基于多节点社团意识系统的属性图聚类算法

石 铠 任泺锟 彭一鸣 李慧嘉
(中央财经大学管理科学与工程学院 北京 100081)

摘 要 属性图用属性向量描述节点,用边描述节点间的关系。为了把节点划分为具有紧密联系的社团,一种有效的方法是对属性图进行聚类。聚类方法有不同的标准,如节点连接度和属性相似度。虽然社团一般是围绕紧密的连边和相似的属性值的节点形成,但是目前的方法都只关注了这两种数据形式中的一种。通过给每个节点赋予一个自治域,提出一个准确且可延展的多节点系统用于提取属性图中的重叠社团。首先,引入带有可调带宽因子的核函数用于测度每个节点的影响力,具有最高局部影响力的节点可以被看作领导节点。其次,提出一种新颖的局部扩展策略,使每一个领导节点能够吸收属性图中相关性最强的跟随者。接着,设计了多节点社团意识系统,该系统为节点之间的充分沟通提供了必要的条件,从而能够得出最优的重叠社团结构。社团中的节点不仅互相联系紧密,而且也有相似的属性。该算法的计算复杂度在特定带宽条件下近似于连边数目的线性函数。最后,基于标准属性图和真实属性图的实验验证了该系统的有效性和高效性。

关键词 聚类,属性图,多节点意识系统,中心度,重叠节点
中图法分类号 TP393 **文献标识码** A

Attributed Graph Clustering Algorithm Based on Cluster-aware Multiagent System

SHI Kai REN Luo-kun PENG Yi-ming LI Hui-jia
(School of Management Science and Engineering, Central University of Finance and Economics, Beijing 100081, China)

Abstract Existing methods to partition nodes into clusters with tight correlations is to apply clustering techniques on attributed graphs based on node connectivity or attribute similarity. In this paper, we comprehend each node as an autonomous agent and developed an accurate multi-agent system to extract overlapping clusters in attributed graphs. First, a kernel function with bandwidth factor is introduced to measure the influence of each agent, and those agents with highest local influence are selected as leader agents. Next, a novel local expansion strategy is proposed, by which each leader agent absorbs the closest followers in the graph. Then, the cluster-aware multiagent system was designed so that the optimal overlapping cluster configuration can be uncovered. Our method is highly efficient, whose computational time nearly linearly depends on the number of edges. Finally, the proposed method is demonstrated on synthetic benchmark graphs and real-life attributed graphs to verify the systematic performance.

Keywords Clustering, Attributed graph, Multiagent cluter-aware system, Centrality, Overlapping nodes

1 引言

在数据挖掘技术中,聚类分析[1-2]是聚合相似对象、分离相异对象的一项技术。传统的聚类分析方法通常将研究对象当作向量来处理。现如今网络知识库和在线社交网络变得十分流行,内含属性信息和网络信息的数据源在近几年内增长迅速。为了能够在这些属性图中找到更有实际意义的特性,最好同时考虑属性特征和拓扑特征。本文主要致力于研究如何在复杂信息中提取有意义的社团,这一方法也叫做属性图聚类[3]。

尽管属性图聚类是一个很复杂的工作,但属性图聚类的应用十分广泛,比如在在线社交网络中进行社团探测[4]、在无线传感器网络中减少能量消耗[5]、在众包平台中优化任务分配[6]。虽然在过去的一段时间内,学术界在属性图聚类技术上有很多研究成果,包括 CCN[10]、CODICIL[12]、CohsMix[13]、贝叶斯概率模型[14]以及 CESNA[4],但是许多问题仍然没有得到解决。大多数传统社团发现方法忽视了对象的属性,仅仅专注于研究网络拓扑特征。尽管它们能发现有内在粘结结构的社团,但同一个社团中的对象可能会有一些不同的属性值。

本文提出了一种有效的面向自治域的算法[8],该算法将输入的属性图看作分布式的多节点系统,其中存储了每个节点的拓扑信息和属性信息。通过应用多节点系统,可以获得最优重叠的社团结构。同一个社团中的节点不仅互相紧密连接,而且具有相似的属性值。具体来说,首先通过引入带有一个可调节的带宽因子 δ 的影响力函数,可以使用多重解析方法找出具有最高局部影响力的领导节点;接着将社团的两种属性即调整紧密度(Adjusted Tightness)和同质度(Homogeneity)联系在一起。本文精心设计了多节点社团意识系统(Cluster-aware Multiagent System)。该系统由物理层、沟通

本文受国家自然科学基金(71401194),中央财经大学"青年英才"培育支持计划(QYP 1603)资助。
石 铠(1996-),男,硕士生,主要研究方向为社会网络;任泺锟(1995-),女,硕士生,主要研究方向为数据挖掘;彭一鸣(1997-),男,硕士生,主要研究方向为社会网络;李慧嘉(1985-),男,博士,副教授,主要研究方向为社会网络、数据挖掘、运筹学,E-mail:Hjli@amss.ac.cn。

