基于局部相似性的复杂网络社区发现方法

刘 旭 [1] 易东云 [1]

摘 要 复杂网络是复杂系统的典型表现形式, 社区结构是复杂网络最重要的结构特征之一. 针对复杂网络的社区结构发现问题, 本文提出一种新的局部相似性度量, 并结合层次聚类算法用于社区结构发现. 相对全局的相似性度量, 本文提出的相似性度量具有较低的计算开销; 同时又能很好地刻画网络的结构特征, 克服了传统局部相似性度量在某些情形下对节点相似性的低估倾向. 为了将局部相似性度量用于社区结构发现, 推广了传统的 Ward 层次聚类算法, 使之适用于具有相似性度量的任意对象, 并将其用于复杂网络社区结构发现. 在合成和真实世界的网络上进行了实验, 并与典型算法进行了比较, 实验结果表明所提算法的可行性和有效性.

关键词 复杂网络, 社区结构发现, 相似性度量, 层次聚类

DOI 10.3724/SP.J.1004.2011.01520

Complex Network Community Detection by Local Similarity

LIU Xu[1] YI Dong-Yun[1]

Abstract Complex networks are a typical form of representation of complex systems. Community structure is one of the most important structural characteristics of complex networks. In this paper, we propose a new measurement of similarity based on local structures for the purpose of detecting communities in complex networks. Compared to the similarity measures based on the entire network, the proposed similarity measure requires less computation and produces good descriptions of the structural characteristics of the networks. Meanwhile, it reverses the tendency of under-estimating produced by some existing similarity measures based on local structures. To utilize our measurement of similarity to the detection of community structures, we also generalize the Ward hierarchical clustering algorithm so that it is applicable to any object that has similarity measurement. And as an application we particularly employ this algorithm to detect community structures in complex networks. The proposed method is tested on both computer-generated and real-world networks, and is compared with the typical algorithms in community detection. Experimental results verify and confirm the feasibility and validity of the proposed method.

Key words Complex networks, community structure detection, similarity measurement, hierarchical clustering

复杂网络是复杂系统的重要表示形式之一, 已成为计算机、物理、数学、生物、社会学及复杂性科学等多学科的研究热点[1−3]. 通过将系统的组成部分用网络顶点集合 V 描述, 各组成部分间复杂关系用边集合 E 描述, 便得到系统的图或网络表示[4]$G = (V, E)$. 图或网络在数学表示方面并无实质区别, 之所以采用网络的称谓是为了强调研究对象的复杂特性或数据分析方面的特性. 例如科学家合作网络[5−6] 中节点表示论文作者, 边表示合作关系; 电子邮件网络[7] 中节点表示邮件地址, 边表示邮件收发关系. 对复杂网络在各个领域的应用感兴趣的读者可参考 Costa 的详尽综述[8].

一般将网络分为 4 种类型[2], 即信息网络、技术网络、社会网络以及生物网络, 它们普遍存在着小世界 (Small world)[9−10]、无标度 (Scale free)[11] 等宏观 (Macroscopic) 结构特征. 小世界特性是指网络顶点间的平均最短路径相对网络的规律而言极小, 例如某欧洲运营商旗下的 450 多万个用户的手机通讯网络的平均最短路径仅为 12.31[12], 任意两个用户只需大约 12 次跳转就能联系到对方. 然而, 无论仿真还是真实的动态扩散或搜索过程都显示小世界网络同时具有速度慢的特性, 即网上的扩散或搜索都未能有效利用网络的最短路径[13−14], 其中一个重要的原因就是网络的社区结构[12].

社区结构是复杂网络的一种介于宏观和微观之间的介观 (Mesoscopic) 结构特性[15−16], 它是网络节点的一种组织方式, 使得社区内部节点间的连接密度高于社区之间的连接密度[1,17]. 社会网络[18−19] 具有典型的社区结构, 人们按照职业、兴趣、居住地址等形成自然的团体, 团体内部成员拥有密切的相互关系. 实际上, 在信息网络、技术网络及生物网络中同样存在着社区结构[2,20], 它们和网络的功能及动态过程如网络同步[21] 有着密切的联系.

收稿日期 2011-04-15 录用日期 2011-07-16
Manuscript received April 15, 2011; accepted July 16, 2011
国家自然科学基金 (60902089, 61005003) 资助
Supported by National Natural Science Foundation of China (60902089, 61005003)
1. 国防科学技术大学理学院数学与系统科学系 长沙 410073
1. Department of Mathematics and Systems Science, College of Science, National University of Defense Technology, Changsha 410073

社区结构发现就是辨识网络社区的过程, 它对网络结构分析、功能演化及预测有重要的理论意义和实用价值. 目前对于社区结构没有公认的明确定义[20], 一个普遍接受的观点是社区内部节点拥有比随机连接更高的连接密度, 其量化表示就是模块度(Modularity)[22]. 以模块度为优化目标的社区发现是最热门的社区发现方法之一, 但精确求解使模块度最大的社区划分方式是 NP-完全问题[23]. 许多启发式搜索算法被提了出来[24], 包括层次聚类 (Hierarchical)、随机搜索[1]、极值优化[24]、旋转玻璃 (Spin-glass) 模型[25] 及矩阵谱分析[26−27] 等. 其中层次聚类算法分为凝聚型 (Agglomerative) 和分裂型 (Divisive) 两类, 凝聚型算法在初始时刻将每个节点划分为一个单独的划分, 然后每次合并两个划分; 分裂型算法在初始时刻将所有节点归入一个划分, 然后每次将某个划分一分为二, 最后将模块度最大的划分输出作为社区发现的结果. 分裂型算法以逐步移除具有最大介数中心性 (Betweenness centrality) 连接的 GN 算法[17, 28] 为代表, 凝聚型算法以逐步合并使模块度增加最快的两个社区的 CNM 算法为代表[29]. 在一定的意义上, 社区发现就是将距离较近或相似程度较高的点聚在一起的过程, 社区内部节点具有较高的连接密度, 可以看作是彼此距离较近或相似程度较高的一个判据. GN 算法的邻近关系判断准则是基于全局性的介数中心性来进行的, 因而具有很高的计算复杂度; CNM 算法的邻近关系判断准则是基于局部性的连接来进行的, 因而适宜处理规模较大的网络. 总体而言, CNM 算法的局部判据过于粗糙, GN 算法的全部判据过于精细. 因此, 本文提出一种介于二者之间的节点局部相似性度量, 它对网络结构的描述能力较强, 同时又具有相对简单的计算复杂度, 非常适合网络社区发现任务.

本文的贡献有两个方面, 一是定义了一种新的局部相似性度量, 拥有现存相似性度量不具备的良好性质, 使之更适用于网络社区发现; 二是将 Ward 聚类推广到任意具有相似性度量的抽象对象的聚类, 并将其用于网络社区结构发现.

通过对现有的网络节点局部相似性度量[30] 的分析, 我们发现它们对有连接点对间的相似性度量存在低估倾向, 在极端情况下甚至将两个有连接的节点的相似度赋值为零. 因此, 这些度量适用于未知链路预测[30], 但不适用于社区结构发现. 为此, 我们从社区发现任务出发, 构造了一种新的局部相似性度量, 它对有连接的点对给予较高的相似性度量. 此外, 它还对具有排他性的连接赋予较高的权重, 调低连接数较多的节点对其邻居点的相似性度量的贡献. 此外我们将欧氏空间的 Ward 聚类算法推广到所有具有两两距离或相似性度量的对象上来, 通过网络节点的相似性度量定义了网络划分间的 Ward 距离, 并通过 Ward 聚类算法进行网络社区结构发现. 在合成的网络和真实世界网络上进行的实验表明, 本文算法能够接近或超过典型的社区结构发现算法, 证实了所提相似性度量及聚类方法用于复杂网络社区结构发现的可行性和有效性.

