加权复杂网络社团的评价指标及其发现算法分析*

吕天阳[1)2)3)†] 谢文艳[3)] 郑纬民[1)] 朴秀峰[3)]

1) (清华大学计算机科学与技术系, 北京 100084)

2) (审计署审计科研所, 北京 100830)

3) (哈尔滨工程大学计算机科学与技术学院, 哈尔滨 150001)

(2012 年 3 月 31 日收到; 2012 年 5 月 7 日收到修改稿)

节点的聚集现象是复杂网络的重要特性. 以往研究主要发现无权复杂网络中的社团, 较少涉及加权网络的社团发现. 由于加权网络的复杂性远高于无权网络, 一般认为加权网络的社团发现是一个较难的问题. 本文基于统一的数据基础, 从社团评价指标的有效性和现有算法的效果两个角度开展研究. 首先, 总结了加权网络三种常见的社团评估指标, 并在社团大小、密度和局域特点均不同的模拟数据集上分析指标的有效性; 其次, 针对 5 个数据集, 分析现有的 3 种加权复杂网络社团发现算法的效果. 研究表明: 上述指标无论在评价最基本的社团结构, 还是在分析结构复杂的社团时都有较大缺欠; 现有的加权网络社团发现算法的泛化能力不强.

关键词: 复杂网络, 社团发现, 聚集系数, 模块度

PACS: 05.65.+b

1 引 言

1998 年 Watts 和 Strogatz[1] 在 Nature 上发表复杂网络模型成功解释了 “小世界现象” 后, 复杂网络成为诸多领域的基础模型, 用于理解各领域研究对象间复杂的拓扑关系和动力学行为, 如互联网 [2]、病毒传播、社会学 [3] 等领域的研究. 近年来, Science, Nature 等都出现了相关研究成果, 提出了无标度模型 [2,4] 等.

复杂网络的社团发现是复杂网络研究的一个重要方面, 用于理解网络的拓扑结构、挖掘网络的潜在意义及预测网络的行为等. 一般将分析所得的数据簇 (cluster) 称为群 (group) 或社团 (community)[5], 下文统一称为社团. 目前, 复杂网络的社团发现算法研究主要针对无权的复杂网络模型, 提出了 Kernighan-Lin、谱平分算法、GN (Girvon-Newman) 算法、FN 快速 Newman 算法、派系过滤算法等. 对此, 杨博等 [6] 已经做了很好的总结. 程学旗和沈华伟 [7] 对社团结构的研究历程和研究成果进行了总结.

与无权网络相比, 加权复杂网络更能反映实际情况. 例如, 论文的互相引用构成论文作者间的关系网, 作者间互相引用的次数很自然地成为边的权值; 又如, 人与人接触的频度为两者间边的权值, 权值越高病毒在两者间传播的可能性越大. 其中, 网络中边的权也被称为连接的强度, 很多经典研究已经指出了连接强度的重要价值 [3].

由于加权网络的复杂性远高于无权网络, 学者们一般认为加权复杂网络的分析是一个较难的问题. 目前仅对加权复杂网络的建模 [8] 与分析 [9,10] 展开了初步的研究. 在社团发现方面, 一些研究分析单一社团发现算法对特定数据集的效果 [11], 也有文献对几种加权复杂网络的社团发现算法进行简单的综述 [12], 但是缺乏较全面的实验比较. 由于加权网络的复杂性远高于无权网络, 而且两者有着

* 国家科技支撑计划 (批准号: 2009BAH42B02, 2012BAH08B02)、国家自然科学基金 (批准号: 60903080)、中央高校基本科研业务费专项资金 (批准号: HEUCFZ100603) 和黑龙江省教育厅科学技术研究 (批准号: 12513050) 资助的课题.

† E-mail: raynor1979@163.com

明显区别, 因此加权复杂网络中社团的意义、发现方法等都有待深入研究.

对此, 本文力图在统一的测试数据基础上, 分析复杂情况下加权网络中社团的界定, 并比较现有典型的发现方法. 由于社团评估指标是衡量社团发现效果、指导社团发现过程的核心, 本文首先总结整理了现有研究对聚集系数和模块度指标的定义, 同时将强/弱社团的定义拓展到加权网络. 其次, 基于社团大小、密度和局域特点均不同且有大量噪音的模拟数据, 分析上述指标的适用性. 再次, 比较 3 种加权网络社团发现算法分析模拟数据集和 4 种真实数据集的效果, 从而在统一的数据基础上评价现有各方法的优劣.

研究结果表明: 1) 加权复杂网络中社团的准确界定并不容易, 多种聚集系数指标在评价基本社团结构时均存在缺欠, 而且聚集系数、强/弱社团结构、模块度指标并不适用于分析存在复杂局部情况和高噪音的加权网络; 2) 对 5 个数据集实验表明, 现有的典型社团发现算法的泛化能力存在缺欠; 3) 由于社团评价指标的有效性存疑且加权复杂网络准确分类信息的缺乏, 使得客观评价加权网络社团发现效果变得异常困难.

2 加权复杂网络社团的评估指标

社团评估指标是对社团的界定, 是衡量社团发现效果、指导社团发现过程的核心. 常见的无权网络社团评估指标包括: 聚类系数 (clustering coefficient)、强/弱社团 (community in a strong sense and in a weak sense) 和模块度 (modularity). 聚集系数给出网络中各节点的邻居节点间的平均紧密程度. 强/弱社团结构和模块度函数均基于假定: 一个合理的社团内节点的连接强度高于社团外节点的连接强度. 一般而言, 聚集系数主要用来刻画网络节点的局部聚集程度, 但是也有研究者用其衡量网络整体或其子社团的聚集程度 [13]; 模块度函数作为启发式函数指导聚类过程; 强/弱社团则用于衡量所发现的社团是否合理.

