一种面向大规模社会信息网络的多层社区发现算法

康颖[1),2)]，于博[1)]，林政[1)]，周江[1)]，王伟平[1)]，孟丹[1)]

[1)](中国科学院信息工程研究所，北京，中国，100093)

[2)](中国科学院大学，北京，中国，100049)

摘 要 社区发现旨在挖掘社会信息网络的社区结构，是社会计算及其相关研究的基础。随着交互式社会信息网络规模的快速增长，传统的社区发现算法难以满足大规模网络的可扩展分析需求。多层社区发现算法如 PMetis、Graclus 等虽然可以分析包含数百万节点规模的网络，但是小于 1/2 的粗化缩减比率以及社会信息网络的幂律分布特性极大地制约着该类算法的性能优势。本文提出了一种基于三角形内点同一社区性粗化策略的多层社区发现算法 TMLCD。TMLCD 不仅以大于 1/2 的粗化缩减比率加快了大规模社会信息网络的粗化过程，而且从基本拓扑结构上保持了初始网络的社区效应，提高了社区发现精度。基于真实网络如 Youtube、Orkut 等的实验结果表明，TMLCD 的计算精度、内存占用以及运行时间均优于目前典型的多层社区发现算法，适用于富含三角形的社会信息网络分析。

关键词 社会信息网络；社区发现；三角形

中图法分类号

A Multilevel Community Detection Algorithm for Large-scale Social Information Networks

KANG Ying[1),2)]， YU Bo[1)]， LIN Zheng[1)]， ZHOU Jiang[1)]， WANG Wei-Ping[1)]， MENG Dan[1)]

[1)](Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China)

[2)](University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract Community detection aims at mining the community structures of Social Information Networks (SINs), which is the foundation of other related researches on social computing. Due to the rapid inflation of interactive SINs, traditional community detection algorithms encounter obstacles in analyzing large-scale networks with scalability. Although multilevel community detection algorithms such as PMetis, Graclus etc. have the capability to analyze networks containing millions of nodes, the coarsening shrink rate is less than 1/2 and the SINs follow the power-law distribution, which constrain these algorithms' performance enormously. This paper proposes a multilevel community detection algorithm TMLCD, based on the coarsening policy of triangle's inner nodes belonging to the same community. TMLCD accelerates the coarsening process of large-scale SINs at a

收稿日期：年-月-日*投稿时不填写此项*；最终修改稿收到日期：年-月-日 *投稿时不填写此项*. 本课题得到国家科技支撑计划项目(2012BAH46B03)；国家核高基项目(2013ZX01039-002-001-001)；中国科学院先导专项(XDA06030200)；国家“八六三”高技术研究发展项目(2012AA01A401)资助. 康颖，女，1984年生，博士在读，主要研究领域为数据挖掘、社区发现等，E-mail: kangying@iie.ac.cn. 于博，男，1985年生，博士，助理研究员，主要研究领域为无线网络、无线传感器网络、分布式计算、查询处理等，E-mail: yubo@iie.ac.cn. 林政，女，1984年生，博士，助理研究员，主要研究领域为自然语言处理、情感分析等，E-mail: linzheng@iie.ac.cn. 周江，男，1980年生，博士，助理研究员，主要研究领域为分布式文件系统、容错系统、大数据存储系统等，E-mail: zhoujiang@iie.ac.cn. 王伟平，男，1975年生，博士，教授，博师生导师，CCF会员(E200012474M)，主要研究领域为数据流、高性能数据库、并行处理等，E-mail: wangweiping@iie.ac.cn. 孟丹，男，1965年生，博士，教授，博师生导师，主要研究领域为高性能计算机体系结构、分布式文件系统、系统安全等，E-mail: mengdan@iie.ac.cn.

Tel:18610848607，E-mail: kangying@iie.ac.cn

coarsening shrink rate of greater than 1/2, and preserves the community effect of initial network from the point of basic topological structure and improves the accuracy of community detection. Experimental results from the real networks like Youtube, Orkut etc. indicate that, TMLCD outperforms the currently typical multilevel community detection algorithms in terms of computing precision, memory occupation and running time. It is obvious that TMLCD is appropriate for analyzing the SINs rich in triangles.

