_{uvw} 为社区 S_i 的一个三角形割. 若社区 S_i 包含且仅包含三角形 Δ_{uvw} 的 2 个端点,则称 Δ_{uvw} 为社区 S_i 的一个非对称三角形割. 社区 S_i 的所有三角形割的集合记作 $\Delta_{S_i\text{-cut}}$,所有非对称三角形割的集合记作 $\Delta_{S_i\text{-asycut}}$. 对节点 u,令 $\Delta_S(u)=\{\Delta_{uvw}\mid u\in S,v\in S,w\in S,\Delta_{uvw}\in\Delta_S\}$ 表示社区 S 中包含节点 u 的三角形的集合,$\Delta_{S\text{-cut}}(u)=\{\Delta_{uvw}\mid u\notin S,\Delta_{uvw}\in\Delta_{S\text{-cut}}\}$ 表示节点 u 与社区 S 形成的三角形割的集合,$\Delta_{S\text{-asycut}}(u)=\{\Delta_{uvw}\mid u\notin S,v\in S,w\in S,\Delta_{uvw}\in\Delta_{S\text{-asycut}}\}$ 表示节点 u 与社区 S 形成的非对称三角形割的集合.

1.2 社区适应度

Kannan 等[20]提出了社区导电性指标,如式(1)所示,用来衡量社区 S 与网络其他部分的连接强度. 通常,社区的导电性 $\delta(S)$ 越大,其社区结构越不明显:

$$\delta(S)=1-\frac{E_{\text{out}}(S)}{\min\{\text{vol}(S),\text{vol}(V-S)\}}. \tag{1}$$

Baumes 等[10]通过社区内部连边密度与社区向外连边密度的比值来衡量所在社区结构显著性,如式(2)所示:

$$f(S)=\frac{w_S^{\text{in}}}{w_S^{\text{in}}+w_S^{\text{out}}}, \tag{2}$$

式中,S 为检测到的某社区,w_{in}^S 表示社区 S 的内部连边密度,w_{out}^S 表示社区 S 的向外连边密度,如式(3)所示:

$$\begin{cases} w_S^{\text{in}}=\dfrac{2E_{\text{in}}(S)}{|S|\times(|S|-1)}, \\ w_S^{\text{out}}=\dfrac{E_{\text{out}}(S)}{|S|\times(|V|-|S|)}. \end{cases} \tag{3}$$

式(1)和(2)给出的社区适应度评价指标考虑了社区内所有节点的连边情况对社区显著性的影响. Clauset[11]认为社区内与其他社区存在连接的节点连边情况才会对社区显著性产生影响,给出了局部模块度评价指标,如式(4)所示,其中 $B_S\subseteq S$ 表示社区 S 中与其他社区有连接的节点集合,也称作 S 的边界节点:

$$\mu(S)=\frac{|E(B_S,S-B_S)+2E_{\text{in}}(B_S)|}{\text{vol}(B_S)}. \tag{4}$$

2 非对称三角形割指标

上述社区适应度指标通常仅考虑社区内部或社区间的连边数对社区显著性的影响,实际上,具有相同顶点数和边数的社区网络拓扑属性有很大的差异,因此其社区显著性也应有较大差异. 除社区内外的连边数对社区结构产生影响外,这些连边所形成不同的网络拓扑结构对社区结构也有很大影响. 为了更好地度量网络拓扑结构对社区结构的影响,本文基于三角形割对社区适应度进行定义(见定义 2),同时对节点与社区的关系进行定义.

定义 2(基于非对称三角形割的社区适应度)　设 $\Omega=\{S_1,S_2,\cdots,S_K\}$ 是图 G 的一种社区检测结果,社区 S_i 内的三角形集合为 Δ_{S_i},S_i 的非对称三角形割集合为 $\Delta_{S_i-\text{asycut}}$,定义社区 S_i 的适应度 $\tau(S_i)$ 为:

$$\tau(S_i)=\frac{|\Delta_{S_i}|}{|\Delta_{S_i-\text{asycut}}|+|S_i|}. \tag{5}$$

基于非对称三角形割的社区适应度考虑了社区内部所形成的三角形数以及该社区拥有的切割三角形数. 本文所给的社区适应度指标可以很好地应用于重叠社区的适应度计算. 图 1 中给出了包含 2 个具有重叠节点的社区图例,其中节点 12 和 13 为重叠节点. 表 1 给出了采用 Kannan 指标 $\delta(S)$、Baumes 指标 $f(S)$、Clauset 指标 $\mu(S)$ 和本文给出的非对称三角形割指标 $\tau(S)$ 的计算结果. 可以看出,$\tau(S)$ 对重叠社区的识别结果更好.