由于加权网络社团既要考虑节点间的连通度, 也要考虑边的权值, 因此有必要分析无权网络中社团的评估指标在加权网络中的有效性. 例如, 无权条件下的强/弱社团和模块度的基本假设在加权网络中就不一定成立. 如图 1 中圆形节点和方形节点由于相互联结的强度高, 形成了两个较明显的社团; 但是在不考虑权的情况下, 两社团之间的连通度并不低于每个社团的内部.

为此, 本节首先总结上述聚集系数和模块度函数在加权网络中的计算方法, 并将强/弱社团指标推广到加权网络中.

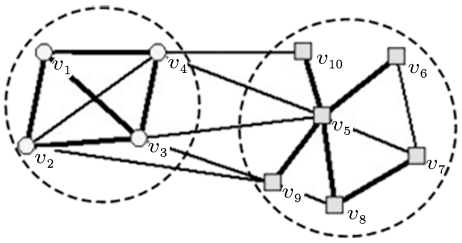

图 1 加权复杂网络社团的示例 (图中粗线权值为 3, 细线权值为 1)

2.1 聚集系数

无权网络中第 i 个节点 v_i 的聚类系数 C_i 的定义如下:

$$C_i = \frac{\sum\limits_{j,k} a_{ij}a_{jk}a_{ki}}{k_i(k_i-1)},\ k_i \neq 0,1, \tag{1}$$

其中, k_i 表示节点 v_i 的度, v_j, v_k 为与 v_i 相连的任意两个节点, a_{ij} 取值 1 或 0, 表示节点 v_i 与 v_j 是否有边相连. 可见, 聚类系数 C_i 实质上等于与节点 v_i 相连的三角形的数量除以与 v_i 相连的三元组的数量 [13].

在定义 (1) 式的基础上, 相关研究给出了加权复杂网络中聚类系数的定义. 其中, 文献 [9] 中总结比较了加权复杂网络聚类系数的 7 种定义. 限于篇幅本文着重分析其中有代表性的 3 种聚类系数, 参见 (2)[14], (3)[15] 和 (4) 式 [16], 分别记为 $C^Z_{w,\,i}$, $C^O_{w,\,i}$ 与 $C^B_{w,\,i}$. 文献 [17—19] 中方法与 (2) 式相似, 文献 [20] 中方法同 (4) 式相似, 本文不再详述.

$$C^Z_{w,i} = \frac{\sum\limits_{j,k} w_{ij}w_{jk}w_{ki}}{\Big(\sum\limits_j w_{ij}\Big)^2 - \sum\limits_j w^2_{ij}}, \tag{2}$$

$$C^O_{w,i} = \frac{\sum\limits_{j,k} (w_{ij}w_{jk}w_{ki})^{1/3}}{k_i(k_i-1)}, \tag{3}$$

$$C^B_{w,i} = \frac{1}{s_i(k_i-1)} \sum_{j,k} \frac{w_{ij}+w_{ik}}{2} a_{ij}a_{jk}a_{ki}, \tag{4}$$

本文中, w_{ij} 表示节点 v_i 与 v_j 间连边的权值, s_i 表示所有与节点 v_i 相连的边的权值之和.

这些加权网络聚集系数试图从不同粒度分析加权复杂网络的聚集程度. 在最小粒度, 能够给出单一节点及其周边邻居节点间的紧密程度; 通过衡量加权网络中部分或全部节点聚集系数的平均值, 可以给出更粗粒度下网络的聚集程度. 理论上, 连接越紧密、连接强度越高则聚集系数越高.

2.2 模块度指标

模块度函数 (Q 函数) 是 Newman[11] 提出的衡量网络划分质量的标准, 并对加权复杂网络进行了推广:

$$Q = \frac{1}{2s}\sum_{ij}\left[w_{ij} - \frac{s_i s_j}{2s}\right]\delta(G_i, G_j), \qquad (5)$$

其中, G_i 表示节点 v_i 所在的社团; 当 G_i 与 G_j 相同时, $\delta(G_i, G_j) = 1$, 否则为 0; s 表示网络中边的权值总和. 模块度函数表示实际情况下社团内部连接强度与随机连接情况下的社团内两个节点连接强度的差异. Q 值取值越接近 1, 社团结构越明显, 实际网络中, Q 值最大值一般位于 0.3—0.7.

2.3 强/弱社团指标

如图 1 所示, 强/弱社团指标 [21] 的基本假定并不适用于加权网络, 本节针对加权复杂网络推广强/弱社团的定义.

对于复杂网络 G, 认为其子网络 $G_{\rm sub}$ 中的节点 v_i 的强度由两部分构成, 即 $s_i(G_{\rm sub}) = s_i^{\rm in}(G_{\rm sub}) + s_i^{\rm out}(G_{\rm sub})$; 其中 $s_i^{\rm in}(G_{\rm sub})$ 表示 v_i 与子网络 $G_{\rm sub}$ 中其他节点之间所有边的权值和, $s_i^{\rm out}(G_{\rm sub})$ 指 v_i 与不属于子网络 $G_{\rm sub}$ 的节点之间所有边的权值和. 如果对任意节点, 子网络满足 $s_i^{\rm in}(G_{\rm sub}) > s_i^{\rm out}(G_{\rm sub}), \forall v_i \in G_{\rm sub}$, 则称 $G_{\rm sub}$ 为网络 G 的强社团结构; 如果 $G_{\rm sub}$ 满足 $\sum\limits_{v_i\in G_{\rm sub}} s_i^{\rm in}(G_{\rm sub}) > \sum\limits_{v_i\in G_{\rm sub}} s_i^{\rm out}(G_{\rm sub})$, 则称 $G_{\rm sub}$ 为网络 G 的弱社团结构.