Key words Social Information Networks; Community Detection; Triangle

1 引言

随着 Web2.0 概念的出现和不断深入，大量用户交互式社会媒体如微博、论坛、社交网络、社会新闻、维基等不断涌现，基于这些共享社会媒体平台产生的新型信息网络称为社会信息网络。社会信息网络具有强社区效应，社区发现是分析社会信息网络社区结构的一种有效方法，其广泛的应用场景引起了社会、经济、信息安全等多个学科领域的普遍关注。深入挖掘社会信息网络社区可为圈内好友推荐、个性化信息导向、未来经济走势预测、网络舆情监控等提供技术支持。

社区发现本质上是聚类在社会网络研究应用上的延伸。从定性的角度讲，社区发现就是发现一个网络中内部连接紧密且外部连接稀疏的子团。传统的社区发现算法有谱分析方法[1]、分层凝聚法[2]、Girvan-Newman 分裂法[3]、基于模块度的方法[4]以及结合其他学科理论如统计学[5]、遗传学[6]、信息论[7]等衍生的社区发现算法。这些算法虽然不断改进优化以适应新的数据计算需求，但 $\Theta(n^2)$或更高的计算复杂度制约着其分析大规模网络的能力。

为了能够扩展挖掘大规模网络的社区结构，近年来，研究人员提出了一些随机近似算法进行社区发现，如 Charicar 等[8]在内存有限的情况下，通过流模式扫描大规模网络一次或几次得到其近似网络拓扑结构，从而加快了社区发现过程；Satuluri 等[9]采用随机抽样去边稀疏化的方法降低了网络规模，可在一定误差范围内高效识别出大规模网络的社区结构。虽然上述两种算法提高了大规模网络社区发现的效率，但随机近似计算过程会给实际问题求解带来较大误差，使得这些算法存在一定的计算不确定性和应用局限性。

多层社区发现是目前大规模网络社区结构分析方法中应用广泛且性能较好的一种，如 Karypis 等提出的 KMetis[10]、PMetis[11]以及 Dhillon 等提出的 Graclus[12]等。该类方法不直接计算大规模网络的社区结构，而是先将网络规模通过迭代粗化的方法逐层缩减，然后分析粗化后生成的小规模网络的社区结构，再通过反粗化来逆推出原大规模网络的社区结构。已有的多层社区发现算法在进行粗化时均采用一种极大匹配选边粗化策略[10]，即通过一定方法尽可能多的从网络中选取出可被粗化的边，然后将被选边上的两个端点融合形成一个复合节点来实现粗化。这种粗化策略存在以下三个方面的问题：(1)极大匹配选边的过程是一个 NP 难问题，其所选边的数目最多为原网络边数的一半，严重制约着该类算法的粗化缩减比率[10]（Coarsening Shrunk Ratio 为$|G_{i+1}|/|G_i|$，其中$|G_i|$和$|G_{i+1}|$分别代表粗化前、后网络中的边数；粗化缩减比率越高，粗化过程越快，存储反粗化所需的中间过渡图信息量越少。目前已有算法的粗化缩减比率实际值均小于1/2），进而影响其计算时间复杂度和存储空间占用；(2)被选作粗化边上的两个端点，其自然社区归属性不确定，但因复合而被纳入同一个社区将直接影响初始社区发现结果，因而反粗化阶段需要大量的调优工作来提高算法整体的计算精度；(3)社会信息网络的节点度一般服从幂律分布[13](power-law)，度高的节点会以其中心占优势屏蔽掉很多可被选作粗化的边，提早结束粗化过程，加剧算法的时间和空间复杂度，限制其分析大规模网络时的性能优势。

针对上述问题，本文提出一种基于三角形的多层社区发现算法 TMLCD (Triangle-Based Multilevel Community Detection)。该算法在粗化阶段采用一种新的粗化策略——基于三角形内点同一社区性的粗化策略，即将网络中遍历到的三角形三个顶点收缩融合形成一个复合节点来实现粗化。三角形因其三个顶点完全互连而具有强社区性[14]，选择三角形作为粗化对象不仅确保了被复合节点同属一个社区，使得经逐层迭代粗化后生成的小规模网络可以保持原始大规模网络的基本社区结构，提高初始社区发现精度，而且可以节省大量反粗化调优工作并以高精确度反推出原始大规模网络的社区结构。实验结果表明，在分析富含三角形的大规模社会信息

网络时，TMLCD 以大于 1/2 的粗化缩减比率加快了网络粗化过程，同时降低了反粗化所需存储的中间过渡图信息量，且能够发现幂律分布网络中的社区结构，极大地提升了多层社区发现算法的性能。

2 基于三角形的多层社区发现算法

多层社区发现的设计理念源于网络社区结构在不同划分度要求下呈现出的层次嵌套模式[15]，是一种高效分析大规模网络且扩展性能很好的方法。据 Abou-Rjeili 等[13]分析，即使在最坏情况下，多层社区发现的计算结果均优于直接分析大规模网络的计算结果。多层社区发现的计算过程一般由粗化、初始社区发现、反粗化并调优三个阶段组成[11]，如图 1 所示。本文秉承多层分析模型的特点并融入基于三角形内点同一社区性的粗化策略，提出一种可扩展分析大规模社会信息网络的多层社区发现算法——基于三角形的多层社区发现算法 TMLCD (Triangle-Based Multilevel Community Detection)，下面将具体阐述 TMLCD。