1 网络结构的局部相似性度量

设网络 $G = (V, E)$ 具有 n 个节点和 m 条边, 其顶点集为 $V = \{V_1, \cdots, V_n\}$, 边集合为 $E = \{E_j | E_j \in V \times V, j = 1, \cdots, m\}$. 这里只考虑简单网络, 即无向 (Undirected)、无权 (Unweighted) 网络, 其邻接矩阵 A 的元素取值为 0 或 1, V_i 与 V_j 有边相连时 $A_{ij} = 1$, 否则 $A_{ij} = 0$. 顶点集 V 的一个覆盖或划分是指 V 的一个分组方式 $C = \{C_1, \cdots, C_c\}$, 使得任意 $V_i \in V$ 都存在唯一的一个 $C_k \in C$ 包含 V_i, 且 $\bigcup_{k=1}^{c} C_k = V$, $C_i \bigcap C_j = \phi, \forall i \neq j$. 模块度 $Q(C)$ 定义为[22, 28]

$$Q(C) = \frac{1}{2m} \sum_{k=1}^{c} \sum_{V_i, V_j \in C_k} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \quad (1)$$

其中, $k_i = \sum_{j=1}^{n} A_{ij}$ 是 V_i 拥有的连接数目, c 是划分 C 中的社区的数目.

模块度的作用有两个方面. 一方面, 用以评价划分的好坏, 对于给定网络的不同划分 $C_{(1)}$ 和 $C_{(2)}$, 如果 $Q(C_{(1)}) > Q(C_{(2)})$, 认为划分 $C_{(1)}$ 好于 $C_{(2)}$, 特别地, 使得 $Q(C)$ 取最大值的划分是网络 G 的最优划分; 另一方面, 用于评价网络社区结构的强弱, 对于不同的网络 G_1 和 G_2, 其最优划分为 C_{G_1} 及 C_{G_2}, 若 $Q(C_{G_1}) > Q(C_{G_2})$, 则认为网络 G_1 的社区结构强于 G_2. 通常社区结构较明显的网络的模块度介于 $0.3 \sim 0.7$ 之间[28].

1.1 模块度与相似性度量

事实上, 通过模块度可定义 G 中任意两个点的相似度, 将其相似度矩阵记为 S^Q, 则 $S^Q_{ij} = A_{ij} - \dfrac{k_i k_j}{2m}$. 当 V_i 和 V_j 没有边相连时 $A_{ij} = 0$, 一定有 $S^Q_{ij} < 0$; 当 V_i 和 V_j 有边相连时, 只要 $k_i k_j < 2m$, 就有 $S^Q_{ij} > 0$, 对于稀疏网络而言, $k_i k_j < 2m$ 是很容易满足的条件. 因此, 模块度鼓励将相似度大的点归入相同的划分, 将相似度小的点归入不同的划分. CNM 算法等模块度优化算法相当于根据上述相似度来进行社区结构发现, 每次合并相似度最大的两个社区, 使得最终的社区划分中位于同一社区的顶点相似性尽可能的大.

实际上, S^Q 是一种较为粗糙的度量, 这可

以从以下三个方面来理解. 首先, 按照模块度相似性的定义, 每个点和自身的相似度为负, 即 $S^Q_{ii} = -\dfrac{k_i^2}{2m} < 0$; 其次, 两个没有连接的顶点的相似度仍然与它们的连接数有关, 即 $S^Q_{ij} = -\dfrac{k_i k_j}{2m}$; 最后, 虽然考虑了参与相似度计算的两个点的除直接连接外的连接信息, 相似度却与其具体连接方式无关, 只与它们的乘积, 即 $k_i k_j$ 有关. 此外, 大多数情况下, S^Q 都不是一个稀疏矩阵, 这使得直接存储 S^Q 代价极高, 不适宜处理大型网络.

鉴于 S^Q 的上述缺陷, 我们需要寻找另外一种相似性度量, 使之能够提供尽可能丰富的结构信息, 同时又不必有太高的计算复杂度.

1.2 局部相似性度量

V_i 的邻居点集记为 $N(i)$, 即 $N(i) = \{V_j | A_{ij} = 1\}$. V_i 的星型邻域子图记为 $St(i)$, 它由 V_i 及其邻居点集构成, 即 $St(i) = \{V_i\} \bigcup N(i)$.

图 1 显示著名的空手道 (Karate) 俱乐部网络[31], 它是社区发现算法的标准测试网络之一. 图中用深色标记出了节点 18 的星型邻域子图 $St(18)$, 它包含 4 个节点, 即 $St(18) = \{V_{18}, V_{19}, V_{21}, V_{25}\}$; 右下角用深色标记了节点 6 和节点 7 的星型邻域子图的并集 $St(6) \bigcup St(7) = \{V_1, V_5, V_6, V_7, V_{28}, V_{33}\}$; 用正方形标记了它们的交集 $St(6) \bigcap St(7) = \{V_1, V_6, V_7, V_{33}\}$.

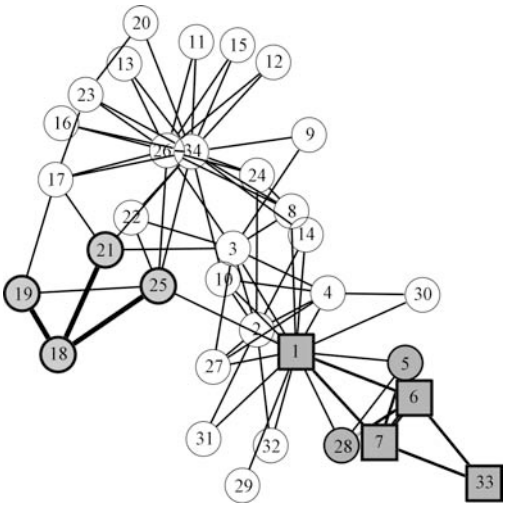


图 1 Karate 网络中的星型邻域子图

Fig. 1 Star neighborhood subgraph in the Karate network

不难看出, 集合 $St(i) \bigcap St(j)$ 反映了 V_i 和 V_j 的联系紧密程度. 此外, 节点连接数越多, 它出现在其余节点的星型邻域子图的次数也越多, 为相应点对的相似性带来的信息量越小; 反之, 如果一个节点只有少数的几个连接, 则为相应的点对带来的相似性信息量越大. 基于上述考虑, 定义 V_i 和 V_j 的局部相似性度量为

$$S_{ij} = \frac{\sum\limits_{V_e \in St(i) \bigcap St(j)} \frac{1}{k_e}}{\sqrt{\sum\limits_{V_e \in St(i)} \frac{1}{k_e}} \sqrt{\sum\limits_{V_e \in St(j)} \frac{1}{k_e}}} \tag{2}$$

连接较多的节点拥有较多的邻居, 而式 (2) 中的分母消除了这种尺度效应, 使得 $S_{ij} \in [0, 1]$. $S_{ij} = 0$ 等价于节点 V_i 和 V_j 没有长度小于 3 的路径相连. $S_{ij} = 1$ 有两种情况, 其中一种情况是节点 V_i 和自身的相似度, 即 $S_{ii} = 1$; 另外一种情况是 $A_{ij} = 1$ 且 $k_i = k_j = 1$, 即 V_i 和 V_j 相连且两个节点均只有一个连接. 换言之, 式 (2) 对于具有排它性的连接给予较高的权重, 如果两个节点间存在较多的连接, 但若这些连接通过一个中介节点到达第三方节点的数量越多, 单个这样的连接对上述节点相似性度量的贡献就越少.