文献 [21] 指出: 如果划分一个网络得到的所有子网络中, 只有一个子网络满足强/弱社团的定义, 则这种划分就不正确或者该网络不具备社团结构. 因此, 该指标可以用来判断一个网络是否具备社团结构, 也可以用来判定社团划分是否正确. 例如, 图 1 中的两类不同节点构成的子网络满足强社团的定义.

3 基于模拟数据集的社团指标分析

依据第 2 节加权网络社团评价指标的计算方法, 就可以利用模拟或真实数据集衡量评价指标的合理性, 从而为判定社团发现算法的有效性奠定基础.

3.1 基于三元结构的聚集系数指标分析

文献 [9] 针对模拟社团分析了在拓扑结构不变的前提下, 不同聚集系数与权值变化的关系. 本文从两个方面补充文献 [9] 的研究: 1) 3.1 节针对最基本的三元结构, 同时分析权值与拓扑对聚集系数的影响; 2) 3.2 节分析聚集系数对高噪音的敏感度.

由定义 (2)—(4) 式可见, 计算聚集系数的重要基础是加权网络中的三元结构, 其他一些研究也以三元结构为基础分析连接强弱的意义. 例如在 Granovetter[3] 于 1973 年发表的经典论文中假定: 当 v_i 与 v_j, v_i 与 v_k 均为强连接时, v_j 与 v_k 应存在连接关系且为强连接.

鉴于三元结构在网络研究中的重要意义, 本节首先给出加权网络中在权值和连接性两方面具典型性的三元结构, 从而在最细粒度上衡量局部聚集程度, 参见图 2. 图 2 中沿 x 轴分布的各类型三元结构体现了连通程度的递变, 沿 y 轴分布的各类型三元结构体现了在连通程度相同的前提下其边权值的递变. 为了简化起见, 本文仅考虑边的权值存在高、低两种情况, 并设 $w_{\rm h} = 3w_{\rm l}$. 在不考虑边权的复杂变化的情况下, 这 10 种基本结构可以通过简单的叠加构成任意形态的加权复杂网络.

聚集系数 $C_{w,i}^{Z}$, $C_{w,i}^{O}$ 与 $C_{w,i}^{B}$ 对前述 10 种基本三元结构的计算结果如表 1 所示. 由表可见, $C_{w,i}^{Z}$ 和 $C_{w,i}^{O}$ 能够明显地区分全互联的三元结构在边权不同时的聚集程度, $C_{w,i}^{B}$ 则不能区分连通度相同而权值不同的结构. 但是 $C_{w,i}^{Z}$, $C_{w,i}^{O}$ 与 $C_{w,i}^{B}$ 均不能区分非连通的三元结构.

上述分析表明: 现有的加权复杂网络聚集系数计算方法不能很好地区分微观网络的聚集程度. 简单推论即可发现: 在分析由上述基本微观结构叠加构成的复杂网络的整体聚集程度时, 也将存在较大缺欠. 例如, 完全由类型 1、类型 5、类型 8 和类型 10 分别构成的四个加权网络, 在采用 $C_{w,i}^{B}$ 方法计算时, 其聚集系数相同, 虽然其聚集的强度明显不同. 其他两个指标虽然识别力较好, 但也存在类

似情况, 例如整个网络由类型 2, 3, 4, 6, 7, 9 构成, 但任何三点均不连通.

3.2 基于高噪音模拟数据的社团指标分析

以往研究中, 通常采用较理想化的数据集剖析各类指标. 本文采用一个含有大量噪音及复杂局域情况的加权网络模拟数据集测试前述三个指标. 图 3(a) 为模拟数据集邻接矩阵的可视化结果. 从图可见, 模拟数据明显分属 6 个类别, 但是各类的大小差异明显, 其中 $n_1 = 400$, $n_2 = 250$, $n_3 = 150$, $n_4 = 100$, $n_5 = 50$ 和 $n_6 = 50$. 同时, 各类别内部和类别间的强度不同, 例如从图 3(a) 和 (b) 都可观察到, 属于两个最大类别 G_1 与 G_2 的数据之间的关联强度高于小类别 G_5 或 G_6 内部的关联强度. 可见, 这一模拟数据既体现了社团的聚集现象, 也反映了社团大小差异较大、聚集程度差异较大、噪音强度高等特点.