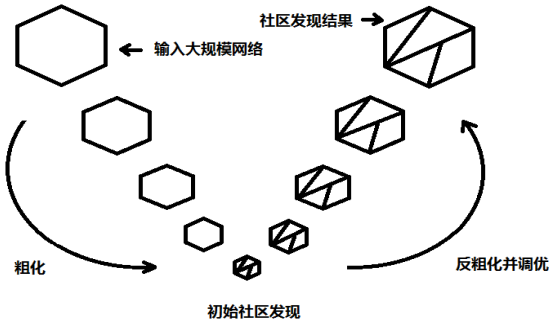


图 1 多层社区发现的计算过程

2.1 设计基础

三角形是复杂网络的基本构成元素，其以最简单的强社区结构，成为网络社区分层嵌套模式中的基本社区单元，是嵌套树中的叶子社区或者叶子社区的有机组成部分。研究表明，85%的大规模社会信息网络富含三角形[16]，具体表现为朋友圈中的三个互粉好友、论坛中的三则跟帖评论等，其共同的社会属性如兴趣爱好、热点事件关注等，从几何上直观展示了三角形的内在社区性。相对于高阶完全子图，大规模社会信息网络包含了更加丰富的三角形，为 TMLCD 粗化过程提供了必要的粗化源。

另一方面，随着粗化层级的不断深入，网络中的高阶多边形会逐渐降阶演变成三角形，为 TMLCD 的逐层迭代粗化提供可持续的粗化源。例如图 2 中所示，粗化前社区 1 内的节点 6、7 与节点 1、2、3 或者节点 1、3、4 构成一个五边形，当由节点 1、2、4 构成的三角形被融合粗化形成一个复合节点时，上述五边形将降阶为下一层粗化图上的四边形，即如图 2 中粗化后的社区 1 内由节点 3、4、6、7 所构成的四边形（图中仅给出粗化局部效果图，未展示全图粗化结果），原因是五边形与被粗化三角形共边；同理，粗化前社区 2 内由节点 9、10、12、13 构成的四边形也会随着由节点 8、9、13 构成的三角形被粗化而降阶为下一层粗化图上的三角形。如此粗化迭代，高阶多边形逐步向低阶多边形演化，甚至转变成三角形，为后续粗化过程提供粗化对象，从而保证了 TMLCD 逐层迭代粗化的可持续性。

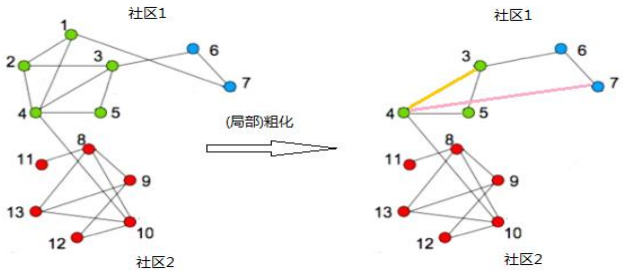


图 2 高阶多边形的降阶过程

2.2 算法实现

本文先将基于链接结构的大规模社会信息网络表示成图 $G_0=(V_0, E_0)$，其中 V_0、E_0 分别表示节点集和边集，节点 $v \in V_0$ 表示个人、组织等信息，边 $e \in E_0$ 表示节点之间某种无向链接关系，如好友关系、合作关系等，边权重 ω_0 取单元值。社区就是图上内连边数大于外连边数的局部强关联子图。

TMLCD 算法的具体实现过程如下：

2.2.1 粗化阶段

遍历大规模社会信息网络图 G_0 上的三角形，将每次遍历遇到的三角形三个顶点融合粗化形成一个复合节点 v_c，相应连接边参照复合节点作归一化处理，且累加求边权重之和得到 ω_c，v_c 和 ω_c 将成为下一层粗化图的基本组成元素。当图 G_0 遍历完全生成第一层粗化图 G_1 时，基于图 G_1 继续上述三角形遍历粗化过程生成第二层粗化图 G_2，然后再基于图 G_2 继续生成下一层粗化图 G_3，如此循环迭代，生成一系列规模逐渐减小的粗化图 $G_0, G_1, \cdots, G_m$，满足节点数逐层递减$|V_0|>|V_1|>\cdots>|V_m|$。当图 G_m 的规模小于一定阈值（该阈值一般为设定内存容量大小）时，粗化迭代过程终止，且定义此时生成的粗化图 G_m 为最简粗化图。

TMLCD 算法的多层粗化是基于单层粗化迭代实现的，本文将该单层粗化过程定义为三角形遍历粗化算法 TTCA (Triangle Traversing Coarsening Algorithm)，其具体实现如算法 1 所示。算法中图 G 以邻接表形式给出，节点信息包括标号、节点度 $d(v)$ 和标志位 $F(v)$， $N(v)$表示节点 v 的相邻节点集；为了避免图上三角形遍历的随机性且使算法适合于分析节点度服从幂律分布的网络，算法开始时先作排序 *SortedDegree (Adj(G))*。另外，TMLCD 逐层迭代粗化生成的中间过渡图均需被存储，为后续反粗化还原、调优处理提供必要的信息。

算法 1. *TTCA*.