以 Karate 网络为例可以计算, $St(6) \bigcap St(7) = \{V_1, V_6, V_7, V_{33}\}$, $S_{6,7} = 0.76$; $St(18) \bigcap St(19) = \{V_{18}, V_{19}, V_{25}\}$, $S_{18,19} = 0.79$, 显然 $S_{18,19} > S_{6,7}$. 可以看出, 虽然 V_6 和 V_7 有较多的共同邻居, 但其邻居有较多的连接, 如 V_1 有多达 16 个连接, 因此对 V_6 和 V_7 相似性的贡献较小; V_{18} 和 V_{19} 有较少的共同邻居, 但其邻居有较少的连接, V_{25} 仅有 6 个连接, 因此对 V_{18} 和 V_{19} 相似性的贡献较多, 最终导致 $S_{18,19} > S_{6,7}$. Karate 网络的相似度矩阵如图 2 所示, 黑色表示相似度为 1, 如对角线部分, 白色表示相似度为 0, $0 \sim 1$ 间的相似度用不同深度的灰色表示. 沿着相似度矩阵的对角线, 我们可以看到 4 个较为密集的非零矩阵块, 显示 Karate 网络的节点按照局部相似性度量大概分成了 4 个不同的区域, 但彼此间没有清晰的界限. 稍后我们将用层次聚类算法定量地给出这 4 个社区的划分.

1.3 需要新的相似性指标的原因

网络节点的相似性度量已有较为系统的研究[30], 大致分为两类: 即全局性相似性度量和局部相似性度量. 由于下面将指出的原因, 它们均不适合用于复杂网络的社区结构发现.

全局相似性度量有 Katz 指标[32]、LHN 指标[33] 以及基于随机游走的平均通勤时间指标 (Average commute time, ACT)[34] 等, 更多的全局相似性度量可参考文献 [30], 其共同特点是利用网络的邻接矩阵 A 来计算任意两点的相似度. Katz 指标由矩阵 $S^{\mathrm{Katz}} = (I - \beta A)^{-1} - I$ 来定义任意节点间的相似度, 其中 $0 < \beta < 1$ 是一个自由参数, 以保证矩阵逆的存在性. LHN 指标利用矩阵 $S^{\mathrm{LHN}} = 2m\lambda_1 D^{-1}$

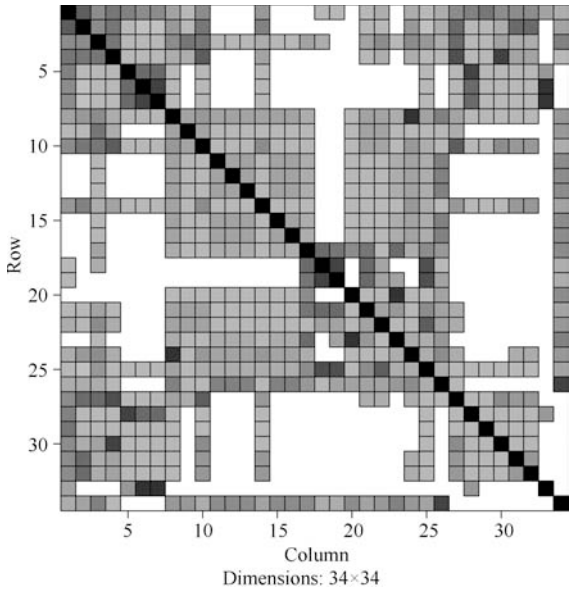


图 2　Karate 网络的相似度矩阵

Fig. 2　Similarity matrix for the Karate network

$\left(I-\frac{\phi A}{\lambda_1}\right)^{-1}D^{-1}$ 来给出任意节点间相似度, 其中 λ_1 是邻接矩阵 A 的最大特征根, $D=\mathrm{diag}\{k_1,\cdots,k_n\}$ 是由节点连接数组成的对角矩阵, $0<\phi<1$ 是一个自由参数, 以保证矩阵逆的存在性. 根据随机游走理论, V_i 和 V_j 间的平均通勤时间为[33] $2m(l^+_{ii}+l^+_{jj}-2l^+_{ij})$, 其中 l^+_{ij} 表示矩阵 $L^+=(D-A)^+$ 的相应元素, V_i 和 V_j 的 ACT 相似度定义为[34] $S^{\mathrm{ACT}}_{ij}=\frac{1}{l^+_{ii}+l^+_{jj}-2l^+_{ij}}$, 基于 ACT 还可进一步定义余弦相似性[30] $S^{\cos+}_{ij}=\frac{l^+_{ij}}{\sqrt{l^+_{ii}l^+_{jj}}}$. 不难发现, 全局相似性指标大多涉及矩阵求逆或矩阵特征根运算, 具有较高的计算复杂度, 不适宜较大规模的网络. 此外, Katz 指标和 LHN 指标引入的自由参数是为了矩阵求逆, 而不是为了描述节点的相似性.

局部相似性度量有 Jaccard 指标[35]、Salton 指标[36]、资源分配 (Resource allocation, RA) 指标等[37], 这里仅列举部分代表性度量, 详尽信息可参考文献 [30], 其共同特点是利用节点的邻域子图 $N(i)$. Jaccard 指标定义节点 V_i 和 V_j 的相似性度量为[35] $S^{\mathrm{Jaccard}}_{ij}=\frac{|N(i)\bigcap N(j)|}{|N(i)\bigcup N(j)|}$, 其中 $|N(i)|$ 表示集合 $N(i)$ 的元素个数. Salton 指标定义[36] $S^{\mathrm{Salton}}_{ij}=\frac{|N(i)\bigcap N(j)|}{\sqrt{k_ik_j}}$. 资源分配指标定义[37] $S^{\mathrm{RA}}_{ij}=\sum_{e\in N(i)\bigcap N(j)}\frac{1}{k_e}$. 上述指标能够给出尚未连接的节点间的相似度, 在链路预测 (Link predication) 方面获得了有效的应用[30]. 不难发现, 局部相似性指标紧紧围绕节点的邻居节点来定义相似度, 这里没有列出的局部相似性指标也有类似特性[30].

由于 $V_i\notin N(i)$, 因此上述局部相似性指标对于本来就存在连接的节点的相似性有低估倾向. 以图 1 所示 Karate 网络中的邻居节点 V_{18} 和 V_{21} 为例, 由于 $N(18)=\{V_{19},V_{21},V_{25}\}$, $N(21)=\{V_3,V_{17},V_{18},V_{34}\}$, 因此 $N(18)\bigcap N(21)=\phi$, $|N(18)\bigcap N(21)|=0$, 于是 $S^{\mathrm{Jaccard}}_{18,21}=S^{\mathrm{Salton}}_{18,21}=S^{\mathrm{RA}}_{18,21}=0$. 两个相邻节点的相似度为 0. 因此, 上述指标对于链路预测可能合适, 但用于社区结构发现则在相当大程度上是不恰当的. 因此, 我们需要在相似性度量中引入星型邻域子图, 以修正对已存在连接的低估倾向. 以上述点对为例, 本文提出的相似性度量给出 $S_{18,21}=0.58$, 给二者相似性以很高的评价 (与其他点对相似性的比较见图 2).