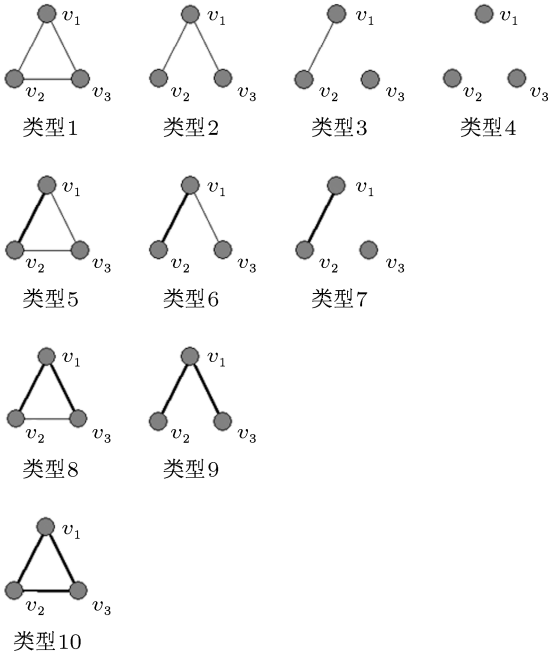

图 2 加权复杂网络中社团的基本三元结构

表 1 10 种基本三元结构的聚集系数考察

	1	2	3	4	5	6	7	8	9	10
$C^Z_{w,i}$	w_l	0	0	0	$\frac{2w_\mathrm{l}+w_\mathrm{h}}{3}$	0	0	$\frac{w_\mathrm{l}+2w_\mathrm{h}}{3}$	0	w_h
$C^O_{w,i}$	w_l	0	0	0	$\sqrt[3]{w_\mathrm{l}^2 w_\mathrm{h}}$	0	0	$\sqrt[3]{w_\mathrm{h}^2 w_\mathrm{l}}$	0	w_h
$C^B_{w,i}$	1	0	0	0	1	0	0	1	0	1

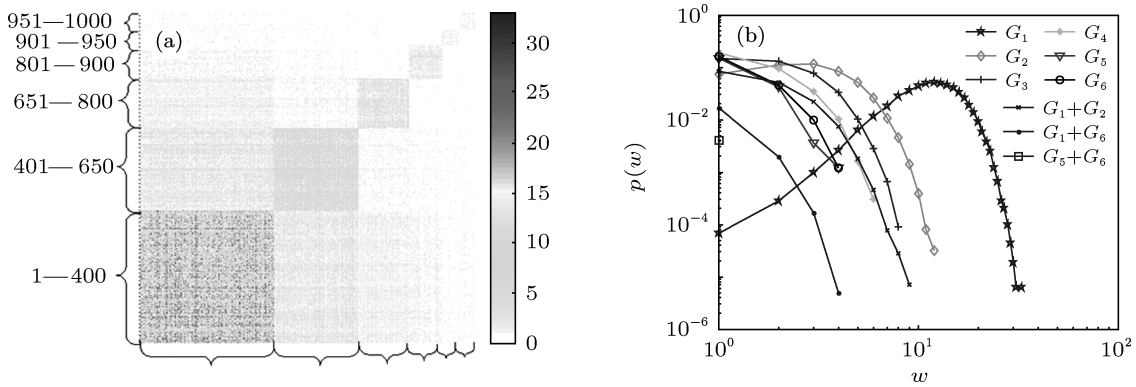

图 3 高噪音模拟加权复杂网络概况 (a) 模拟加权复杂网络邻接矩阵可视化结果, 各点颜色深浅代表边连接强度的高低; (b) 权值分布图, x 轴为权的大小, y 轴为具有相应权值的边在社团或子网络中所占的比重

首先, 分析模拟数据集的聚集系数. 按 (2)—(4) 式计算网络整体的聚类系数分别为 0.2042, 0.0883, 0.7499, 其中各个类别内部的聚集系数、由 G_1 和 G_2 中全部节点及节点间边构成的子网络 $G_1 + G_2$ 与整个网络 G 的聚集系数的比较见图 4. 可见, $G_1 + G_2$ 的聚集系数虽然小于 G_1 的

聚集系数, 但在 $C_{w,i}^{Z}$ 和 $C_{w,i}^{B}$ 的度量方式下均高于 G_2, G_3, G_4, G_5, G_6 的聚集系数, 可见 G_1 与 G_2 间噪音连边的影响巨大, 导致一个并非真实社团的子网络 G_1+G_2 呈现出更高的聚集程度; 而且 $C_{w,i}^{Z}$ 与 $C_{w,i}^{O}$ 计算的各社团内部的聚集系数差异较小, 区分力不强. 因此, 这三个聚集系数的评价方法并不适合于衡量局域情况复杂的加权网络整体的聚集程度.

其次, 分析模拟数据集的强/弱社团结构. 图 5 分析了模拟数据中各类的 $s_i^{\rm in}(G_{\rm sub})$ 与 $s_i^{\rm out}(G_{\rm sub})$ 的之间的关系. 显然, 只有第一个类别满足弱社团的定义. 按照文献 [21] 中的推论, 这种划分不正确或者该网络不具备社团结构. 可见强/弱社团结构也不适合分析聚集情况复杂且噪音大的网络.