输入：位于磁盘上图 $G(V,E)$的邻接表 $Adj(G)$

输出：经粗化处理后图 $G'(V',E')$的邻接表 $Adj(G')$

1. $Adjs(G)$ =*SortedDegree* ($Adj(G)$);
2. INIT($F(v)$);
3. WHILE ($Adjs(G)$中有未读数据){
4. *buffer* =READ ($Adjs(G)$);
5. FOREACH (node v in *buffer*) DO
6. IF ($F(v)$>1) THEN
7. CONTINUE;
8. END IF
9. $N(v)$ =v 的相邻节点集在内存? 直接赋值: 从磁盘读入内存;
10. FOREACH (node u in $N(v)$) DO
11. IF ($d(u)$<$d(v)$ or $F(u)$>1) THEN
12. CONTINUE;
13. END IF
14. $N(v) \cap N(u)$=v 和 u 的相邻节点集交集在内存? 直接赋值: 从磁盘读入内存;
15. FOREACH (node w in $N(v) \cap N(u)$) DO
16. IF ($d(w)$<$d(v)$ or $d(w)$<$d(u)$ or $F(w)$>1) THEN
17. CONTINUE;
18. ELSE
19. 融合三角形形成复合节点 v_c，相应连接边聚合并求权重和 ω_c；设置复合结点 $F(v_c)$ += 1，被复合节点 $F(v)$=2;
20. BREAK;
21. END IF
22. END FOR
23. END FOR
24. END FOR }
25. 由复合节点、聚合边和未被粗化的节点、边联合构成粗化图 $Adj(G')$;

2.2.2 初始社区发现阶段

调用社区发现算法对最简粗化图 G_m 进行初始社区发现。原则上，任何高效的社区发现算法均可用于初始社区发现，因为图 G_m 包含的数据量已经完全被内存容纳。但随着迭代粗化的深入，原始大规模社会信息网络图 G_0 中的边权重开始由单元值逐渐累加求和变成了图 G_m 中的数值，因此，边的权重成为决定初始社区发现质量的重要因素，选择初始社区发现算法必须考虑边权重的可计算性。本文采用 Pizzuti[17]提出的基于遗传学社区发现算法 GA-Net 进行初始社区发现。GA-Net 运用遗传学机制缩小最优解目标搜索空间以降低算法的时间复杂度，具体地，首先通过选择交叉算子将解空间锁定在使目标函数值趋优的范围内，避免全图逐点搜索，再结合变异算子将解空间适度变换到全局范围内，避免陷入局部最优解而提早终止计算过程，最终找到全局最优解。GA-Net 是一种社区自动检测方法，不需要任何关于社区的先验知识（如社区数目 k 等），也不局限于分析凸社区结构。

GA-Net 算法定义了一个全局意义上的社区发现目标函数，如公式 (1)所示：

$$Q_c = \sum_{i}^{k} \frac{\sum_{i \in I} (a_{iJ})^r}{|I|} \times V_{C_i} \tag{1}$$

其中，Q_c 为公式化的社区发现定义，求解 Q_c 的最大值即社区发现过程，$C=\{C_1, C_2, \cdots, C_k\}$为图上的 k 个子图社区，C_j 代表某一子图社区，其矩阵表示形式为 $C_j=(I, J)$， V_{Ci} 为 C_j 的容量， a_{iJ} 为 C_j 第 i 行的均值。由于 GA-Net 仅限于基于链接结构的图上计算，为了拓展分析加权图 G_m，本文将边权重 ω 融入 a_{iJ} 的如下定义式(2)中：

$$a_{iJ} = \frac{1}{|J|} \sum_{j \in J} \omega_{ij} a_{ij} \tag{2}$$

式中 a_{ij} 取布尔值表示节点间链接与否，ω_{ij} 为对应边权重。经过逐层迭代粗化，边权重 ω_{ij} 由初始的单元值累加求和逐渐转变成为图 G_m 中的数值；ω_{ij} 值越大，表示节点间的关联度越高，其社区同属性越高，因此，ω_{ij} 成为初始社区计算的重要因素。

2.2.3 反粗化并调优阶段

初始社区发现结果并不能直接呈现出原始大规模社会信息网络的社区结构，需经过逐层反粗化还原出图 G_0 上的社区结构，即将最简粗化图 G_m 上分析得到的社区结果通过中间过渡粗化图 G_{m-1} →