从图 2 还可以看出, 相似度矩阵具有稀疏特性, 许多不具备直接连接的点对被赋予零相似度, 这在大型网络中有更加显著的体现, 这是全局性相似性度量无法做到的. 以模块度相似性为例, 任意两个节点都具有非零相似性 $A_{ij}-\frac{k_ik_j}{2m}$, Katz 指标等也有类似倾向, 随着网络规模增大, 相似性矩阵快速膨胀, 大大限制了能够处理的网络规模.

通过上面的讨论, 由于现有全局相似性度量有较高计算复杂度以及局部相似性度量对直接连接点对相似性的低估倾向, 使得二者均不适用于复杂网络社区结构发现. 而本文提出的局部相似性度量具有较低的计算开销, 同时也修正了传统相似性度量的偏性, 因而较好地刻画了网络节点的相似性.

2　基于层次聚类的社区结构发现

我们采用与 CNM 算法类似的策略, 开始时刻将每个节点归入单独的划分, 然后每次迭代时选择相似性最大的两个划分进行合并, 直至形成一个单一划分, 合并过程形成一个划分序列, 将模块度取值最大的划分输出作为社区发现的结果.

上一节讨论的相似性指标 (2) 度量了节点间的相似性, 它使得连接越多、连接的排它性越强的点对具有较高的相似度, 可等价地认为相似性高的点对具有较近的 "距离", 这里的距离是指相似性指标的一个减函数, 它甚至可以不必取值为正. 以下称相似性指标的减函数为节点间的广义距离. 我们将在传统的 Ward 聚类[38] 的基础上, 通过将传统的欧氏距离替换为广义距离进行社区结构发现.

2.1　Ward 层次聚类

设 $\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n$ 是欧氏空间的 n 个数据, 层次聚类时将其按照欧氏距离逐次合并, 最终形成聚类树的一个过程. 初始时刻每个数据点单独作为一个划分, 即 $C_i=\{\boldsymbol{x}_i\},i=1,\cdots,n$, 每个划分的中心为数据本身, 即 $\boldsymbol{c}_i=\boldsymbol{x}_i,i=1,\cdots,n$. 聚类过程将逐

步合并这些划分, 设待合并的划分为 C_i 和 C_j, 合并形成的新划分记为 $C_{i\bigcup j}$, 其中心

$$\boldsymbol{c}_{i\bigcup j}=\frac{N_i\boldsymbol{c}_i+N_j\boldsymbol{c}_j}{N_i+N_j} \tag{3}$$

其中, $N_i=|C_i|$ 表示划分 C_i 中的数据个数. 每个划分中的所有数据到其中心的距离反映了数据聚集的紧密程度, 记

$$W(C_i,\boldsymbol{c}_i)=\sum_{\boldsymbol{x}_k\in C_i}d(\boldsymbol{x}_k,\boldsymbol{c}_i) \tag{4}$$

这里 $d(\cdot,\cdot)$ 是向量间的欧氏距离. 令 $W(C,\boldsymbol{c})=\sum_{k=1}^{c}W(C_k,\boldsymbol{c}_k)$, 则 $W(C,\boldsymbol{c})$ 的取值越小, 划分 $C=\{C_1,\cdots,C_c\}$ 对数据 $\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n$ 的分割就越有效. 在 Ward 聚类合并划分的过程中, 总是选取使得 $W(C,\boldsymbol{c})$ 增加最少的两个划分进行合并. 合并 C_i 和 C_j 为 $C_{i\bigcup j}$ 得到的划分 C' 为将 C_i 和 C_j 从划分 C 中减去, 替换为 $C_{i\bigcup j}$ 的结果, 即

$$C'=C\backslash\{C_i,C_j\}\bigcup\{C_{i\bigcup j}\}$$

则 $W(C',\boldsymbol{c}')-W(C,\boldsymbol{c})=W(C_{i\bigcup j},\boldsymbol{c}_{i\bigcup j})-W(C_i,\boldsymbol{c}_i)-W(C_j,\boldsymbol{c}_j)$ 反映了 C_i 和 C_j 的距离, 称为 Ward 距离, 记为 $dw(C_i,C_j)$. 它的一个等价表达式为[38]

$$dw(C_i,C_j)=\frac{N_iN_j}{N_i+N_j}d(\boldsymbol{c}_i,\boldsymbol{c}_j) \tag{5}$$

Ward 层次聚类就是每次合并 Ward 距离最小的两个划分, 并按照式 (3) 更新划分中心. 由于初始时刻 $W(C,\boldsymbol{c})=0$, 每次合并 $W(C,\boldsymbol{c})$ 后的 $W(C,\boldsymbol{c})$ 取值都会增加 $dw(C_i,C_j)$, 因此 $W(C,\boldsymbol{c})$ 被形象地称为聚类数的高度 (Height). 图 3 展示了采用上述方法对 Karate 网络进行分析得到的聚类树, 每个节点的纵坐标都反映了其左右子树对应的社区划分合并后 $W(C,\boldsymbol{c})$ 的取值, 其下竖线的高度值即其左右子树对应划分的 Ward 距离.

2.2 Ward 聚类的推广

由于 Ward 聚类通过最小化数据划分间的 Ward 距离来聚类的, 如果能够直接获取 d_{ij}, 甚至无需知道 $\boldsymbol{x}_i$ 的具体取值. 这样, 聚类分析可以对任何具有距离度量的对象进行, 无需限于欧氏空间的数据. 在广义的 Ward 聚类中, 只需知道 n 个抽象对象 $\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n$ 的两两距离 d_{ij}, 并将划分 C_i 的中心看作抽象的对象 $\boldsymbol{c}_i$, 并按照

$$d(\boldsymbol{x}_j,\boldsymbol{c}_i)=\frac{1}{N_i}\sum_{\boldsymbol{x}_k\in C_i}d_{jk} \tag{6}$$

计算 $\boldsymbol{c}_j$ 和数据点 $\boldsymbol{x}_i$ 的距离. 式 (5) 中 $d(\boldsymbol{c}_i,\boldsymbol{c}_j)$ 如下计算:

$$d(\boldsymbol{c}_i,\boldsymbol{c}_j)=\frac{1}{N_iN_j}\sum_{\boldsymbol{x}_k\in C_i,\boldsymbol{x}_l\in C_j}d_{kl} \tag{7}$$

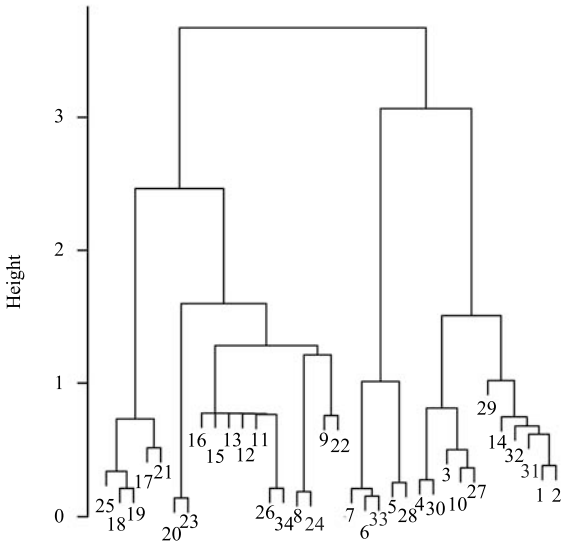


图 3 Karate 网络聚类树状图

Fig. 3 Cluster dendrogram for the Karate network

聚类过程和欧氏空间数据聚类几乎一致, 初始时刻将每个抽象对象视作一个划分, 然后逐次合并 Ward 距离最小的划分, 按照式 (3) 和式 (5) 更新抽象划分中心及聚类树的高度, 涉及相应的距离计算时用式 (6) 和式 (7) 替换, 合并完成时输出聚类树. 图 3 是将推广的 Ward 算法应用于 Karate 网络的结果, 从中可以看出 Ward 距离增加最快的位置对应 4 个划分, 网络的模块度为 0.4111604, 已十分接近旋转玻璃模型[25] 报告的最优结果 0.4197896. 采用广义距离的 Ward 聚类算法对距离的要求相当宽松, 它甚至不要求 d_{ij} 取值范围为正数. 容易看到, Ward 距离是对象间距离的线性组合, 将距离矩阵整体平移后进行聚类完全不影响聚类的结果. 例如取 $d_{ij}^a=d_{ij}-a$, 其中 a 为任意实数, 则 $dw^a(C_i,C_j)=dw(C_i,C_j)-a$, 因此在层次聚类的任意时刻, 所有划分间 Ward 距离的序关系并未发生改变, 我们将得到相同的聚类结果.