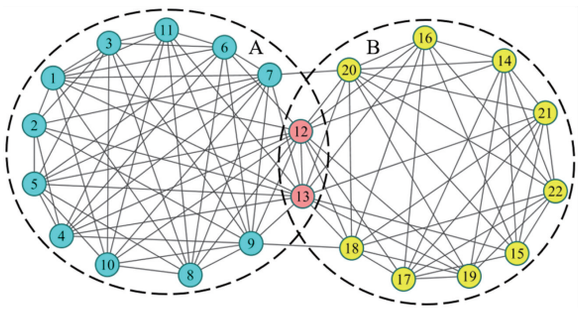

图 1　包含 2 个重叠社区的网络

Fig. 1　Network with 2 overlapping communities

表 1　重叠社区适应度指标比较结果

Table 1　Comparison of overlapping community fitness indicators

适应度函数	A	B	$A-B$	$B-A$	发现社区
$\delta(S)$	0.837	0.851	0.851	0.833	$A-B,B$
$f(S)$	0.888	0.905	0.895	0.919	$A,B-A$
$\mu(S)$	0.810	0.774	0.837	0.812	$A-B,B-A$
$\tau(S)$	2.367	2.32	1.461	1.437	A,B

定义 3(节点对社区的归属度)　设 $\Omega=\{S_1,S_2,\cdots,S_K\}$ 是图 G 的一种社区检测结果,则节点 v 对社区 S_i 的归属度 $\zeta(v;S_i)$ 定义为:

$$\zeta(v;S_i)=\begin{cases}\dfrac{|\Delta_{S_i}(v)|}{|\Delta_{S_i}|}, & v\in S_i;\\ 0, & v\notin S_i.\end{cases} \tag{6}$$

定义 3 根据社区 S_i 中以节点 v 为端点的三角形数占社区内总三角形数的比例给出了社区内节点 v 对当前社区 S_i 的归属度. 显然,节点对社区的归属度越大,则节点留在当前社区的可能性越大.

定义 4(节点与社区的连接强度)　设 $\Omega=\{S_1,S_2,\cdots,S_K\}$ 是图 G 的一种社区检测结果,则节点 v 与社区 S_i 的连接强度 $\xi(v;S_i)$ 定义为:

$$\xi(v;S_i)=\begin{cases}\dfrac{|\Delta_{S_i-\text{cut}}(v)|}{|\Delta_G(v)|}, & v\notin S_i;\\ 0, & v\in S_i.\end{cases} \tag{7}$$

定义 4 根据社区外节点 v 与社区 S_i 所形成的切割三角形数占网络中包含节点 v 的总三角形数的比例给出了社区外节点 v 与社区 S_i 的连接强度. 显然,节点与社区的连接强度越大,则节点加入当前社区的可能性越大.

3　基于非对称三角形割重叠社区检测算法

3.1　初始化社区

首先将所有度不小于 2 的节点及其邻居节点组成初始社区. 由于网络中的节点度通常服从幂律分布,即存在少量大度节点与网络中大部分节点间存在连边,这些大度节点形成的初始社区往往包含网络中的一些小规模社区,导致最终无法发现这些小规模社区. 为了提高社区发现准确率,首先移除社区内节点数超过网络总节点数 60%的初始社区. 记所得的 K 个初始社区集合为 $\Omega^0=\{S_1,S_2,\cdots,S_K\}$,其中初始社区 $S_i(1\leqslant i\leqslant K)$ 包括中心节点 v_{S_K}与其邻居节点.