G_{m-2}→⋯→G_1 逐层反粗化还原到图 G_0 上，以得到原始大规模社会信息网络的社区结构。在图 G_{i+1} 反粗化还原图 G_i 时，通常将复合节点展开的各节点简单地归属到复合节点所属社区内。

对于基于极大匹配选边粗化策略的多层社区发现算法，由于复合节点反粗化还原过程会产生随机自由节点（即社区归属性不确定的节点），若不作适度调优，社区发现结果误差将在粗化图层间以链式反映逐级递增，进而降低原始图 G_0 上社区发现结果的准确性。因此，每次反粗化均需采用调优算法如 KL (Kernighan-Lin)、BKL (Boundary Kernighan-Lin)[18]等重新调整随机自由节点的社区归属性。而对于本文提出的 TMLCD 算法，由于采用基于三角形内点同一社区性的粗化策略，即从三角形所具有的内在社区性出发，将其三个顶点融合实现粗化，从基本拓扑结构上保持了初始网络的社区效应，因此，反粗化还原展开的各节点可直接归属到复合节点所属的社区内，不会产生随机自由节点。但存在一种特殊情况，即被粗化的三角形位于重叠社区部分。如图 3 所示，由节点 1、2、3 构成的三角形位于重叠社区部分，其因强社区性同时属于社区 1 和社区 2。如果优先选择该三角形作融合粗化，社区 1 和社区 2 会因过早聚合而形成一个大社区，模糊掉一定划分度要求下的自然社区结构。因此，与基于极大匹配选边粗化策略的多层算法不同，TMLCD 反粗化调优工作的重点不是重新调整由复合节点展开的节点的社区归属性，而是分离因社区重叠而被聚合的大社区，从而发现不同划分度要求下的自然社区结构。

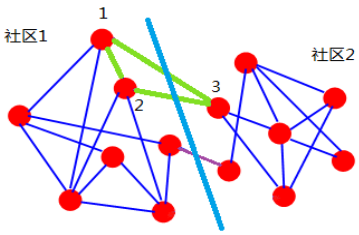


图 3 位于重叠社区部分的三角形

3 算法性能分析

3.1 复杂度分析

基于三角形内点同一社区性的粗化策略使得 TMLCD 算法在反粗化阶段无需特别调优处理，且初始社区发现阶段所需分析的图规模远远小于原始大规模社会信息网络，因此，TMLCD 的复杂度主要取决于粗化阶段，即 TTCA 的有限次迭代粗化。回顾算法 1 中 TTCA，假设图 G 包含 n 个节点和 m 条边，第 1 行中的排序过程可在 $\Theta(nlog(n))$时间内完成，且空间复杂度为 $\Theta(n)$；当数据从磁盘写入内存，第 5 行至第 10 行通过两层嵌套循环遍历图上节点以及度大于该节点的相邻节点来遍历图上的边，因此时间复杂度为 $\Theta(m)$；第 14 行相邻节点集的交集计算完全可以在 $\Theta(m^{1/2})$时间内完成，这是因为节点已排序，因此第 15 行最内层循环的时间复杂度也为 $\Theta(m^{1/2})$；这三层嵌套循环的计算时间复杂度为 $\Theta(m^{3/2})$= $(\Theta(m)+\ \Theta(m^{1/2}))$且占用 $\Theta(n+m)$大小的存储空间。另外，算法中所描述的“遍历”图上三角形并不需要“完全遍历”图上的每一个三角形，如图 4 所示，若三角形 A 被融合粗化，与之相邻的三角形 B 和 C 会随之消失，因为边(1,4)、(3,4)以及边(1,5)、(3,5)会因节点 1 与节点 3 复合而被聚合形成一条边。因此，后续遍历选择粗化对象时，可通过标志位 F(v)的值来判断并忽略所有与被粗化三角形相邻的三角形。综上，TTCA 算法的时间复杂度将远小于 $\Theta(m^{3/2})$。

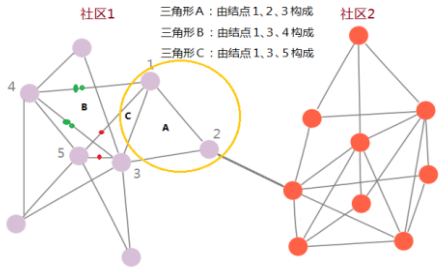


图 4 邻边三角形被聚合单边化

3.2 性能提升

TMLCD 算法聚焦于社会信息网络图上的三角形并将被遍历选定的三角形三个顶点融合形成一个复合节点来实现粗化，从基本拓扑结构上保持了生成粗化图 G_{i+1} 与被粗化图 G_i 的社区结构一致性($0 \leqslant i \leqslant m$)，即经过 m 次迭代粗化后生成的最简粗化图 G_m，其拓扑结构近似于初始图 G_0，使得在图 G_m 上分析得到的社区结果可以最大限度地还原出原始大规模社会信息网络图 G_0 上的社区结构，减小分析误差，提高初始社区发现精度，且反粗化无需调优界定被展开节点的社区归属性，降低了系统的时间开销。另外，相较于基于极大匹配选边粗化策略的多层社区发现算法，基于三角形内点同一社区性粗化策略的 TMLCD 算法可以大于 1/2（因为三角形三个顶点被融合粗化形成一个复合节点）的粗化缩减比率加快网络粗化过程，减少所需保存的中