2.3 基于广义 Ward 聚类的社区结构发现

根据上一小节的分析, 只需提供网络节点间的距离信息就可对其进行层次聚类, 然后在聚类树合适的位置进行切割便可得到社区结构. 节点间的距离可以取为相似性度量的单调减函数, 使得相似性

越高的节点距离越近. 因此, 不妨令

$$D_{ij} = 1 - \sqrt{S_{ij}} \tag{8}$$

表示 V_i 和 V_j 间的距离. 根据前面的分析, 我们完全可以取 $D_{ij} = -\sqrt{S_{ij}}$ 表示 V_i 和 V_j 的 "距离", 对社区发现的结果不会有任何影响. 这里将其变换为正数, 只是为了符合距离的传统定义.

算法维持一个 Ward 距离矩阵, 初始时刻其取值由式 (8) 给出. 当合并 C_i 和 C_j 时, 为节约存储将 C_j 并入 C_i, 对 $\forall k \neq i, k \neq j$ 更新矩阵 D 的 i 行如下:

$$D_{ik} = \frac{(N_i + N_k)D_{ik} + (N_j + N_k)D_{jk} - N_k D_{ij}}{N_i + N_j + N_k} \tag{9}$$

然后删除 D 的 j 行和 j 列. 由于对称性矩阵 D 的 i 列与其 i 行取值相同, 可不予存储. 将聚类过程中得到的划分序列记为 $C_{(n)}, C_{(n-1)}, \cdots, C_{(1)}$, 其中 $C_{(k)}$ 表示该划分将节点集合 V 分为 k 个子集合. 最后输出使得模块度取值最大的划分 C_m 作为社区发现的结果. 上述过程用伪码描述如下:

算法 1.

1) 初始化: $C = \{C_1, \cdots, C_n\}$, 其中 $C_i = \{V_i\}$. D_{ij} 按照式 (8) 初始化. $C_{(n)} \leftarrow C$.

2) DO

3) 求矩阵 D 下三角部分最小值, 得 $D_{ij} = \min D$.

4) for $k = 1 : |C|$

5) if $k \notin \{i, j\}$

6) if $k < i$

7) $D_{ik} \leftarrow \frac{(N_i+N_k)D_{ik}+(N_j+N_k)D_{jk}-N_k D_{ij}}{N_i+N_j+N_k}$;

8) else

9) $D_{ki} \leftarrow \frac{(N_i+N_k)D_{ik}+(N_j+N_k)D_{jk}-N_k D_{ij}}{N_i+N_j+N_k}$;

10) end if

11) end if

12) end for

13) $C \leftarrow C \setminus \{C_i, C_j\} \bigcup \{C_i \bigcup C_j\}$;

14) $C_{(|C|)} \leftarrow C$;

15) 删除 D 的 j 行和 j 列;

16) UNTIL $|C| = 1$

17) $C_{(\text{tree})} \leftarrow \{C_{(n)}, \cdots, C_{(1)}\}$;

18) $C_m \leftarrow \arg\max_{C \in C_{(\text{tree})}} Q(C)$;

19) 返回 C_m 作为社区划分结果.

不难看出上述算法涉及矩阵循环, 因此具有的复杂度为 $\mathrm{O}(n^2)$, 慢于 CNM 的 $\mathrm{O}(n\log^2 n)$. GN 算法的时间复杂度为 $\mathrm{O}(nm)$[28], 由于网络的边数通常大于顶点数, 即 $m > n$, 因此本文方法实际时间复杂度低于 GN 算法. 注意到矩阵 S 的初始化只需每个节点附近的局部搜索, 其实际执行速度明显快于 GN 算法.

3 实验

为了测试本文所提算法的性能, 我们利用计算机合成网络和真实世界网络进行了社区发现实验. 合成网络按照已知的社区结构产生, 可以测试社区发现算法得出的社区结构的有效性. 实验在一台配置了 Intel i5 430 M 2.27 GHz CPU 和 2 GB DDR3 内存的笔记本上进行, 操作系统为 Windows XP, 使用 R 语言编程实现. 进行了与常见典型社区发现算法的比较实验, 参与对比的算法包括:

1) CNM 算法[29]: 采用快速贪婪规则合并划分得出社区结构, 是凝聚型算法的典型代表, 记为 Fastgreedy;

2) GN 算法[28]: 采用逐次移除介数中心性最大边的方法得出社区划分, 是分裂型算法的典型代表, 记为 Betweenness;

3) 标签传播算法 (Label propagation)[39]: 初始时刻每个点都有唯一的标签, 然后按照随机顺序轮询, 当前节点在其邻居中选择最流行的标签替换自己原有标签, 若这样的标签有多个则随机地选择其一. 它是线性时间复杂度算法的典型, 记为 Label;

4) 模块度矩阵谱聚类[26]: 以模块度矩阵最大特征根的对应特征向量来逐次将网络节点进行二分, 是谱聚类的典型代表, 记为 Eigenvector;

5) 旋转玻璃模型[25]: 以 q-状态 Potts 模型来表示网络节点, 相邻节点间存在相互作用, 这种作用决定着系统的能量状态, 通过寻求该物理系统能量最低的状态来进行社区结构发现. 它含有参数 q, 表示算法允许的最大可能的社区数目, 默认取值为 25. 它是高精度复杂算法的典型, 记为 Spinglass;

6) 本文提出的基于局部相似性度量的社区结构发现方法, 即算法 1, 记为 Similarity.

3.1 合成网络

我们采用 Newman 建议的社区发现基准测试网络[40], 该网络有 128 个节点, 分为 4 组, 每组 32 个节点. 我们在每个节点随机生成 16 个连接, 其中 z_o 个连接组间节点, 剩下 $16 - z_o$ 个连接组内节点. 令 $z_o \in \{1, \cdots, 12\}$, 当 $z_o = 1$ 时组内连接最密集, 社区结构最明显; 当 $z_o = 12$ 时组内连接最稀疏, 社区结构最模糊. 文献常常只考虑 $z_o \leq 8$ 的情形[1, 17, 28], 因为 $z_o = 8$ 时意味着每个点在组内外的连接数相同. 实际上, 由于组内仅有 32 个点而组外有 96 个点, 因此组内连接密度仍然高于组间连接. 因此, 我们报告了 $z_o \leq 12$ 的情形, 当 $z_o = 12$ 时组内连接和组间连接密度相同, 社区结构是最模糊的.