为了移除对社区归属度低的节点,需要计算社区中所有节点对当前社区的归属度. 随着社区规模的逐步扩大,计算量也不断增长. 为了减少计算代价,仅考虑那些最有可能被移除社区的节点集,称为社区 S 的边缘节点集 B_S. 显然节点 v_i 是初始社区 S_i 的中心节点,不应从初始社区中移除,此时初始社区 S_i 的边缘节点集 $B_{S_i}=S_i-\{v_{S_i}\}$. 为了方便进行社区扩展,定义社区 S 的外围节点集为与 S 中节点相邻且不在 S 中的节点集合,即 $W_S=\{v\mid u\in S,v\notin S,(u,v)\in E\}$.

3.2 高重复率社区处理

社区节点的移除和扩展阶段是迭代进行的,两个社区重复节点越多,经过移除阶段和扩展阶段后社区组成相同的可能性越高,会导致不必要的重复计算. 定义社区 S_i 和 S_j 的重复率为:

$$O(S_i,S_j)=\frac{|S_i\cap S_j|}{\min\{|S_i|,|S_j|\}}. \tag{8}$$

利用式(8),计算第 l 次迭代产生的社区间的重复率,假设社区 S_i 和 S_j 重复率 $O(S_i,S_j)\geqslant 0.6$,则删除规模较小的社区,记去重后第 l 次迭代产生的社区集合为 $\Omega^l=\{S_1,S_2,\cdots,S_K\}$.

3.3 社区中节点的移除和扩展

3.3.1 移除阈值和扩展阈值

本阶段需根据节点对社区的归属度(或连接强度)对社区进行节点移除(或扩展),考虑到节点周围局部拓扑结构在网络中的差异性,此处采用与文献[13]类似的方法,为每个节点分别设置相应的移除阈值和扩展阈值.

对任意节点 $v\in G$,计算其与当前社区集合 Ω^l 中每个社区 S 的归属度 $\zeta(v;S)$ 和连接强度 $\xi(v;S)$. 由于归属度和连接强度取值区间为[0,1],因此将[0,1]区间平均分割为 20 份,每份取值区间长度为 0.05. 根据节点 v 对 Ω^l 中每个社区的归属度值,统计落在每个区间中的社区个数. 为描述方便,令 BC_i 表示位于区间$[0.05i,0.05i+0.05)$中的社区个数,其中 $0\leqslant i<20$. 令 $r_{\max}=\underset{0\leqslant i<20}{\operatorname{argmax}}\{BC_i\}$ 为社区个数非零的最右区间号. 若某区间 x 满足条件 $BC_x<BC_{r_{\max}}$且,节点 v 的移除阈值设置为 $0.05\times\zeta_v^{\text{cut-off}}$,其中

$$\zeta_v^{\text{cut-off}}=\begin{cases}\underset{0<j<r_{\max}}{\operatorname{argmax}}\{BC_j\leqslant BC_{r_{\max}},BC_j<BC_{j-1}\}, & \text{otherwise};\\ 0, & \text{if } BC_j\geqslant BC_{j-1} \text{ for } 0<j<r_{\max};\\ r_{\max}, & \text{if } r_{\max}=0.\end{cases}$$

3.3.2 社区节点的移除

首先根据式(5)计算当前社区的适应度值,删除那些对社区归属度低于移除阈值且不降低社区适应度的边缘节点. 假设当前考虑社区为 S,根据式(5),其社区适应度为:$\tau(S)=\frac{|\Delta_S|}{|\Delta_{S-\text{asycut}}|+|S|}$. 考虑移除边缘节点 v,原社区 S 中包含 v 的三角形 $\Delta_S(v)$ 变成社区 $S-\{v\}$ 的非对称三角形割,而 $\Delta_{S-\text{asycut}}(v)$ 中的三角形不再是社区 $S-\{v\}$ 的非对称三角形割,因此移除 v 后社区适应度为:$\tau(S-\{v\})=\frac{|\Delta_S|-|\Delta_S(v)|}{|\Delta_{S-\text{asycut}}|+|\Delta_S(v)|-|\Delta_{S-\text{asycut}}(v)|+|S|-1}$. 进而可得,当节点 v 和社区 S 所形成的三角形和非对称三角形割满足$|\Delta_S(v)|>\frac{|\Delta_S|}{|\Delta_S|+|\Delta_{S-\text{asycut}}|+|S|}\times(|\Delta_{S-\text{asycut}}(v)|+1)$时,有 $\tau(S-\{v\})<\tau(S)$.