层和应用层 3 部分组成，并且独立于使用的局部扩展策略，所以能够被应用在其他的图聚类情形下。最后，基于该系统提出了精确、快速的面向自治域的算法，该算法在处理大型属性图时具有良好的可延展性。

2 问题的制定

2.1 准备知识

定义属性图 $G=(V,E,X)$，其中节点集 $V=\{v_1,\cdots,v_n\}$，边集 $E=\{e_{ij}\}$，属性矩阵 $X=[x_{ij}]_{n\times p}$，x_{ij} 是节点 v_i 的第 j 个属性值。因此，属性矩阵的第 i 行 X_i 可被看作节点 v_i 的属性向量。本文主要分析二元属性，也就是当第 i 个节点表现出第 j 个属性时，$x_{ij}=1$，否则 $x_{ij}=0$。我们的目标是利用属性图的聚类方法找出优良的 K 个社团，即在 G 中有 $P:=\{C_1,\cdots,C_K\}$。同时，一个完美的聚类框架应该在以下两条性质中取得平衡：1)社团内节点连接良好；2)社团内节点具有相似的属性值。

为了更好地解释以上两条性质，本文将以一个特定的社团 C_K 为例来介绍以下概念。

定义 1(紧密度) 社团 C_K 的紧密度可以被定义为：

$$T(C_K)=n_k(n-n_k)\left(\frac{2L_k^{in}}{n_k^2}-\frac{L_k^{out}}{n_k(n-n_k)}\right)=\frac{2(n-n_k)}{n_k}L_k^{in}-L_k^{out} \tag{1}$$

其中，n_k 是社团 C_K 内的节点个数，L_k^{in} 是社团 C_K 内边的数量，L_k^{out} 是社团 C_K 间边的数量。因子 $n_k(n-n_k)$ 对很小和很大的社团有惩罚作用，同时可以产生更多的均衡解。

定义 2(均匀性) 基于 Havrda-Charvat 关于二元离散概率分布的广义熵定理，C_K 的均匀性可以被定义为：

$$\psi(C_K)=-\sum_{j=1}^{p}c_{kj}(1-c_{kj}) \tag{2}$$

其中，$c_k=(c_{k1},\cdots,c_{kp})$ 定义了社团 C_K 的属性中心，元素 c_{kj} 是第 j 个属性为 1 的概率。类似地，一个较大的 $\psi(C_K)$ 表明 C_K 上的节点有相同的属性值。

基于调整后的紧密度和均匀性的定义，在给定属性图的条件下，一个好的 K 聚类划分应该满足：对于在 P 中的每一个 c_k，$T(C_K)$ 和 $\psi(C_K)$ 都应该尽可能大。

2.2 社团领导节点的识别

这里把属性图看作一个物理系统，在这个系统内所有的节点都互相影响。推断这种影响产生了一种作用力，这种作用力使任意两个节点间都相互直接连接。同时，这种作用力还要满足以下 3 条性质：1)相比其他节点，领导节点应该有更高的局部中心性；2)随着两个节点物理距离的不断增加，这种吸引力将迅速减弱；3)两个节点越相似，它们之间的相互吸引就越大。基于以上 3 条性质，一个度量每个节点相互影响的方法就是应用 Gaussian 核函数 $K(\cdot)$。通过将可调整带宽引入核函数 $K(\cdot)$，每个节点的影响区域可以由带宽因素 δ 控制。在属性图中，每一个节点 v_i 的影响可以被定义为：

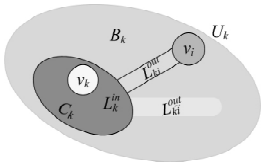
$$\begin{aligned}&K(v_i,\delta)=\frac{1}{|\Gamma_i(\delta)|}\sum_{v_j\in\Gamma_i(\delta)}f_{ij}(\delta)\\&f_{ij}(\delta)=\frac{X_iX_j^{\mathrm{T}}}{||X_i|\cdot|X_j||}\mathrm{e}^{-\frac{d_{ij}^2}{2\delta^2}}\\&\Gamma_i(\delta)=\{v_i\,|\,d_{ij}\leqslant[2\delta]\}\end{aligned} \tag{3}$$

其中，$\delta\in(0,+\infty)$ 是用来控制每个节点的影响区域的带宽因子；$f_{ij}(\delta)$ 是节点 v_i 和 v_j 之间的吸引力；d_{ij} 是节点 v_i 和 v_j 之间的最短路径；$\frac{X_iX_j^{\mathrm{T}}}{||X_i|\cdot|X_j||}$ 是两个属性向量 X_i 和 X_j 的余弦相似度，其值越大，两个节点越相似；$\Gamma_i(\delta)$ 是对节点 v_i 的影响区域，对于每一个特定的 δ，每个节点对其他节点的影响带宽大约是 $[2\delta]$，当其大于 $[2\delta]$ 时，指数方程的值 $\mathrm{e}^{-\frac{d_{ij}^2}{2\delta^2}}$ 将迅速减小为 0，所以可以运用 δ 来控制每个节点的影响区域，并且在仅仅考虑两个相互影响的影响区域为 $d_{ij}\leqslant[2\delta]$ 的节点时计算 $K(v_i,\delta)$。因此如果一个节点的局部影响是最大的，即 $\forall v_j\in\Gamma_i(\delta),K(v_i,\delta)\geqslant K(v_j,\delta)$，那么它就是领导节点。我们还推导出领导节点的数量依赖于带宽因子 δ，δ 越大，在属性图中能提取的领导节点的数目就越少。