设原始的社区划分为 C_o, 算法得出的社区划分

为 C_e, 用归一化互信息 (Normalized mutual information, NMI)[41] 来度量两个划分的差别. 其定义为

$$NMI(C_o, C_e) = \frac{H(C_o) + H(C_e) - H(C_o, C_e)}{\sqrt{H(C_o)H(C_e)}} \tag{10}$$

其中, $H(C)$ 表示划分 C 的香农信息熵. 当 C_e 和原始的社区划分 C_o 完全一致时, $NMI(C_o, C_e) = 1$; 当 C_e 和 C_o 完全不同时, $NMI(C_o, C_e) = 0$.

由于 Label 和 Spinglass 具有随机性, 因此我们共进行了 10 次实验, 报告平均 NMI 取值. 为了使实验可重复, 将随机数种子分别固定为 $1, \cdots, 10$, 每次让 z_o 从 1 变化到 12, 产生 12 个网络, 用不同的方法进行社区结构发现, 实验结果见图 4. 当 $z_o \in \{1, 2, 3, 4\}$ 时 Similarity 和 Betweenness 都能够完全恢复原始社区划分, NMI 取值为 1.0; 带有随机性的 Spinglass 方法得到的 NMI 在 1.0 附近有极小的波动; Fastgreedy 的表现稍差, 在 $z_o = 4$ 时得到的 NMI 约为 0.979; Eigenvector 和 Label 的表现较差, 在 $z_o = 4$ 时其 NMI 分别为 0.939 和 0.910. 随着 z_o 继续增大, 算法的表现逐渐下降, 下降最快的是 Label, 当 $z_o = 5$ 时其 NMI 取值尚在 0.7 左右, 但当 $z_o = 6$ 时其 NMI 急速下降到 0.317, 远远落后于其他算法. 因此, Label 虽然有线性时间复杂度, 但其高效率是以极大牺牲社区发现精度为代价的. 从 $z_o = 7$ 开始, NMI 取值急速下降的有三个算法, 分别是 Betweenness, Fastgreedy 及 Eigenvector, 这时只有 Spinglass 和 Similarity 具有高于 0.9 的 NMI 取值. 当 $z_o = 8$ 时, 每条连接以 50 % 的概率连接组间节点, 社区结构十分模糊, 但 Similarity 的 NMI 仍然为 0.757, 仅次于 Spinglass 的 NMI 取值 0.974; Label 的 NMI 取值为 0.0, Betweenness 得到的 NMI 为 0.455, Fastgreedy 以及 Eigenvector 得到的 NMI 约为 0.6. 当 $z_o = 9$ 时, 仅 Spinglass 的 NMI 维持在 0.9 以上, 其余除 Label 外的所有方法得到的 NMI 均为 0.3 左右. 当 z_o 接近 12 时, 仅 Betweenness 和 Eigenvector 有较好表现, 得益于其所采用的全局性度量, 其余方法的 NMI 均接近 0. 注意此时 Spinglass 并未从其极高的计算量获益, 它与 Fastgreedy 及 Similarity 等低计算量算法的表现几乎相同. 整体上看, 当社区结构较为明显时, Spinglass 和 Similarity 有最好的表现, 当社区结构较为模糊时, Betweenness 和 Eigenvector 有较好的表现. 但从计算复杂度来看, 除 Similarity 仅仅利用网络的局部信息外, 其余三个表现较好的方法都利用了网络的全局性信息, 具有较高的计算复杂度, 尤其是 Betweenness 和 Spinglass, 其执行时间远远超过 Similarity; Eigenvector 得益于只需要计算最大特征根, 可以利用 ARPACK[42] 等线性代数软件包加速, 其计算复杂度稍低.

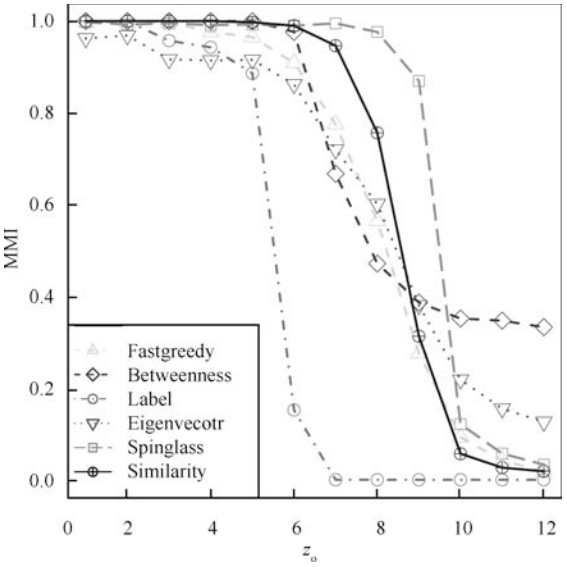


图 4 算法精度比较

Fig. 4 Comparison of different algorithms

3.2 真实世界网络

社区发现的真正意义在于真实网络的社区结构发现, 通过在合成的网络上验证我们方法的有效性, 还需要进一步在真实网络上测试算法的性能. 真实网络的社区发现比合成网络更具有挑战性, 由于不能预知其社区结构, 因此只能从其返回的社区划分的模块度来进行比较.

我们使用的网络包括空手道俱乐部网络 (Karate)[31]、海豚网络 (Dolphins)[43]、爵士音乐家合作网络 (Jazz)[44]、美国大学生 2000 赛季美式足球联赛网络 (Football)[17]、美国航空网络 (Us.air, http://vlado.fmf.uni-lj.si/pub/networks/data/)、2004 年美国大选政治博客网络 (Polblogs)[45]、拟南芥 (C. elegans) 新陈代谢网络 (Cel)[46]、电子邮件网络 (Email)[7]、蛋白质作用网络 (Ppi)[46] 以及网络科学家论文合作网络 (Netscience)[5]. 为了使参与对比的方法可用, 部分网络在进行实验前进行了一定的处理. 其中 Netscience 和 Ppi 网络是不连通的, 因此取其最大连通分支进行实验; Polblogs 网络是有向的, 实验时并未对连接方向加以区分; Cel 网络是加权网络, 实验时忽略了连接的权重; Football 网络存在少数重复连接, 因为不同的球队在赛季中可能比赛多次, 实验时忽略重复连接. 实验结果见表 1.