也即,当节点 v 在社区 S 中所形成的三角形个数大于所形成的非对称三角形割一定比例时,删除该节点会降低社区适应度. 对于该类型节点,尽管其归属度低于移除阈值,也不应从社区移除. 由于节点移除会使社区规模下降,将节点数目少于 3 的社区删除.

3.3.3 社区节点的扩展

首先计算社区 S_i 的外围节点集 W_{S_i}中节点 v 对该社区的连接强度,将连接强度大于其扩展阈值 $\xi_v^{\text{cut-off}}$ 的节点加入社区. 将本阶段新加入社区 S_i 的节点集定义为下一次迭代前该社区的外围节点集. 本轮迭代所得到社区结构 Ω^l 作为下一轮迭代中社区移除阶段的输入.

若两次迭代任何社区的边缘节点集 B_S 和外围节点集 W_S 都未发生改变,停止迭代. 网络中存在节点与其邻居节点未形成三角形,没有为其分配社区,将无社区归属的节点放入其邻居节点所在社区,得到最终的社区 Ω^l,算法终止.

3.4 算法描述

算法 1 给出了所提出的基于非对称三角形割重叠社区发现算法框架.

算法 1　基于非对称三角形割重叠社区发现算法

输入:网络 $G=(V,E)$.

输出:社区发现结果 $\Omega=\{S_i \mid S_i \in \Omega, 1 \leqslant i \leqslant K\}$.

Step 1　令 $t=0, \Omega^t=\varnothing, flag=0$.

Step 2　对图 G 的每个顶点 $v_i \in G(1 \leqslant i \leqslant |V|)$,若 $d(v) \geqslant 2$,则令 $S_i=\{v_i\} \cup N(v_i)$, $\Omega^t=\Omega^t \cup \{S_i\}$, $B_{S_i}=N(v_i)$.

Step 3　对任一 $S \in \Omega^t$,若 $|S|>0.6|V|$,则 $\Omega^t=\Omega^t-\{S\}$.

Step 4　对任意 2 个社区 $S_i, S_j \in \Omega^t$(假设 $|S_i| \leqslant |S_j|$),若 $O(S_i,S_j) \geqslant 0.6$,则 $\Omega^t=\Omega^t-\{S_i\}$.

Step 5　对任一社区 $S_i \in \Omega^t$, $rnodes=\varnothing$.

Step 5.1　对节点 $v \in B_{S_i}$,计算节点 v 对社区的归属度 $\zeta^t(v;S_i)$ 和移除阈值 $\zeta_v^{\text{cut-off}}$. 若 $\zeta^t(v;S_i)<\zeta_v^{\text{cut-off}}$ 且 $\tau(S_i)<\tau(S_i-\{v\})$,则令 $rnodes=rnodes \cup \{v\}$, $flag=1$.

Step 5.2　令 $S_i=S_i-rnodes$.

Step 6　对任一社区 $S_i \in \Omega^t$,令其外围节点集为 $W_{S_i}=\{v \mid u \in S_i, v \notin S_i, (u,v) \in E\}$, $anodes=\varnothing$.

Step 6.1　对节点 $v \in W_{S_i}$,计算节点 v 对社区的连接强度 $\xi(v;S_i)$ 和扩展阈值 $\xi_v^{\text{cut-off}}$. 若 $\xi(v;S_i) \geqslant \xi_v^{\text{cut-off}}$,则令 $anodes=anodes \cup \{v\}$, $flag=1$.

Step 6.2　令 $S_i=S_i \cup anodes$, $B_{S_i}=anodes$.

Step 7　若 $flag=1$,则令 $\Omega^{t+1}=\Omega^t$, $t=t+1$, $flag=0$,返回 Step 3;若 $flag=0$,则将未分配社区节点加入其邻居节点所在社区.