2.3 领导节点的社团扩展

在确定隐藏在属性图中所有的领导节点后，社团 K 的数目就被固定。因此，通过在每一个领导节点采用局部扩展策略，就可以得到 K 个社团的划分。从领导节点 v_k 的角度来看，可以将待解的属性图划分为 3 个区域：归属于领导节点 v_k 的社团 C_K、边界区域 B_K 和巨大的未知区域 U_K。其中 C_K 包括了 v_k 最重要的追随节点，B_K 和 C_K 紧密联系，即 $B_K=\{v_j\,|\,v_j\notin C_K\wedge v_i\in C_K\wedge e_{ij}\in E\}$；$U_K$ 对 C_K 不可见，同时 U_K 可以被定义为 $U_K=\{v_j\,|\,v_j\notin C_K\cup B_K\}$。在以上设定条件下，基于一些促进 $\hat{T}(C_K)$ 和 $\psi(C_K)$ 的数值尽可能高的预定义标准，v_k 可以用 C_K 来进行局部扩展。更重要的是，我们认为 B_K 和 C_K 对 v_k 都是可见的，因此 v_k 的可行区域可以被定义为 $V_K=\{C_K\cup B_K\}$。

假设领导节点 v_k 已经将其所在的社团拓展到 C_K，现在就可以求得一个如图 1 所示的边界节点 $v_i(v_i\in B_K)$。



注：C_K 是由领导节点 v_k 扩展而成的局部社团，B_K 是社团 C_K 的边界区域，U_K 是较大的未知区域，L_k^{in} 是 C_K 中的社团内连边数，L_{ki}^{out} 是 C_K 和边界节点 v_i 之间的连边数，$L_{\tilde{k}i}^{out}$ 是社团 C_K 和 B_K 中除 v_i 以外对节点之间的连边数。

图 1 局部扩展模型

接下来将区分 3 种类型的连接：在 C_K 社团内的连接(定义为 L_k^{in})；在 C_K 和 v_i 之间(定义为 L_{ki}^{out})；在 C_K 和 B_K 中的其他节点之间(定义为 $L_{k\tilde{i}}^{out}$)，此外 $L_k^{out}=L_{ki}^{out}+L_{k\tilde{i}}^{out}$。为了简化计算，利用 L_k^{in} 和 k_i(v_i 的节点度)表示内在连接的数目：$L_k^{out}=aL_k^{in}=bk_i$；$L_{ki'}^{out}=cL_k^{in}$，$a\geqslant\frac{1}{L_k^{in}}$，$b\geqslant\frac{1}{k_i}$(因为对于任意的在 B_K 中的 v_i 都至少在 C_K 中有一个邻接节点)，$c\geqslant0$。基于以上设定，现在社团 C_K 调整后的紧密度为：

$$T(C_K)=\frac{n-n_k}{n_k}2L_k^{in}-(a+c)L_k^{in} \tag{4}$$

在 C_K 吸收了 v_i 节点后，调整后的紧密度变为：

$$T(C_K\cup v_i)=\frac{n-n_k-1}{n_k+1}2L_k^{in}(1+a)-(cL_k^{in}+k_i-bk_i) \tag{5}$$

准则 1 为了确保一个较高水平的调整后的紧密度值，一个领导节点 v_k 只能吸收 B_K 上的边界节点，这会提高现在

的$\hat{T}(C_K)$。在合并 v_i 节点后,调整后的紧密度的收益可以被定义为:

$$\begin{aligned}\Delta\hat{T}_{C_k}(v_i)&=\hat{T}(C_K\bigcup v_i)-\hat{T}(C_K)\\&=\frac{n-n_k-1}{n_k+1}2L_k^{in}(1+a)-(cL_k^{in}+k_i-bk_i)-\\&\quad\frac{n-n_k}{n_k}2L_k^{in}-(a+c)L_k^{in}\\&=2n\frac{bk_in_k-L_k^{in}}{n_k(n_k+1)}-k_i\end{aligned}\tag{6}$$