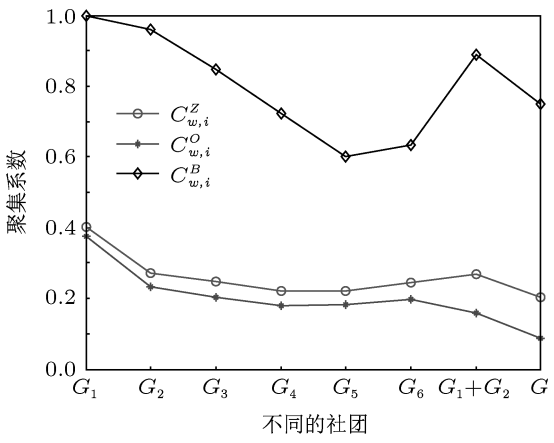

图 4 模拟数据集各个类别内部及两个最大类别间的聚集系数

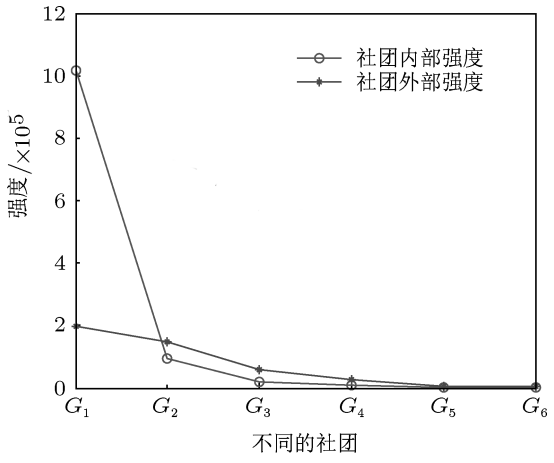

图 5 各类别的 $k_i^{\rm in}(G_{\rm sub})$ 与 $k_i^{\rm out}(G_{\rm sub})$ 的比较分析

再次, 分析模拟数据集的模块度指标. 由于模块度指标通常用来指导聚类过程, 理想情况下: 噪音在真实社团中所占的比例越高, Q 值应越低. 对于图 3(a) 所示的模拟数据集, 其 6 个真实社团的 Q 值的平均值仅为 0.1485, 当去除所有社团间起干扰作用的噪音连边后, Q 值为 0.1955, 说明社团间的噪音连边对 Q 值存在一定的影响.

本文通过向真实社团中移入其他社团的节点作为噪音节点, 同时保留真实社团间的起干扰作用的噪音连接, 进一步分析 Q 函数所受的影响. 具体方法为: 从每个真实类别随机抽取一定比例的节点, 按比例分配到其他类别中, 重新计算新得的社团的 Q 值. 这些抽取的数据将成为新社团中的噪音, 理论上抽取节点越多则各社团中噪音所占比例越高, Q 值应越低. 实验的交换比例从 10%开始递增, 每次递增 10%, 直至 100%; 为避免抽样的随机性对实验结果的影响, 对每一比例重复实验 20 次, 最终采用 Q 值的平均值.

图 6(a) 给出了在保留社团间的噪音连边的前提下, 按照不同比例交换各社团所属节点后, 新得到的包含噪音节点的社团的 Q 值平均值. 表 2 给出 20 次随机抽样实验的 Q 值的统计值, 可见最大值与最小值之间差异很小, 故图 6(a) 可以表征其趋势. 图 6(b) 分析在去除社团间的噪音连边的前提下, 再按一定比例添加噪音, 按不同比例交换社团间的节点时其 Q 值的变化情况.

值得注意的是, 图 6 表明 Q 函数存在两个局部最优解, 其中一个对应最理想结果. 具体地, 当社团内交换节点数量为 70%左右时 Q 值最低, 交换比例继续增大, Q 值反而上升趋势. 其原因为: 当社团内交换节点比例为 70%左右时, 社团被重新分配得最为混杂; 当交换大于 70%后, 其趋势为小社团中节点合并至大社团, 而大社团被分割到几个小社团中; 当交换节点比例为 70%与 100%时, 其可视化图见图 7(a) 与图 7(b).

综上分析, 聚集系数能对整体网络的聚集情况进行粗略评估, 但对小社团聚集情况的评估还有待商榷; 聚集系数、强/弱社团和 Q 函数在分析聚集情况复杂且噪音大的加权网络时均面临较大困难. 社团评价指标的缺欠导致加权网络中社团界定更为模糊.

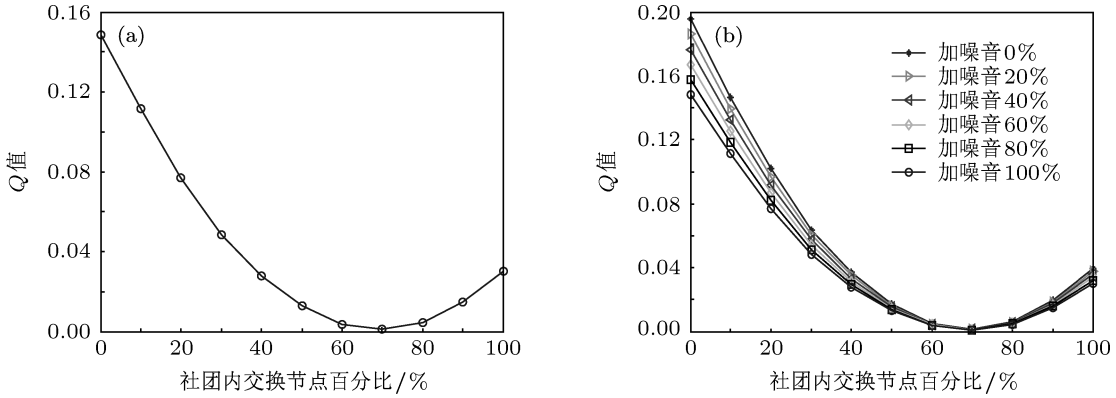

图 6 交换不同社团内节点的比例与 Q 值的关系 (a) x 轴为交换社团内节点的比例; (b) 不同噪音下交换社团内节点对 Q 值的影响, x 轴为交换社团内节点的百分比