表 1 第 1 列展示了网络的基本统计数据, Nodes 对应网络的节点数, Edges 对应网络的连接数. 对每种方法, 报告了其社区发现的模块度、社区数目

及执行时间, 例如 Karate 网络用 Similarity 算法得出的模块度为 0.41, 社区数为 4, 运行时间为 0.01 秒, 在表 1 中表示为 0.41/4/0.01 s. 考虑到 Spinglass 和 Betweenness 算法较高的复杂度对应着相对较好的社区发现结果, 表 1 对每个网络将模块度取值前两位用粗体强调, 以下称相应的算法取得领先结果. 总体而言, Betweenness 和 Spinglass 取得领先的次数最多, 分别达到 8 次和 9 次, 这得益于它们的全局性度量. Betweenness 基于所有点对的最短路径, Spingass 需要计算整个网络的最低能量状态, 因此均具有极高的计算复杂度. Similarity 采用局部相似性度量, 取得了 5 次领先, 尤其是在网络 Us.air 的社区发现, Betweenness 和 Spinglass 均给出了接近零的模块度, Similarity 仍然取得不错的表现. Eigenvector 和 Label 相对来说表现最差, Fastgreedy 相对较好, 有 4 次领先. 另外, 从社区数目的角度, Betweenness 倾向于给出较多数目的社区, 如将 Polblogs 网络分为 205 个网络, 显著高于其余方法, Similarity 得出的社区数为 2, 但二者的模块度取值非常接近. 从该网络的实际特征来说, 分为两个社区是合理的, 它对应 Polblogs 的两大阵营, 即民族党支持者和自由党支持者[45]. 此外, 仅 Label 给出较为接近的社区数, 其余方法给出的社区数均在 10 个以上, 显示相应的方法有过拟合 (Over fitting) 倾向. 运行时间方面, 随着网络节点和边的增加, 各个算法的运行时间出现了显著的增长. 总的看来, Fastgreedy 和 Label 比较适宜处理大型网络, 但 Label 获得的模块度得分偏低. Eigenvector 由于采用了 ARPACK 加速特征根的计算[26, 42], 在处理规模较大网络时消耗时间增加较慢. 但是, 我们可以从表 1 看出, 该方法获得的模块度取值偏低, 其速度的增加是以社区发现的性能为代价的. 消耗时间最多的是 Similarity 和 Betwenness, 但 Similarity 在大部分网络上的时间消耗显著少于 Betweenness. Betweenness 在 Ppi 和 Netscience 两个网络上的时间消耗显著小于 Similarity, 一个可能的原因是这两个网络在移除介数中心性较大的若干连接后迅速分裂为较小的部分, 使得后续的全局介数中心性的计算复杂度大为降低. 这点可以从网络的模块度取值间接反映出来, 以 Netscience 为例, 除 Eigenvector 得出的模块度为 0.78 外, 其余方法得到的社区结构的模块度取值均超过 0.8, 表明该网络的社区结构非常显著. 此外, Spinglass 的运行时间随网络规模的增大以较慢的速度增加, 一个重要的原因是其预先设定了最大的可能社区数目 (在本文的实验中, Spinglass 最大可能的社区数设为固定的 25. 如果将 25 改为较大的数目, 将显著地增加其运行时间. 限于篇幅, 这里省略相应结果).

4 总结与展望

本文定义了一种新的网络节点相似性度量, 它只用到网络节点的局部连接信息, 即星型邻域子图. 传统的全局性相似性度量, 如基于随机游走平均通勤时间的相似性度量[34] 涉及矩阵求逆运算, 使得相似性的计算具有很高的计算复杂度, 另一个全局性度量 Kata[32] 指标涉及额外的人为参数, 这些特性在本文的度量中都被成功地避免了. 传统的局部相似性度量, 如 Jaccard 指标[35] 及资源分配指标[37] 等倾向于低估存有连接点对间的相似性, 因而适用于链路预测[37] 但不适用于网络社区结构发现. 本文提出的相似性指标很好地避免了上述问题, 是适于进行社区结构发现的局部相似性指标.

通过对 Ward 层次聚类算法的分析, 发现其优化目标是待处理数据距离的线性函数, 进而提出将聚类对象拓展到具有两两距离的抽象对象. 通过网络节点相似性度量的减函数来定义节点间的广义距

表 1 各算法社区发现结果

Table 1 Community detected by different algorithms

Network	Nodes/Edges	Fastgreedy	Betweenness	Label	Eigenvector	Spinglass	Similarity
Karate	34/78	0.38/3/0.00 s	0.40/5/0.01 s	0.37/3/0.00 s	0.38/5/0.00 s	**0.42**/4/1.02 s	**0.41**/4/0.01 s
Dolphins	62/159	0.50/4/0.00 s	**0.52**/5/0.05 s	0.51/4/0.00 s	0.49/6/0.02 s	**0.53**/5/1.42 s	**0.52**/4/0.04 s
Jazz	198/2 742	**0.44**/4/0.01 s	**0.41**/39/34.31 s	0.28/2/0.00 s	0.35/8/0.13 s	**0.44**/5/9.05 s	0.40/4/0.88 s
Football	115/613	0.55/6/0.00 s	**0.60**/10/0.94 s	**0.58**/12/0.00 s	0.45/13/0.04 s	**0.60**/10/1.59	**0.60**/10/0.22 s
Us.air	332/2 126	0.19/4/0.00 s	0.14/124/30.28 s	0.00/3/0.00 s	**0.27**/16/0.80 s	0.03/24/19.12 s	**0.30**/10/3.16 s
Polblogs	1 222/16 714	**0.43**/10/0.25 s	**0.42**/205/6 374 s	**0.43**/3/0.05 s	0.27/62/5.58 s	**0.43**/16/157.69 s	0.38/2/111.81 s
Cel	453/2 025	0.39/9/0.00 s	**0.40**/38/61.31 s	0.33/8/0.00 s	0.32/28/0.99 s	**0.44**/10/19.62 s	0.37/14/6.81 s
Email	1 133/5 451	0.49/16/0.06 s	**0.53**/61/838.03 s	0.01/4/0.02 s	0.42/45/2.32 s	**0.58**/12/76.62 s	0.45/16/122.69 s
Ppi	1 458/1 948	**0.82**/32/0.00 s	**0.81**/39/89.72 s	0.74/168/0.02 s	0.72/115/1.17 s	**0.81**/25/59.03 s	0.79/61/238.89 s
Netscience	379/914	**0.84**/19/0.00 s	**0.84**/18/1.95 s	0.81/38/0.00 s	0.78/33/0.54 s	**0.84**/21/11.65 s	**0.82**/17/3.92 s

离, 设计了推广的 Ward 层次聚类算法用于网络社区结构发现. 该算法对距离度量的要求极低, 甚至不要求距离必须取正值, 只需将 "距离" 取为节点相似性度量的单调减函数即可. 此外, 对于具有稀疏特性的网络, 本文提出的相似度矩阵也是稀疏的, 可仅存储非零元素以减少内存消耗.

这里提出的相似性度量还可从更多的角度服务于网络数据分析. 例如可以考虑对相似性度量矩阵 S 的谱分析, 考察它对于揭示网络结构的作用, 矩阵 S 也许是现有 5 种谱分析矩阵[27] 的有力竞争者. 另外, 由于本文在 Ward 聚类中采用矩阵的存储方法, 难以适用于大型网络, 进一步可以考虑在 Ward 聚类中的 S 矩阵改用高效的数据结构如堆 (Heap)[29] 来存储, 取代本文采用的矩阵存储方式, 然后探索基于堆结构的社区结构发现方法. 堆结构使得矩阵最大值查询为常数时间, 矩阵元素更新或删除为对数时间, 从而将算法的执行效率提高到 $O(n\log^2 n)$[29], 使之更加适合处理大型网络.

此外, 本文主要关注基于网络顶点相似性的聚类. 最近的研究表明, 基于边的分析同样可以揭示网络的社区结构[47]. 基于边的聚类允许网络的顶点属于多个社区, 形成所谓的有重叠 (Overlapping) 的社区结构[48], 这也将是我们未来可能的工作方向.