因此,如果 $\Delta\hat{T}_{C_k}(v_i)>0$,则边界节点 v_i 可以加入 C_K 中。

准则 1 表明:如果就度数而言,一个小的节点连接了一个 C_K 社团,则将该小节点纳入将会提高调整后 C_K 的紧密度。此外,一旦一个节点 v_i 被社团 C_K 吸收,吸收后的 C_K 的均匀性即变为:

$$\begin{aligned}\psi(C_K\bigcup v_i)&=-\sum_{j=1}^{p}c'_{kj}(1-c'_{kj})\\&=-\sum_{j=1}^{p}\frac{c_{kj}n_k+x_{ij}}{n_k+1}(1-\frac{c_{kj}n_k+x_{ij}}{n_k+1})\\&=-\sum_{j=1}^{p}\frac{(n_k^2c_{kj}+n_kx_{ij})(1-c_{kj})+n_kc_{kj}(1-x_{ij})}{(n_k+1)^2}\end{aligned}\tag{7}$$

因此,纳入 v_i 后均匀性的增加值为:

$$\begin{aligned}\Delta\psi_{C_k}(v_i)&=\psi(C_K\bigcup v_i)-\psi(C_K)\\&=-\sum_{j=1}^{p}[\frac{n_k^2c_{kj}(1-c_{kj})+n_kx_{ij}(1-c_{kj})+n_kc_{kj}(1-x_{ij})}{(n_k+1)^2}-c_{kj}\\&\quad(1-c_{kj})]\\&=\sum_{j=1}^{p}\frac{-n_kc_{kj}(1-x_{kj})-n_kx_{ij}(1-c_{kj})+(2n_k+1)c_{kj}(1-c_{kj})}{(n_k+1)^2}\\&=\sum_{j=1}^{p}\frac{n_kx_{ij}(2c_{kj}-1)+n_kc_{kj}-2n_kc_{kj}^2+c_{kj}-c_{kj}^2}{(n_k+1)^2}\\&=\sum_{j=1}^{p}\frac{n_k(2c_{kj}-1)(x_{ij}-c_{kj})+c_{kj}(1-c_{kj})}{(n_k+1)^2}\end{aligned}\tag{8}$$

如果按照以上的方法求解,即利用 $\Delta\psi_{C_k}(v_i)>0$ 从领导节点 v_k 扩展社团,这就是一个冷启动问题。如同在图 2 中展示的一样,一旦我们在社团 C_K 中放置 v_k,属性中心 c_K 就是 v_k 的属性向量,$c_K=X_k$。因此,$\Delta\psi_{C_k}(v_i)$可以被写作:

$$\Delta\psi_{C_k}(v_i)=\sum_{j=1}^{p}\frac{n_k(2x_{kj}-1)(x_{ij}-x_{kj})}{(n_k+1)^2}\tag{9}$$

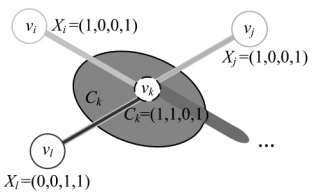


图 2　冷启动问题

从式(9)中可以发现 $\forall v_i\in B_K$,$\Delta\psi_{C_k}(v_i)\leqslant 0$,等式只在边界节点 v_i 和 v_k 有着完全相同的属性时才成立,因为对于社团扩展的条件 $\Delta\psi_{C_k}(v_i)>0$ 过于严格。一个可能避免这种冷启动问题的解决措施是放宽条件 $\Delta\psi_{C_k}(v_i)>0$。注意到式(9)可以被写作:

$$\begin{aligned}\Delta\psi_{C_k}(v_i)&=\sum_{j=1}^{p}\frac{n_kx_{ij}(2c_{kj}-1)}{(n_k+1)^2}+\\&\quad\sum_{j=1}^{p}\frac{n_kc_{kj}(1-2c_{kj})+c_{kj}(1-c_{kj})}{(n_k+1)^2}\end{aligned}\tag{10}$$

其中,第二项对于一个给定的社团 C_K 是固定的,最大化 $\Delta\psi_{C_k}(v_i)$就是最大化第一项。接下来有如下的定义:

定义 3(调整后的均匀性增量)　纳入 v_i 后,调整后的均匀性增量可以被定义为:

$$\Delta\hat{\psi}_{C_k}(v_i)=\sum_{j=1}^{p}\frac{n_kx_{ij}(2c_{kj}-1)}{(n_k+1)^2}\tag{11}$$

当且仅当调整后的均匀性增量是正的,即 $\Delta\hat{\psi}_{C_k}(v_i)>0$,边界节点 v_i 方可被纳入 C_K 中。

定义 4(局部扩展策略)　假定有一个已知的领导节点为 v_k 的社团,局部扩展策略可以被定义为:首先,计算 v_k 与 B_K 上边界点的吸引力 $f_{ki}(\delta)$,并且选择一个拥有最高吸引力的节点作为候选节点 v_i^*。基于准则 1 和准则 2,如果$\Delta\hat{T}_{C_k}(v_i^*)>0$ 和 $\Delta\psi_{C_k}(v_i^*)>0$ 同时成立,则 v_i^* 可以被看作 v_k 的跟随节点,接下来将 v_i^* 纳入 C_K 中,否则 v_i^* 将在 B_K 中被移除。一旦 v_i^* 纳入 C_K 中,那些由 v_i^* 连接的但是位于 U_K 的节点就会被纳入 B_K 中,接下来就会计算 v_k 和新纳入节点之间的吸引力。这个算法将会一直重复,直到 B_K 变为空集,在这种情况下,归属于领导节点 v_k 的整个社团将会被找到。