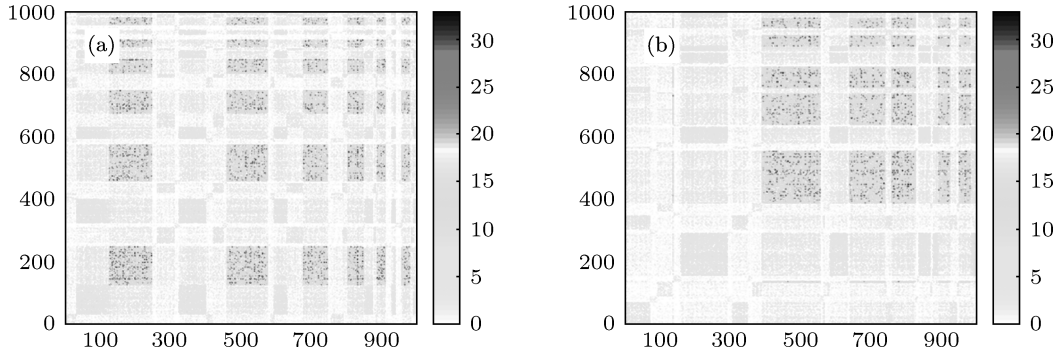

图 7 交换不同社团内节点的比例与 Q 值的关系 (a) 交换社团内 70%的节点; (b) 交换社团内 100%的节点

表 2 交换社团内节点重复实验 20 次的相关统计值

	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Q 值均值	0.1114	0.0771	0.0483	0.0280	0.0129	0.0037	0.0011	0.0046	0.0149	0.0303
最大值	0.1134	0.0792	0.0510	0.0291	0.0137	0.0043	0.0017	0.0053	0.0161	0.0311
最小值	0.1099	0.0751	0.0465	0.0271	0.0123	0.0033	0.0007	0.0040	0.0139	0.0294

4 加权复杂网络社团发现算法实验分析

目前, 主要是通过考虑权值信息, 将无权网络的社团发现算法推广到加权复杂网络. 本文采用三种算法: 基于最小割的谱聚类算法, 基于无权多重图的 GN 算法 [11] 和 FN 算法.

谱聚类算法采用矩阵分析技术将求割函数的问题转化为求网络拉普拉斯矩阵第二小特征值对应的特征向量问题, 从而根据特征向量将网络进行递归划分. 对于 n 个节点的网络, 基于最小割的谱聚类算法的时间复杂度为 $O(n^3)$.

基于无权多重图的 GN 算法采用无权多重图的思想将 GN 算法推广到加权网络, 采用无权多重图的思想, 将边介数除以相应边的权值作为新的边介数, 再移除边介数最大的边. 对于 n 个节点 m 条边的网络, 时间复杂度为 $O(m^2n)$.

FN 算法是一种凝聚层次聚类算法, 最初时各个节点单独成一个社团, 再根据模块度定义函数, 即 Q 函数, 每次按 Q 值增大最大或减少最小的方

向进行合并得到新的社团. 在加权网络中, 采用 2.2 节中的 Q 函数聚类. 对于 n 个节点 m 条边的加权网络, 时间复杂度是 $O((m+n)n)$.

4.1 数据集分析

实验中应用的五个数据集分别是模拟数据集、PSB 数据集、Lesmis、科学家合作网 SCN 和 News 数据集. 其中: 模拟数据集为模拟用户反馈的异质数据集, 即图 3 所示数据集; PSB 数据集为依据三维模型检索系统获得的反馈信息, 构建的三维模型间的语义关系图 [22]; Lesmis[23] 是悲惨世界人物关系的数据集, 节点表示人物, 节点间的边表示这两个人物在同一场景中出现, 边的权值表示共同出现的次数; 科学家合作网 SCN 数据集 [24] 包括 1589 个科学家, 本文选取由 379 个科学家构成的最大连通子图; News 数据集为根据文献 [25] 中的思想对文献 [26] 中的文章构造的数据集, 统计正文中各个单词出现的次数, 保留出现次数大于 5 的有用名词作为节点, 对出现在同一段话中的词连边, 边的权值表示两个名词同时出现的次数.

上述加权复杂网络中, 以往研究通常采用的 Lesmis, SCN 和 News 数据集缺乏先验分类信息作为 ground truth 评价社团发现效果, 仅本文提出的模拟数据集和 PSB 数据集存在原始分类信息. 值得注意的是, 数据集的这一缺欠并不是个例, 这导致精确的量化评价社团发现效果极为困难. 而只能对小数据集采取专家观察的方法, 对于大型数据集则无能为力.

4.2 实验分析

本文从聚集系数、强/弱社团结构、Q 值、信息熵 (Entropy) 和纯度 (Purity) 这几个指标来衡量三种聚集算法, 表 3 给出其实验结果的统计值. 其中, 采用 (2)—(4) 式计算各个社团内的聚集系数, 应用各个社团的聚集系数的均值来考察划分社团的聚集现象; Q 值的计算采用 (5) 式的计算方法; 通过计算聚类结果中满足弱社团结构的社团数来体现强/弱社团结构. Entropy 与 Purity 两个指标的定义为

$$\mathrm{Entropy}=\sum_{i=1}^{k}\frac{n_i}{n}\left(-\frac{1}{\log l}\sum_{j=1}^{l}\frac{n_i^j}{n_i}\log\frac{n_i^j}{n_i}\right), \quad (6)$$

$$\mathrm{Purity}=\sum_{i=1}^{k}\frac{1}{n}\max_j(n_i^j), \quad (7)$$

其中, l 为网络中的真实社团数目, k 为此次划分的社团数目, n_i^j 表示本属于第 j 个社团而算法求得的结果却将其分到第 i 个社团中的节点数目. 在算法的聚类结果与真实社团结构完全符合时, Purity 值为 1, Entropy 值为 0. 即, $\mathrm{Entropy}\to 0$, $\mathrm{Purity}\to 1$ 时, 说明聚类效果好.