间过渡粗化图层数和每层粗化图的信息量，降低了多层社区发现算法的空间占用。

为了进一步提升社区发现性能，TTCA 算法采用了对节点的度进行排序并依次从度最低的节点开始遍历图上三角形的方法。这样做有两方面的原因：一方面，节点度服从幂律分布的网络分布极不均衡，中心社区和边缘社区的密度相差较大[13]，若粗化阶段随机选择粗化三角形，则会造成度高的节点以其中心占优势将网络先向中心聚合粗化而模糊掉网络的边缘特征，从而覆盖掉很多有意义的边缘低密度社区；另一方面，度高的节点所邻接的三角形较多，选择不同的三角形会引起图遍历粗化的随机性。相反地，度低的节点所邻接的三角形较少，当其中一个三角形被选作粗化时，与该粗化三角形相邻的三角形的顶点会被逐渐标记（$F(v)$）而变的不可选，这样从边缘向中心推进，逐渐减少度高节点所邻接的三角形中可被选作粗化对象的数目，降低图 G 遍历粗化的随机性；而且，从度最低的节点开始遍历，优先选择粗化边缘社区内的三角形，可以突出边缘低密度社区效应，弱化占优节点的中心覆盖性，进而提高 TMLCD 发现自然社区的能力。

3.3 过粗化现象

过粗化是一种极端现象，一般不会发生，但当网络中被选作粗化对象的三角形很多都存在于高阶极大完全子图时，就会产生过粗化现象。例如图 5 所示，粗化前社区 A、B、C 是三个高阶极大完全子图，若将其每个子图中包含的三角形均作融合粗化，就会产生粗化后的一个简单三角形，原有的社区效应完全消失。这种导致社区内连边数目大大减少或消失，使得其与社区间连边数目相当，模糊掉社区边界，进而抵消网络局部聚集特性（即社区内边密度高于社区间边密度的社区效应）的粗化现象就是过粗化。过粗化将严重影响社区发现质量，因此，本文在设计 TMLCD 算法时，采用设置标志位 $F(v)$的值来控制一个节点所在三角形中被选作粗化对象的数目，有效防止了过粗化现象的产生。

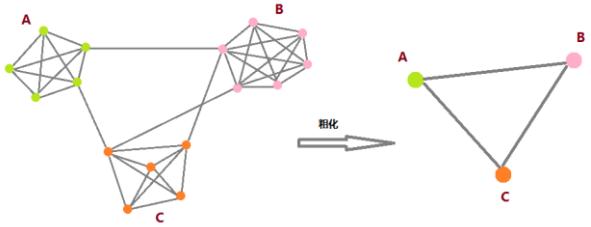


图 5 过粗化现象的产生

4 实验

4.1 实验数据及设置

TMLCD 算法是基于三角形内点同一社区性粗化策略实现的，因此本实验分成两组进行测试：第一组将对比于极大匹配粗化策略，测试基于三角形内点同一社区性粗化策略对单层粗化过程性能的改善；第二组将对比于目前可有效发现大规模网络社区的算法 Graclus、PMetis 和 GEM，测试 TMLCD 的多层社区发现性能。所有实验均运行于配置为 2.67GHz CPU、24 核处理器且内存为 48GB 的 Linux 机器上，算法基于 Java 语言实现。

实验数据采用斯坦福大学提供的真实社会信息网络数据负载[19]，并根据实验需求分成三组，前两组用于第一组测试，后一组用于第二组测试，其每组数据的具体信息分别列于表 1、表 2 和表 3 中。

表 1 实验数据集（一）

数据-图	节点数	边数	三角形数
Collaboration-G_C	23,133	93,497	173,361
Email-G_E	36,692	183,831	727,044
Flickr-G_F	81,306	768,619	3,985,776

表 2 实验数据集（二）

数据-图	节点数	边数	$N_v=\alpha d^{\beta}$（幂律分布函数，N_v 为节点数，d 为节点度，α、β 为参数）	
			α	β
Actor	513,165	1,637,860	2.56	-1.91
Google	216,872	328,917	1.39	-2.26

表 3 实验数据集（三）

数据-图	节点数	边数	三角形数
DBLP	317,080	1,049,866	2,224,385
Amazon	334,863	925,872	667,129
Youtube	1,134,890	2,987,624	3,056,386
Orkut	3,072,441	117,185,083	627,584,181
LiveJournal	3,997,962	34,681,189	177,820,130
Twitter	17,069,982	476,553,560	1,566,092,367