3　面向自治域进行属性聚类划分

本节首先引入多节点社团意识系统(Cluster-Aware Multiagent System,CAMAS);再给出一个属性聚类划分的分布式算法;最后进行算法复杂度的分析。

3.1　多节点社团意识系统

一个多节点社团意识系统被定义为 $CAMAS=\{A,n,\delta\}$,其中,$A=\{A_1,\cdots,A_n\}$是节点集合,n 是多节点社团意识系统中节点的总数,δ 是带宽因子。该系统中的每一个节点可以用元组表示为$\langle T_i,E_i,\Theta_i,\Gamma_i,K_i,\xi_i,C_i,B_i,V_i,n_i,l_i,c_i\rangle$,其中,$T_i=\{j|x_{ij}=1\}$是属性识别信号 j 的集合,j 由属性向量 X_i 给出;$E_i=\{j|A_i$ 与 A_j 相互连接$\}$是 A_i 的邻居集合;Θ_i 是 A_i 的信息池;$\Gamma_i=\{\langle j,d_{ij},T_j\rangle|\forall v_{ij}\in\Gamma_i(\delta)\}$是存储 A_i 影响域内拓扑属性信息的数据池;K_i 是节点 A_i 的影响力;ξ_i 是布尔变量,A_i 是一个领导节点时其值为真,否则,其值为假;$C_i=\{j|A_j\in C_i\}$是由 A_i 领导的社团;$B_i=\{\langle j,d_{ij},f_{ij}|A_j\in B_i\rangle\}$存储了 B_i 中边界节点的信息;$V_i=\{\langle j,d_{ij},f_{ij}|A_j\in V_i\rangle\}$存储了 V_i 中节点的信息;$n_i=|C_i|$是 C_i 中节点的数量;l_i 是 C_i 中内连边的数量;c_i 是社团 C_i 的属性中心。利用记号 T_i 把 $f_{ij}(\delta)$,$K(v_i,\delta)$和 $\Delta\hat{\psi}_{C_k}(v_i)$分别改写为:

$$f_{ij}(\delta)=\frac{|T_i\bigcap T_j|}{\sqrt{|T_i|\cdot|T_j|}}\mathrm{e}^{-\frac{d_{ij}^2}{2\delta^2}}\tag{12}$$

$$K(v_i,\delta)=\frac{1}{|\Gamma_i|}\sum_{\langle j,d_{ij},T_j\rangle\in\Gamma_i}\frac{|T_i\bigcap T_j|}{\sqrt{|T_i|\cdot|T_j|}}\mathrm{e}^{-\frac{d_{ij}^2}{2\delta^2}}\tag{13}$$

$$\Delta\hat{\psi}_{C_k}(v_i)=\sum_{j\in T_i}\frac{n_k(2c_{kj}-1)}{(n_k+1)^2}\tag{14}$$

定义 5　多节点社团意识系统的全局目标的实现条件是所有领导节点的边界节点的集合为空集。

$$\forall\xi_i=\text{“真”},B_i=\emptyset\tag{15}$$

定义 6(沟通机制)　为实现全局目标,每个节点可以与其他节点在预设好的沟通框架下交流。每一个节点 A_i 的标识符 i 表示 A_i 的地址。其他节点(如 A_j)若知道 A_i 的标识符,就可

以向 A_i 发送信息。信息可以被定义为一个五元组$\langle \mathbb{S},\mathbb{T},\mathbb{P},\mathbb{R},\mathbb{D}\rangle$,其元分别表示根节点的标识符、目标节点的标识符、传播步长、需求集和数据集。在本系统中,有两种信息类别:1)广播信息,被用来更新每个节点的自治域。广播信息的形式是$\langle \mathbb{S},\emptyset,\mathbb{P},\emptyset,T_{\mathbb{S}}\rangle$。节点 A_i 从邻居收到一条广播信息时,会更新它的数据池,并且决定是否将这条信息转播给它的邻居。2)要求回复的信息,用于节点之间的信息复查。它的形式是$\langle \mathbb{S},\mathbb{T},\emptyset,\mathbb{R},\emptyset\rangle$。当目标节点收到了一条要求回复的信息时,它会把信息$\langle \mathbb{S},\mathbb{T},\emptyset,\emptyset,\mathbb{D}\rangle$回复给根节点。