Step 8　结束.

3.5　时间复杂度分析

对网络 $G=(V,E)$,其中 $n=|V|$, $m=|E|$. 社区初始化阶段需扫描图 G 的邻接表,因此时间消耗为 $O(n+m)$. 计算每个节点所在的三角形数的平均代价为 $O(n \times \overline{d}^2)=O(2\overline{d}m)$,其中,$\overline{d}$ 为网络节点平均度.

假设在社区发现过程,社区平均规模为 A_{comm},则计算一个社区的三角形和非对称三角形割的平均代价为 $O(A_{\text{comm}}^3)$. 由于网络中最多包含 n 个社区,所以此阶段总代价最多为 $O(n \times A_{\text{comm}}^3)$. 假设每个节点属于的平均社区数为 A_{vc},则每次迭代中,所有节点进行移除和扩展总代价为 $O(n \times A_{vc})$.

综上,本文所提算法的总时间复杂度为 $O(2\overline{d}m+tn \times A_{\text{comm}}^3+tn \times A_{vc})$,其中 t 为算法迭代次数.

4　实验结果与分析

在 7 个带社区标签的网络上将所提算法 ATCO 与其他 7 个经典重叠社区检测算法 Ego_Networks[16]、Egonet_Splitter[16]、Kclique[5]、Node_Perception[15]、SLPA[9]、FOCS[13] 和 CoreExp[19] 等进行实验比较,实验结果通过重叠标准互信息[21] 和 F_1 指标[22] 进行评价,数据如表 2 所示.

表 2　数据集基本情况

Table 2　Basic information of datasets

数据集	$\|V\|$	$\|E\|$	社区
Karate	34	78	2
Dolphin	62	159	2
Football	115	613	12
Adjnoun	109	425	2
Internet_partnerships	219	631	3
Amazon	334 863	925 872	75 149
DBLP	317 080	1 049 866	13 477

表 3 给出了各算法重叠标准互信息的比较结果,可以看出 ATCO 在大多数网络上 OvmNMI 明显优于对比算法. 特别地,在 Amazon 和 DBLP 两个网络规模较大的数据集上 Ego_Networks、Ego_Splitter 和 Node_Perception 等算法发现的社区数远大于真实社区数目,导致 OvmNMI 值很低. Kclique 和 SLPA 算法在网络非常稀疏的情况下,社区发现质量明显较差. 由于 FOCS 算法考虑了节点在网络中的差异,为每个节点设立了不同的扩展阈值,因而在大规模网络上取得了较好的结果. 本文算法 ATCO 不仅考虑了节点阈值的差异性,还利用非对称三角形割对节点和社区的关系进行定义,实验结果得到了进一步提升. 表 4 给出了各算法 F_1 分数的比较结果,可以看出本文算法 ATCO 和对比算法 FOCS 都取得了较好的结果,在多数网络上本文算法明显优于其他算法. SLPA 算法的 F_1 分数在规模较小网络上表现优于本文算法 ATCO,随着网络规模的增大,其性能急剧下降. 这是由于规模较小网络中社区构成也相对简单,SLPA 可快速得到比较稳定的社区发现结果;随着网络规模的增大,社区构成变得复杂,在有限的迭代次数内无法得到稳定的社区发现结果,导致其社区发现性能大幅度降低. CoreExp 算法首先发现一些内部连接稠密的核社区,通过扩展核社区进行社区发现,仅能发现内部稠密连接的社区.