表 3 给出三个社团发现算法对于不同数据集在不同评价指标下的聚类效果, 表中 "—" 表示因缺乏分类信息作为 ground truth 而导致无法计算的数据. 由表 3 可见, $C_{w,i}^{Z}$ 与 $C_{w,i}^{O}$ 计算的各算法社团内部的聚集系数均值在不同数据集上无规律, 不能表现各算法的优越性; $C_{w,i}^{B}$ 指标体现了各个算法的社团的平均聚集情况为 $C(\mathrm{FN})>C(\mathrm{GN})>C($谱聚类), 说明 FN 算法得到的社团内部节点的平均聚集情况较高, 但是结合表 1 的分析结果, $C_{w,i}^{B}$ 指标对宏观聚集情况的评价是否恰当则存疑. 因此, 如果需要采用聚类系数衡量加权网络整体的聚集程度, 恰当的指标仍有待研究.

在 PSB 和 SCN 数据集中各个算法的弱社团结构均比较多, 这与数据集本身的特性有关, 即在拓扑结构上观察, 能看出较明显的社团结构, 且社团之间的连边强度较低, 见图 8. 此结论进一步验证了弱社团结构对局域情况复杂网络的不适应性. 对于不同的数据集 FN 算法所计算的 Q 值均比较大, 与此算法本身基于 Q 函数有关; 对 SCN 数据集, 其 GN 算法的 Q 值要大于谱聚类的, 也体现了基于边介数的 GN 算法在此类数据集上的优势. 而对于模拟数据和 PSB 数据计算的 Entropy 与 Purity 两个指标, 谱聚类算法要优于 FN 算法和 GN 算法. 从算法复杂度角度来看, $O(\mathrm{FN})<O($谱聚类$)<O(\mathrm{GN})$.

通过上述分析可见: 1) 现有加权复杂网络聚集算法的泛化能力存在欠缺, 尚不存在适用于多种拓扑结构的聚类算法, 其中, 谱聚类算法的聚类精度较高, 且具有较严格的数学证明, 但是其复杂度高; FN 算法复杂度有所下降, 但是由于其基于 Q 函数, 趋向于将小社团向大社团合并, 因此其聚类精度有所下降; 而 GN 算法具有很高的复杂性, 对社团之间的边数较小的网络较为适用; 2) 由于社团评价指标的有效性存疑且加权复杂网络准确分类信息的缺乏, 使得客观评价社团发现效果变得异常困难.

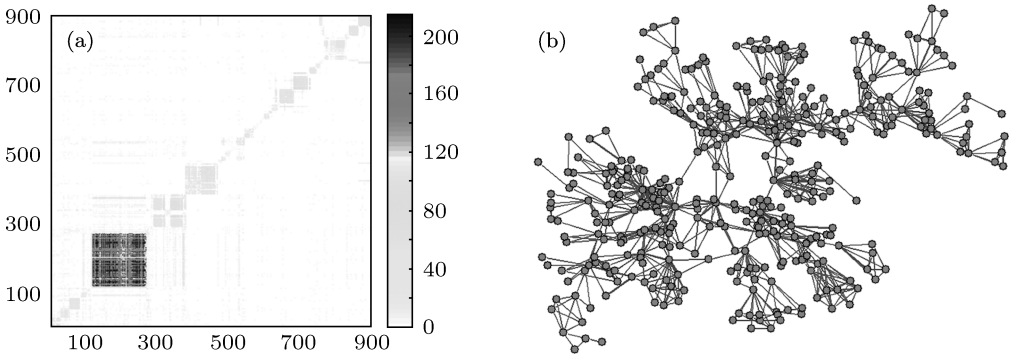

图 8 部分数据集可视化图 (a) PSB 数据集; (b) SCN 数据集

表 3 各社团发现算法比较

		聚集系数均值			Q 值	满足弱社团结构的社团数	Entropy	Purity	社团数	离群点
		$C^Z_{w,i}$	$C^O_{w,i}$	$C^B_{w,i}$						
模拟数	谱聚类	0.2303	0.1795	0.6424	0.0744	2	0.3962	0.6440	6	0
据集	FN	0.2370	0.1661	0.7590	0.0788	1	0.5434	0.5680	2	18
PSB 数	谱聚类	0.5026	0.3891	0.8805	0.2349	7	0.1363	0.8490	50	0
据集	FN	0.4382	0.2874	0.9040	0.2613	4	0.5257	0.4719	4	96
Lesmis	谱聚类	0.2435	0.1732	0.4212	0.5067	4	—	—	11	0
	FN	0.3620	0.2969	0.6755	0.5822	5	—	—	5	0
	GN	0.2192	0.1566	0.5746	0.2281	1	—	—	4	0
SCN	谱聚类	0.3048	0.2774	0.4174	0.3843	17	—	—	78	49
	FN	0.3189	0.2034	0.7590	0.857	21	—	—	21	0
	GN	0.3601	0.3204	0.5033	0.6269	23	—	—	44	58
News	谱聚类	0.3939	0.2548	0.8205	0.2009	0	—	—	3	0
	FN	0.4563	0.3118	0.9033	0.22	0	—	—	4	0
	GN	0.0516	0.0319	0.1505	0.0607	1	—	—	5	6

5 结 论

本文总结、整理并适当扩展了加权复杂网络中聚集系数、模块度指标和强/弱社团三个指标的定义, 并基于含有大量噪音及复杂局域情况的加权网络模拟数据集对上述指标的进行了分析, 指出现有的社团评价指标并不适用于存在复杂局部情况和高噪音的加权网络. 应用 3 种加权网络社团发现算法分析模拟数据集和 4 种真实数据集的效果, 结果表明: 加权复杂网络中社团的准确意义并不明确; 现有的典型社团发现算法的泛化能力均存在缺欠.