可以注意到,每个节点保持的信息池 Θ_i 是基于队列结构的,并且处理的信息将会从 Θ_i 中剔除。

3.2 面向自治域的计算流程

基于多节点社团意识系统,进一步提出一个面向自治域的算法(Autonomy-oriented computing,AOC),如算法 1 所示。

算法 1 面向自治域的社团扩展算法

输入:属性图 $G=(V,E,X)$,带宽因子 δ

输出:K 层划分 $P=\{C_1,\cdots,C_k\}$

1. 生成多节点社团意识系统:$\forall A_i\in A, A_i\leftarrow\langle T_i,E_i,\emptyset,\emptyset,\emptyset,\emptyset,\emptyset,\emptyset,\emptyset,\emptyset,\emptyset,\emptyset\rangle$;
2. $P\leftarrow\emptyset$;
3. **循环**:对于 $\forall A_i\in A$
4. $\forall j\in E_i$,向 Θ_j 中添加$(\langle i,\emptyset,1,\emptyset,T_i\rangle)$;
5. 若 $\Theta_i\neq\emptyset$
6. 从 Θ_i 发布顶层信息$\langle \mathbb{S},\emptyset,\mathbb{P},\emptyset,T_{\mathbb{S}}\rangle$;
7. 如果 $i\neq\mathbb{S}\wedge\exists(\mathbb{S},d_{\mathbb{S}},T_{\mathbb{S}})\in\Gamma_i\wedge\mathbb{P}\leqslant\lfloor 2\delta\rfloor$
8. 向 Γ_i 中添加$\langle \mathbb{S},\mathbb{P},T_{\mathbb{S}}\rangle$,并且向 Θ_j 中添加$\langle \mathbb{S},\emptyset,\mathbb{P}+1,\emptyset,T_{\mathbb{S}}\rangle$;
9. 否则,如果 $i\neq\mathbb{S}\wedge\exists(\mathbb{S},d_{\mathbb{S}},T_{\mathbb{S}})\in\Gamma_i\wedge\mathbb{P}<d_{\mathbb{S}}\leqslant\lfloor 2\delta\rfloor$
10. $\langle \mathbb{S},d_{\mathbb{S}},T_{\mathbb{S}}\rangle\leftarrow\langle \mathbb{S},\mathbb{P},T_{\mathbb{S}}\rangle$,并且向 Θ_j 中添加$\langle \mathbb{S},\emptyset,\mathbb{P}+1,\emptyset,T_{\mathbb{S}}\rangle$;
11. 使用式(13)计算 K_i;
12. 如果 $K_i\ \forall\langle j,d_{ij},T_j\rangle\in\Gamma_i,K_i\geqslant K_j$
13. $\forall\xi_i=$“真”;
14. **循环**:对于 $\forall A_i\in A$
15. 如果 $\xi_i=$“真”
16. $A_i\leftarrow\langle T_i,E_i,\emptyset,\Gamma_i,K_i,$“真”$,\{i\},\{\langle j,d_{ij},f_{ij}\rangle\mid\forall j\in E_i\},\{j,d_{ij},f_{ij}\},1,0,X_i\rangle$;
17. 若 $B_i\neq\emptyset$
18. 从 B_i 中发布顶层元组$\langle j^*,d_{ij}^*,f_{ij}^*\rangle$;
19. 如果 $\Delta\hat{T}_{C_k}(v_i)>0\wedge\Delta\overset{\wedge}{\psi}_{C_k}(v_i)>0$
20. $B_i\leftarrow B_i\cup\{\langle k,d_{ik},f_{ik}\rangle\mid k\in E_{j*}\wedge k\notin C_i\}$;
21. $C_i\leftarrow C_i\cup\{j^*\}$并更新 V_i,n_i,l_i,c_i;
22. 否则
23. $B_i\leftarrow B_i-\langle j^*,d_{ij}^*,f_{ij}^*\rangle$
24. $P\leftarrow P\cup C_i$;

3.3 复杂度分析

本文提出的面向自治域的算法适用于节点可以独立运作并且没有中央节点控制的环境。这种环境使得确定运算时间非常困难,因为所有的节点能够同时运作。为了简化估计时间复杂度的任务,假定每个时间点考虑一个节点。

给定一个 n 个节点、m 条边、p 维属性的属性图,算法 1 对单层环境的计算复杂度在最坏情况下是 $O(n^2\langle p\rangle+K|V_i|\log|B_i|)$,其中$\langle p\rangle$是属性向量中非 0 元素的平均数量。

4 实验

本节展示多节点社团意识系统算法在 AGM 标准图[9]中摘取重叠集群的有效性。同时,本文将该算法和最先进的一些算法在有效性和可测性上进行了比较。3 种图聚类法被用于对比实验:SCD[19]和 Bigclam[20]是两种只考虑了图拓扑结构的、有效且可测的社团探测方法;CESNA[4]和 EDCAR[15]是两种最近提出的属性化图聚类方法,结合了图结构和节点属性;第 3 种算法是基于 AOC 的图聚类法,本节选取了 Yang 的方法[17]和 AOCCM[18]。在本文算法中,算法 1 中的属性向量已经被转化为事物集(Transaction Set),包括那些在相应属性维度展示的属性。为了寻找细密的社团结构,将带宽因素 δ 定为 0.5。以上实验的计算机环境为四核 E5-2650v2 (2.6GHz)处理器,128GB 缓存,240GB SSD disk,600GB SAS disk 和 2.6.32-5-amd64 内核的 Debian Linux 系统。