4.2 TTCA性能评测

为了对比分析基于三角形内点同一社区性粗化策略实现的单层粗化算法 TTCA 的性能，本组测试先将已有多层社区发现算法中基于极大匹配选边粗化策略的单层粗化过程定义为极大匹配选边

粗化算法 MMSEA。具体实验步骤如下：先对表 1 中 3 个规模相对较小的社会信息网络数据（这里仅需对比单层粗化算法 TTCA 和 MMSEA 的性能，因此数据规模要求可适度降低）构建无向图 G_C、G_E 和 G_F；然后分别运用 TTCA 和 MMSEA 对图 G_C、G_E 和 G_F 进行粗化得到图 G_{TC}、G_{TE}、G_{TF} 和图 G_{MC}、G_{ME}、G_{MF}，并要求经两种算法粗化后得到的图的规模相当；再采用社区发现遗传算法 GA-Net 对在所有粗化图进行社区发现并比对结果。

由于表 1 中的数据规模较小，因此本测试可采用 GA-Net 直接分析三个网络图 G_C、G_E 和 G_F 上的社区结构，并将该社区发现结果作为基准社区以对比 TTCA 和 MMSEA 性能。GA-Net 的具体参数设定如下：种群规模为 500，一共遗传 50 代，每代交叉率为 0.8，突变率为 0.2，遵循轮转选择法则，并按照种群规模 10%的比例选择精英个体直接复制到下一代。为了对比 TTCA 和 MMSEA 粗化结果对社区发现精度的影响差异，这里采用归一化互信息 *NMI* [20] (Normalized Mutual Information)进行测试对比。*NMI* 是一种相似度度量指标，由 Danon 等[20]证明可用于度量两种社区划分结果之间的相似程度，即若两个社区发现结果相同，*NMI* 值为 1，若完全不同，*NMI* 值为 0，其定义如公式(3)所示：

$$NMI(\Omega,C)=\frac{I(\Omega,C)}{[H(\Omega)+H(C)]/2} \tag{3}$$

其中 Ω 和 C 分别代表两种不同社区发现结果，即 $\Omega=(\omega_1,\omega_2,\cdots,\omega_k)$，$C=(c_1,c_2,\cdots,c_j)$，$I$ 代表互信息，H 代表信息的熵，其形式化定义分别如公式(4)和公式(5)所示：

$$\begin{aligned} I(\Omega,C)&=\sum_k\sum_j P(\omega_k\cap c_j)\log\frac{P(\omega_k\cap c_j)}{P(\omega_k)P(c_j)}\\ &=\sum_k\sum_j\left(\left|\omega_k\cap c_j\right|/N\right)\log\frac{N\left|\omega_k\cap c_j\right|}{\left|\omega_k\right|\left|c_j\right|} \end{aligned} \tag{4}$$

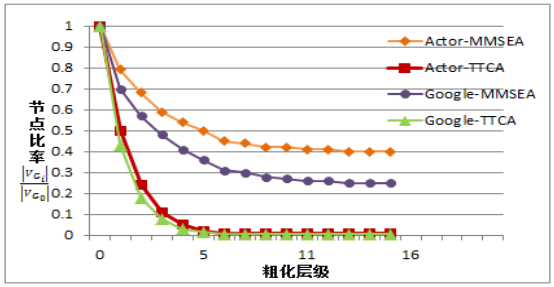
$$\begin{aligned} H(\Omega)&=-\sum_k P(\omega_k)\log P(\omega_k)\\ &=-\sum_k\frac{\left|\omega_k\right|}{N}\log\frac{\left|\omega_k\right|}{N} \end{aligned} \tag{5}$$

实验中，将经 TTCA 粗化得到的图 G_{TC}、G_{TE}、G_{TF} 和 MMSEA 粗化得到的图 G_{MC}、G_{ME}、G_{MF} 分别通过算法 GA-Net 进行社区发现，再基于上述的基准社区计算其各自对应的 *NMI* 值（NMI_{TTCA} 和 NMI_{MMSEA}）并列入表 4 中。对比表 4 中的每一列数据可以看到，NMI_{TTCA} 值均显著优于 NMI_{MMSEA} 值，且较接近于 1，说明在不作调优处理的情况下，基于 TTCA 粗化得到的网络图所发现的社区与基准社区很相近，即基于三角形内点同一社区性的粗化策略充分保持了网络粗化前后其基本社区结构的一致性，因此可以显著改善多层社区发现算法的性能，提高社区发现精度。而 MMSEA 不能保证上述结论，其较低的 NMI_{MMSEA} 值显示，经过 MMSEA 粗化后的社区发现结果与基准社区相差较大，因此反粗化阶段需调用大量的调优工作以提高多层社区发现算法的计算精度。