下一步的研究重点为探索加权复杂网络中社团的含义及通用的加权复杂网络聚类算法.

[1] Watts D J, Strogatz S H 1998 *Nature* **393** 440

[2] Barabási A L, Albert R, Jeong H, Bianconi G 2000 *Science* **287** 2115

[3] Granovetter M 1973 *Am. J. Soc.* **78** 1360

[4] Barabasi A L, Albert R 1999 *Science* **286** 509

[5] Newman M E J, Girvan M 2004 *Phys. Rev.* E **69** 026113

[6] Yang B, Liu D Y, Liu J M, Jin D, Ma H B 2009 *J. Software* **20** 54 (in Chinese) [杨博, 刘大有, Liu Jiming, 金弟, 马海宾 2009 软件学报 **20** 54]

[7] Cheng X Q, Shen H W 2011 *Complex Syst. Complexity Sci.* **8** 57 (in Chinese) [程学旗, 沈华伟 2011 复杂系统与复杂性科学 **8** 57]

[8] Barrat A, Barthélemy M, Vespignani A 2004 *Phys. Rev.* E **70** 066149

[9] Antoniou I E, Tsompa E T 2008 *Dis. Dyn. Nat. Soc.* **2008** 194

[10] Tian L, Di Z R, Yao H 2011 *Acta Phys. Sin.* **60** 028901 (in Chinese) [田柳, 狄增如, 姚虹 2011 物理学报 **60** 028901]

[11] Newman M E J 2004 *Phys. Rev.* E **70** 056131

[12] Li X J, Zhang P, Di Z R, Fan Y 2008 *Complex Sys. Complexity Sci.* **5** 19 (in Chinese) [李晓佳, 张鹏, 狄增如, 樊瑛 2008 复杂系统与复杂性科学 **5** 19]

[13] Wang X F, Li X, Chen G R 2006 *Complex Network Theory and Its Application* (1st Edn.) (Beijing: Tsinghua University Press) p10 (in Chinese) [汪小帆, 李翔, 陈关荣 2006 复杂网络理论及其应用 (第一版) (北京: 清华大学出版社) 第 10 页]

[14] Zhang B, Horvath S 2005 *Stat. Appl. Genet. Mol.* **4** 1128

[15] Onnela J-P, Saramäki J, Kertész J, Kaski K 2005 *Phys. Rev.* E **71** 065103

[16] Barrat A, Barthelemy M, Pastor-Satorras R, Vespignani A 2004 *PNAS* (USA) **101** 3747

[17] Holme P, Park S M, Kim B J, Edling C R 2007 *Physica* **373** 821

[18] Kalna G, Higham D J 2006 *SNANSE* pp45–50

[19] Lopez-Fernandez L, Robles G, Gonzalez-Barahona J M 2004 *Proc. of the 1st Intl. Workshop on MSR* pp101–105

[20] Serrano M A, Boguñá M, Pastor-Satorras R 2006 *Phys. Rev.* E **74** 055101

[21] Filippo R, Claudio C, Federico C, Vittorio L, Domenico P 2004 *PNAS* **101** 2658

[22] Lü T Y, Huang S B, Wu P, Jia Y R 2010 *Proc. SKG* pp211–218

[23] Knuth D E 1993 *The Stanford Graph Base*: *A Platform for Combinatorial Computing* (1st Ed.) (Indianapolis: Addison-Wesley Professional) pp15

[24] Newman M E J 2006 *Phys. Rev.* E **74** 036104

[25] Kevin D, Steven C 2004 *NDPLS* **8** 259

[26] http://www.nti.org/d_newswire/issues/newswires/2001_10_17.html

Analysis of community evaluation criterion and discovery algorithm of weighted complex network*

Lü Tian-Yang[1)2)3)†] Xie Wen-Yan[3)] Zheng Wei-Min[1)] Piao Xiu-Feng[3)]

1) (*College of Computer Science and Technology, Tsinghua University, Beijing* 100084, *China*)

2) (*Audit Research Institute, National Audit Office, Beijing* 100830, *China*)

3) (*College of Computer Science and Technology, Harbin Engineering University, Harbin* 150001, *China*)

(Received 31 March 2012; revised manuscript received 7 May 2012)

Abstract

The clustering of nodes is an important feature of complex network. Previous researches mainly focus on community discovery in unweighted network, with little attention paid to the weighted network because of the complexity of weighted network. The community discovery of the weighted network is believed to be a much more difficult task. In this paper, we perform a study on the effectivenesses of community evaluation criterion and the performances of the existing discovery algorithms. First, we summarize three classical community evaluation criterions of weighted network, and analyze their effectivenesses according to a simulated noisy dataset, which has different community sizes, densities and local characteristics. Second, we adopt five datasets to compare the performances of three typical community discovery algorithms. The study shows that the existing criterions encounter difficulties in evaluating the basic community structure and in evaluating the weighted community with complex structure, and the generalization ability of the typical community discovery algorithm of weighted network is unsatisfactory.

Keywords: complex network, community discovery, clustering coefficient, modularity

PACS: 05.65.+b

* Project supported by the National Basic Research Program of China (Grant Nos. 2009BAH42B02, 2012BAH08B02), the National Natural Science Foundation of China (Grant No. 60903080), the Fundamental Research Funds for the Central Universities, China (Grant No. HEUCF100603) and the Scientific Research Fund of Heilongjiang Provincial Education Department, China (Grant No. 12513050).

† E-mail: raynor1979@163.com