选取的几种算法的有效性可以用以下 6 个指标来衡量,其中有 3 种是无监督的,另外 3 种是有监督的。3 种无监督的指标为覆盖率(Cover)、标准化致密性(NorTi)和标准化同质性(NorHo)。另外 3 种有监督的指标为平均 F1 值(AvgF1)、平均 Jaccard(AvgJa)和调整的 Rand 指数(ARI),这 3 种指标都利用了测试图中真实社团的信息。

在此次实验中,本文在一系列由属性图模型(AGM[9])生成的人工网络中测试了 CAMAS 算法。将图的尺度从 100 线性扩大至 1000。对于每个实验设置,产生 50 个属性图,并将本算法与 CESNA[4]和 EDCAR[15]两种属性图聚类模型进行对比,结果如图 3 所示。

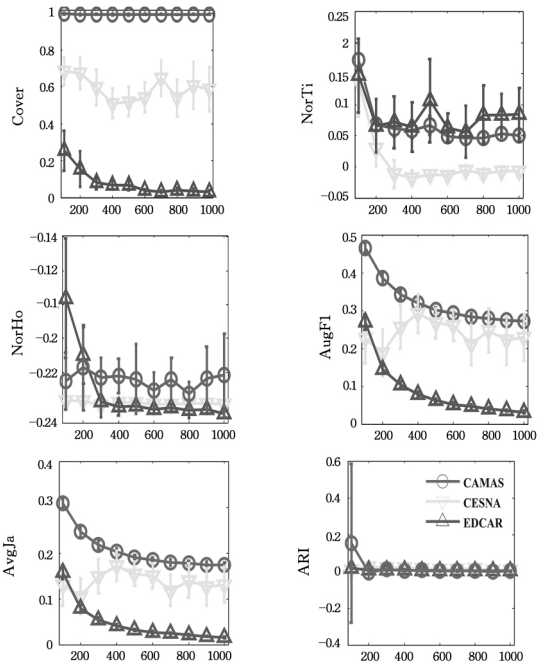


图 3 AGM 标准图上的有效性比较

从图 3 中可以观察到:1)由本算法得到的图划分结果几乎覆盖了所有的节点,CESNA 算法的结果覆盖了图中的一半节点,EDCAR 的覆盖率在所有的测试图中都较低;2)考察 NorTi 指标,EDCAR 在所有的测试图中几乎表现最好,本算法稍逊于 EDCAR,但优于 CESNA 算法;3)考察 NorHo 指标,本算法表现最好,CESNA 和 EDCAR 算法表现相似且都劣于本算法;4)AvgF1与 AvgJa 正相关,对于这两个指标,本算法表现最好,CESNA 和 EDCAR 次之;5)在所有测试的标准图中,ARI 的值都很低。在由 AGM 产生的结果示例中,每个点都根据属性和边的联合分布被赋于了二进制的值,这意味着产生的标准图只

包含两个真实社团。随着图尺度的增加,所有的测试算法都可以探测出 2 个以上的社团,因此 ARI 的测量值应减至一个较低水平。

图 4 进一步展示了不同算法在包含不同数量的边的图中的执行时间。点代表了在给定图中每种方法所花费的时间。由图 4 可以看出,SCD 是效率最高的;基于 Bigclam 和 AOC 的算法如 Yang 和 AOCCM 的运行比 CESNA 快一个梯度,比 EDCAR 在 Twitter 和 Google+上快两个梯度。CESNA 和 EDCAR 延展性不好的主要原因在于,混合两种来源的数据会使算法的效率降低。本文算法尽管也使用了两种来源的数据,但是效率仍近似于只使用图的拓扑结构。通过使用基于 AOC 的技术,本文算法在小型图(例如 Facebook 和 Twitter)上的运行效率甚至比 Bigclam 和 Yang 算法还高。

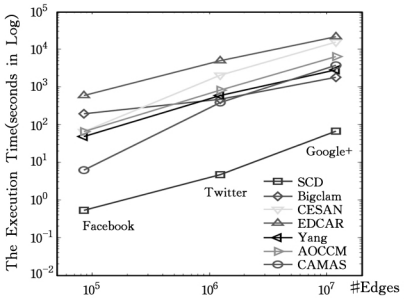


图 4　总体效率的比较

结束语　本文提出一个准确且可延展的多节点系统用于提取属性图中的重叠社团。首先,利用带有可调带宽因子的核函数测度节点的影响力,选出领导节点。其次,局部扩展策略使领导节点能够吸收属性图中相关性最强的跟随者。多节点社团意识系统为节点之间的充分沟通提供了必要的条件,从而能够得出最优的重叠社团结构。社团中的节点不仅互相联系紧密,而且也有相似的属性。该算法的计算复杂度在特定带宽条件下近似于连边数目的线性函数。最后,标准属性图和真实属性图的实验验证了该系统的有效性和高效性。为了使得算法更加高效,一些问题需要更深入的探讨,例如如何改善算法来处理更一般的属性(分类或定量),以及如何在超大规模的属性图上进行聚类划分。