表 3　OvmNMI 的实验比较结果

Table 3　Comparison results on OvmNMI

Network	Ego_Networks	Ego_Splitting	Kclique	Node_Perception	SLPA	FOCS	CoreExp	ATCO
Karate	0.141	0.132	0.165	0.122	0.132	0.243	0.186	0.829
Dolphin	0.102	0.102	0.275	0.096	0.121	0.250	0.484	0.286
Football	0.252	0.302	0.160	0.204	0.302	0.555	0.011 1	0.635
Adjnoun	0.020 3	0.000 461	0.000 461	0.013 2	0.004 63	0.002	0.000 101	0.023 1
Internet_partnerships	0.019 1	0.032 8	0.006 08	0.024 9	0.013 5	0.006 75	0.021 5	0.038 7
Amazon	0.153	0.112	0.094 4	0.104	0.093	0.047 9	0.062 6	0.101
DBLP	0.090 2	0.045 1	0.053 9	0.057 4	0.036 6	0.057 9	0.015 4	0.070 9

表 4　F_1 指标实验比较结果

Table 4　Comparison results on F_1

Network	Ego_Networks	Ego_Splitting	Kclique	Node_Perception	SLPA	FOCS	CoreExp	ATCO
Karate	0.485	0.421	0.446	0.337	0.775	0.54	0.525	0.97
Dolphin	0.145	0.146	0.480	0.168	0.970	0.386	0.820	0.415
Football	0.221	0.737	0.570	0.516	0. 901	0.216	0.161	0.673
Adjnoun	0.083 3	0.061	0.580	0.063 0	0.680	0.127	0.590	0.108
Internet_partnerships	0.047 8	0.028 9	0.102	0.028 5	0.083 0	0.05	0.690	0.083 3
Amazon	0.031 2	0.037 1	0.053 9	0.042 5	0.049 7	0.089 2	0.074 4	0.057 2
DBLP	0.054 1	0.059 7	0.087 5	0.065 2	0.134	0.074 8	0.338	0.071 2

图 2 给出了在 1 000 个节点的 LFR 人工网络上预测重叠节点的准确率的比较结果,其中网络的平均度为 20,最大度为 50,重叠节点数为 100,横轴表示混合参数,纵轴表示预测所得重叠节点数与真实重叠节点数的比值. 可以看出,Ego_Networks 发现的重叠节点数量较多,由于 Ego_Networks 算法发现大量节点数为 3 的小社区,导致识别重叠节点的准确率较高;除此之外,ATCO 算法识别重叠节点准确率最高,且随着混合参数的增加,ATCO 算法发现重叠节点的准确率也较稳定.

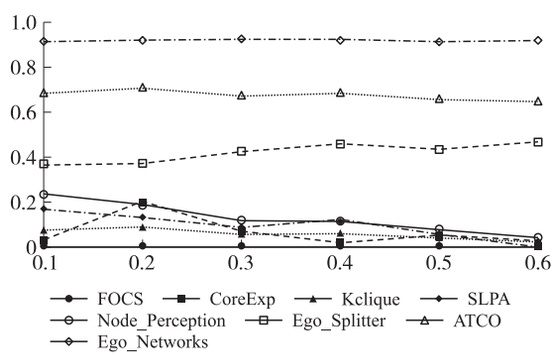

图 2　重叠节点准确率

Fig. 2　Accuracy of overlapping nodes

5　结论

本文提出了一种基于非对称三角形割的重叠社区发现算法 ATCO,在社区质量评价和节点与社区关系度量过程中引入了三角形连接机制. 对社区内节点,根据社区中包含该节点的三角形情况定义节点对社区的归属度;对社区外节点,根据节点与社区形成的非对称三角形割定义节点与社区的连接强度;同时考虑到网络不同部分连接密度差异对节点与社区的归属度或连接强度绝对数值的影响,在社区节点移除和扩展过程中,为每个节点分别设置了不同的移除阈值和扩展阈值,以提高社区发现质量. 在 7 个带社区标签的网络上将 ATCO 与其他 7 个重叠社区检测算法进行比较,通过重叠标准互信息和 F_1 指标进行评价,结果表明 ATCO 算法能较好地发现不同规模网络中的社区结构.

本文利用非对称三角形割进行社区发现,在一定程度上提高了重叠社区发现的质量. 然而,不同的网络中三角形连接情况存在较大差异,如何根据这些差异合理定义社区评价函数以及节点与社区的关系度量,需进一步探索.

[参考文献](References)

[1] JAVED M A,YOUNIS M S,LATIF S,et al. Community detection in networks:a multidisciplinary review[J]. Journal of Network and Computer Applications,2018,108:87-111.