References

1 He Dong-Xiao, Zhou Xu, Wang Zuo, Zhou Chun-Guang, Wang Zhe, Jin Di. Community mining in complex networks — clustering combination based genetic algorithm. *Acta Automatica Sinica*, 2010, **36**(8): 1160−1170
(何东晓, 周栩, 王佐, 周春光, 王喆, 金弟. 复杂网络社区挖掘 — 基于聚类融合的遗传算法. 自动化学报, 2010, **36**(8): 1160−1170)

2 Newman M E J. The structure and function of complex networks. *SIAM Review*, 2003, **45**(2): 167−256

3 Scheffer M. Complex systems: foreseeing tipping points. *Nature*, 2010, **467**(7314): 411−412

4 Newman M E J. *Networks: an Introduction*. New York: Oxford University Press. 2010

5 Newman M E J. Scientific collaboration networks: I. network construction and fundamental results. *Physical Review E*, 2001, **64**(1): 016131

6 Zeng J, Cheung W K, Li C H, Liu J M. Coauthor network topic models with application to expert finding. In: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. Toronto, Canada: IEEE, 2010. 366−373

7 Guimerà R, Danon L, Díaz-Guilera A, Giralt F, Arenas A. Self-similar community structure in a network of human interactions. *Physical Review E*, 2003, **68**(6): 065103

8 Costa L F, Oliveira O, Travieso G, Rodrigues F A, Villas Boas P, Antiqueira L, Viana M P, Correa Rocha L. Analyzing and modeling real-world phenomena with complex networks: a survey of applications. *Advances in Physics*, 2011, **60**(3): 329−412

9 Watts D J, Strogatz S H. Collective dynamics of "small-world" networks. *Nature*, 1998, **393**(6638): 440−442

10 Leij M V, Goyal S. Strong ties in a small world. *Review of Network Economics*, 2011, **10**(2): 1−21

11 Barabási A L, Albert R. Emergence of scaling in random networks. *Science*, 1999, **286**(5439): 509−512

12 Karsai M, Kivelä M, Pan R K, Kaski K, Kertész J, Barabási A-L, Saramäki J. Small but slow world: how network topology and burstiness slow down spreading.*Physical Review E*, 2011, **83**(2): 025102

13 Arenas A, Cabrales A, Danon L, Díaz-Guilera A, Guimerà R, Vega-Redondo F. Optimal information transmission in organizations: search and congestion. *Review of Economic Design*, 2010, **14**(1): 75−93

14 Hu Y Q, Wang Y G, Li D Q, Havlin S, Di Z R. Possible origin of efficient navigation in small worlds. *Physical Review Letters*, 2011, **106**(10): 108−701

15 Arenas A, Diaz-Guilera A, Pérez-Vicente C J. Synchronization reveals topological scales in complex networks. *Physical Review Letters*, 2006, **96**(11): 114102

16 Arenas A, Diaz-Guilera A, Perez-Vicente C J. Synchronization processes in complex networks, *Physica D: Nonline or Phenomena*, 2006, **224**(1−2): 27−34

17 Girvan M, Newman M E J. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 2002, **99**(12): 7821−7826

18 Wasserman S, Faust K. *Social Network Analysis*, Cambridge: Cambridge University Press, 1994.

19 Kwak H, Lee C H, Park H, Moon S. What is Twitter, a social network or a news media? In: Proceedings of the 19th International Conference on World Wide Web. Raleigh, North Carolina, USA: ACM, 2010. 591−600

20 Fortunato S. Community detection in graphs. *Physics Reports*, 2010, **486**(3−5): 75−174

21 Wang X H, Jiao L C, Wu J S. Phase synchronization on small-world networks with community structure. *Chinese Physics B*, 2010, **19**(2): 020501

22 Newman M E J. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 2006, **103**(23): 8577−8582

23 Brandes U, Delling D, Gaertler M, Gorke R, Hoefer M, Nikoloski Z, Wagner D. On modularity clustering. *IEEE Transactions on Knowledge and Data Engineering* 2008, **20**(2): 172−188

24 Duch J, Arenas A. Community detection in complex networks using extremal optimization. *Physical Review E*, 2005, **72**(2): 027104

25 Reichardt J, Bornholdt S. Statistical mechanics of community detection. *Physical Review E*, 2006, **74**(1): 016110

26 Newman M E J. Finding community structure using the eigenvectors of matrices. *Physical Review E*, 2006, **74**(3): 036104

27 Shen H W, Cheng X Q. Spectral methods for the detection of network community structure: a comparative analysis. *Journal of Statistical Mechanics: Theory and Experiment*, 2010, **10**: P10020

28 Newman M E J, Girvan M. Finding and evaluating community structure in networks. *Physical Review E*, 2004, **69**(2): 026113

29 Clauset A, Newman M E J, Moore C. Finding community structure in very large networks. *Physical Review E*, 2004, **70**(6): 066111

30 Lv L Y, Zhou T. Link prediction in complex networks: a survey. *Physica A*, 2011, **390**(6): 1150−1170

31 Zachary W W. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 1977, **33**(4): 452−473

32 Katz L. A new status index derived from sociometric analysis. *Psychmetrika*, 1953, **18**(1): 39−43

33 Leicht E A, Holme P, Newman M E J. Vertex similarity in networks.*Physical Review E*, 2006, **73**(2): 026120

34 Fouss F, Pirotte A, Renders J M, Saerens M. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 2007, **19**(3): 355−369

35 Hamers L, Hemeryck Y, Herweyers G, Janssen M, Keters H, Rousseau R, Vanhoutte A. Similarity measures in scientometric research: the Jaccard index versus Salton′s cosine formula. *Information Processing & Management*, 1989, **25**(3): 315−318

36 Salton G, McGill M J. *Introduction to Modern Information Retrieval*. New York: McGraw-Hill, 1983

37 Zhou T, Lv L Y, Zhang Y C. Predicting missing links via local information. *The European Physical Journal B: Condensed Matter and Complex Systems*, 2009, **71**(4): 623−630

38 Mirkin B. *Core Concepts in Data Analysis: Summarization, Correlation and Visualization*. London: Springer, 2011. 283−313

39 Raghavan U N, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 2007, **76**(3): 036106

40 Newman M E J. Fast algorithm for detecting community structure in networks. *Physical Review E*, 2004, **69**(6): 066133

41 Danon L, Diaz-Guilera A, Duch J, Arenas A. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005, **2005**(9): P09008

42 Sorensen D C. Implicit application of polynomial filters in a κ-step arnoldi method. *SIAM Journal of Matrix Analysis and Applications*, 1992, **13**(1): 357−385

43 Lusseau D. The emergent properties of a dolphin social network. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 2003, **270**(2), 186−188

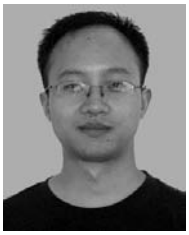
44 Gleiser P, Danon L. Community structure in jazz. *Advances in Complex Systems*, 2003, **6**(4): 565−573

45 Adamic L A, Glance N. The political blogosphere and the 2004 US election: divided they blog. In: Proceedings of the 3rd International Workshop on the Weblogging Ecosystem, New York, USA: ACM, 2005. 36−43

46 Jeong H, Mason S, Barabási A L, Oltvai Z N. Lethality and centrality in protein networks. *Nature*, 2001, **411**(6833): 41−42

47 Ahn Y Y, Bagrow J P, Lehmann S. Link communities reveal multiscale complexity in networks. *Nature*, 2011, **466**(7307): 761−764

48 Gregory S. Fuzzy overlapping communities in networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2011, **2**: P02017



刘 旭 国防科学技术大学理学院数学与系统科学系博士研究生. 2007 年获国防科学技术大学应用数学硕士学位. 主要研究方向为复杂海量信息处理. 本文通信作者.

E-mail: liuxuwork@gmail.com

(**LIU Xu** Ph. D. candidate in the Department of Mathematics and Systems Science, Science College, National University of Defense Technology. He received his master degree from National University of Defense Technology in 2007. His main research interest is complex large-scale data and information processing. Corresponding author of this paper.)



易东云 国防科学技术大学教授. 主要研究方向为智能信息系统.

E-mail: dongyun.yi@gmail.com

(**YI Dong-Yun** Professor at National University of Defense Technology. His main research interest is intellecture information systems.)