参 考 文 献

[1] KAUFMAN L,ROUSSEEUW P J. Finding groups in data. an introduction to cluster analysis[J/OL]. http://library. mpibberlin. mpg. de/toc/z2007_1211. pdf.

[2] FORTUNATO S. Community detection in graphs[J]. Physics Reports,2010,486(3-5):75-174.

[3] BOTHOREL C,CRUZ J D,MAGNANI M,et al. Clustering attributed graphs: Models, measures and methods[J]. Network Science, 2015,3(3):408-444.

[4] YANG J,MCAULEY J,LESKOVEC J. Community Detection in Networks with Node Attributes[J/OL]. http://www-cs. stonford. edu/peole/jure/pubs/cesna-icdm13. pdf.

[5] LI Q,ZHU Q,WANG M. Design of a distributed energy-efficient clustering algorithm for heterogeneous wireless sensor networks[J]. Journal of Software,2006,29(12):2230-2237.

[6] WANG W,JIANG J,AN B,et al. Toward Efficient Team Formation for Crowdsourcingin Noncooperative Social Networks[J]. IEEE Transactions on Cybernetics,2016,pp(99):1-15.

[7] 李钝,李伦,张行进,等. 一种基于结构和属性的图聚类算法研究[J]. 小型微型计算机系统,2016,37(7):1469-1473.

[8] LIU J,JIN X,TSUI K C. Autonomy-oriented computing (AOC): formulating computational systems with autonomous components[J]. IEEE Transactions on Systems Man and Cybernetics-Part A Systems and Humans,2005,35(6):879-902.

[9] III J J P,MORENO S,FOND T L,et al. Attributed Graph Models:Modeling network structure with correlated attributes[J/OL]. http://www. conference. org/proceedings/www2014/proceedings/p831. pdf.

[10] BERLINGERIO M,PINELLI F,CALABRESE F. ABACUS: Apriori-Based Community discovery in mUltidimensional networkS[J]. Data Mining & Knowledge Discovery,2013,27(27): 294-320.

[11] 张昕尧,高宏. 一种新的属性图重叠聚类挖掘算法[J]. 智能计算机与应用,2012,2(5):27-30.

[12] RUAN Y,FUHRY D,PARTHASARATHY S. Efficient Community Detection in Large Networks using Content and Links [C]// International Conference on World Wide Web. 2012:1089-1098.

[13] ZANGHI H,VOLANT S,AMBROISE C. Clustering based on random graph model embedding vertex features[J]. Pattern Recognition Letters,2009,31(9):830-836.

[14] XU Z,KE Y,WANG Y,et al. A model-based approach to attributed graph clustering[C]// ACM Sigmod International Conference on Management of Data. 2012:505-516.

[15] GÜNNEMANN S,BODEN B,FÄRBER I,et al. Efficient Mining of Combined Subspace and Subgraph Clusters in Graphs with Feature Vectors[M]// Advances in Knowledge Discovery and Data Mining. 2013:261-275.

[16] 吴烨,钟志农,熊伟,等. 一种高效的属性图聚类方法[J]. 计算机学报,2013,36(8):1704-1713.

[17] YANG B,LIU J,LIU D. An autonomy-oriented computing approach to community mining in distributed and dynamic networks[J]. Autonomous Agents and Multi-Agent Systems, 2010,20(2):123-157.

[18] BU Z,WU Z,CAO J,et al. Local Community Mining on Distributed and Dynamic Networks From a Multiagent Perspective [J]. IEEE Transactions on Cybernetics,2015,46(4):986-999.

[19] PRATPEREZ A,DOMINGUEZSAL D,LARRIBAPEY J,et al. High quality,scalable and parallel community detection for large real graphs[C]// World Wide Web. 2014:225-236.

[20] YANG J,LESKOVEC J. Overlapping community detection at scale:a nonnegative matrix factorization approach[C]// ACM International Conference on Web Search and Data Mining. 2013: 587-596.

[21] 张素智,张琳,曲旭凯. 基于最短路径的加权属性图聚类算法研究[J]. 计算机应用与软件,2016,33(11):212-214.

[22] LI H J,BU Z,LI A,et al. Fast and accurate mining the community structure:integrating center locating and membership optimization[J]. IEEE Transactions on Knowledge & Data Engineering,2016,28(9):2349-3362.

[23] LI H J,DANIELS J J. Social significance of community structure:Statistical view[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics,2015,91(1):012801.