[2] YANG J,LESKOVEC J. Defining and evaluating network communities based on ground-truth[J]. Knowledge and Information

Systems,2015,42(1):181-213.
[3] 杨蒙蒙,王水花,陈燚,等. 生物地理学优化算法与应用综述[J]. 南京师范大学学报(工程技术版),2018,18(2):50-55.
[4] 王俊淑,张国明,胡斌. 基于深度学习的推荐算法研究综述[J]. 南京师范大学学报(工程技术版),2018,18(4):33-43.
[5] 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.
[6] MAGDON-ISMAIL M, PURNELL J. SSDE-Cluster: fast overlapping clustering of networks using sampled spectral distance embedding and GMMs[C]//Proceedings of the 2011 IEEE 3rd International Conference on Social Computing. Boston,USA: IEEE,2011:756-759.
[7] YANG J,LESKOVEC J. Overlapping community detection at scale:a nonnegative matrix factorization approach[C]//Proceedings of the 6th ACM International Conference on Web Search and Data Mining. Rome,Italy:ACM,2013:587-596.
[8] GREGORY S. Finding overlapping communities in networks by label propagation[J]. New Journal of Physics,2010,12(10): 2011-2024.
[9] XIE J R,SZYMANSKI B K,LIU X M. SLPA:uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process[C]//Proceedings of the IEEE 11th International Conference on Data Mining Workshops. Vancouver,Canada:IEEE,2011:344-349.
[10] BAUMES J,GOLDBERG M K,KRISHNAMOORTHY M S,et al. Finding communities by clustering a graph into overlapping subgraphs[C]//Proceedings of the 2005 IADIS International Conference on Applied Computing. Algarve,Portugal:DBLP, 2005:97-104
[11] CLAUSET A. Finding local community structure in networks[J]. Physical Review E,Statistical,Nonlinear,and Soft Master Physics,2005,72(2 Pt 2):026132.
[12] LANCICHINETTI A, FORTUNATO S, KERTÉSZ J. Detecting the overlapping and hierarchical community structure in complex networks[J]. New Journal of Physics,2009,11(3):033015.
[13] BANDYOPADHYAY S,CHOWDHARY G,SENGUPTA D. FOCS:fast overlapped community search[J]. IEEE Transactions on Knowledge and Data Engineering,2015,27(11):2974-2985.
[14] YADAV N. Community-affinity:measuring strength of memberships of nodes in network communities[D]. Salt Lake City: University of Utah,2015.
[15] SOUNDARAJAN S,HOPCROFT J E. Use of local group information to identify communities in networks[J]. ACM Transactions on Knowledge Discovery from Data,2015,9(3):1-27.
[16] EPASTO A, LATTANZI S, LEME R P. Ego-splitting framework: from non-overlapping to overlapping clusters[C]// Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Halifax,Canada: ACM,2017:145-154.
[17] ZHANG Y,PARTHASARATHY S. Extracting analyzing and visualizing triangle k-core motifs within networks[C]//Proceedings of the 2012 IEEE 28th International Conference on Data Engineering. Arlington,USA:IEEE,2012:1049-1060.
[18] BENSON A R,GLEICH D F,LESKOVEC J. Higher-order organization of complex networks[J]. Science,2016,353(6295): 163-166.
[19] REZVANI M,LIANG W F,LIU C F,et al. Efficient detection of overlapping communities using asymmetric triangle cuts[J]. IEEE Transactions on Knowledge and Data Engineering,2018,30(11):2093-2105.
[20] KANNAN R,VEMPALA S,VETA A. On clusterings-good,bad and spectral[J]. Journal of the ACM,2004,51(3):497-515.
[21] CHAKRABORTY T, DALMIA A, MUKHERJEE A, et al. Metrics for community analysis: a survey[J]. ACM Computing Surveys,2017,50(4):1-37.
[22] ROSSETTT G,PAPPALARDO L,RINZIVILLO S. A novel approach to evaluate community detection algorithms on ground truth[C]//Proceedings of the 7th Workshop on Complex Networks CompleNet. Dijon,France:Springer,2016:133-144.

[责任编辑:严海琳]