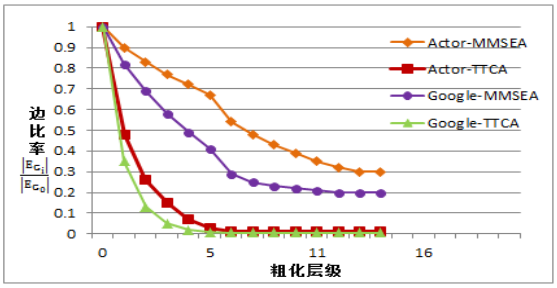
表 4　粗化算法对社区发现结果的影响

NMI	Collaboration	Email	Flickr
NMI_{TTCA}	0.957	0.932	0.941
NMI_{MMSEA}	0.633	0.605	0.619

此外，为了证明 TTCA 在分析节点度服从幂律分布的社会信息网络时性能优于 MMSEA，测试又针对表 2 中列出的两个幂律分布大规模网络 Actor 和 Google，分别采用 TTCA 和 MMSEA 作逐层迭代粗化。两种粗化算法逐层迭代生成的粗化图 G_i 上的节点数和边数，相较于初始网络图 G_0 上的节点数和边数，其比率变化如图 6 所示（该比率是 $|V_{Gi}|/|V_{Go}|$ 或者 $|E_{Gi}|/|E_{Go}|$，而不是粗化缩减比率 $|V_{Gi+1}|/|V_{Gi}|$ 或者 $|E_{Gi+1}|/|E_{Gi}|$，但两者反映的算法粗化本质是一致的）。



(a)　$|V_{Gi}|/|V_{Go}|$



(b)　$|E_{Gi}|/|E_{Go}|$

图 6　幂律分布网络粗化性能的比较

从图 6 中可以看到，同一纵坐标下，经过相同层数的粗化， MMSEA 保留的节点数和边数大于 TTCA；同一横坐标下，要达到相同程度的粗化，MMSEA 需经过更多次迭代粗化操作，这充分说明了采用基于三角形内点同一社区性粗化策略的 TTCA 其粗化幂律分布网络的能力高于 MMSEA，并以大于 1/2 的高粗化缩减比率（该值可通过计算 $(|V_{Gi+1}|/|V_{Go}|)/(|V_{Gi}|/|V_{Go}|)$ 或 $(|E_{Gi+1}|/|E_{Go}|)/(|E_{Gi}|/|E_{Go}|)$ 得到，而 MMSEA 的粗化缩减比率小于 1/2）加快了网络粗化过程。另外，图 6 中 MMSEA 比率折线的渐近值明显高于 TTCA 比率折线的渐近值。实验结果表明：TTCA 选择从度最低的节点开始由边缘向中心逐渐推进式的遍历三角形的粗化方式，有效避免了 MMSEA 极大匹配选边的随机性以及因度高节点的中心占优势而过早终止的粗化过程，适用于分析具有幂律分布特性的大规模社会信息网络。

4.3 TMLCD性能评测

为了评估 TMLCD 性能，本组测试选择如下两个具有代表性的社区评分函数[19]来度量 TMLCD 计算结果的优劣：

1. Internal Density Function (IDF):

$$F(C)_{IDF}=\frac{m_C}{n_C(n_C-1)/2} \tag{6}$$

2. Normalized Cut Function (NCF):

$$F(C)_{NCF}=\frac{b_C}{2m_C+b_C}+\frac{b_C}{2(m-m_C)+b_C} \tag{7}$$

以 $G=(V, E)$表示一个无向图，$n=|V|$为图中的节点数，$m=|E|$为边数，C 代表图上的一个子图社区，n_C 为社区 C 中的节点数，m_C 为社区 C 中的边数，b_C 为社区 C 中位于边界的边数，即 $b_C=|\{(u, v) \in E: u\in C，v\in /C\}|$。上述公式(6)中 $F(C)_{IDF}$ 表示社区 C 内部边的连接密度的大小，其值越大表示社区划分效果越好；公式(7)中 $F(C)_{NCF}$ 表示社区 C 与其他社区之间的分离程度，其值越小表示社区划分效果越好。

本组测试将以表 3 中六个大规模社会信息网络为实验研究对象，通过计算上述评分函数来分别测试 TMLCD、Graclus、PMetis（Graclus 和 PMetis 是两种广泛使用的基于极大匹配选边粗化策略实现的多层社区发现算法）以及 GEM[21]（GEM 是一种社区发现抽样近似算法，并非多层社区发现算法）性能，并对比各算法的时间和空间复杂度，实验结果如图 7、图 8 和图 9 所示。目前的 TMLCD 仅限于分析非重叠性社区结构或者重叠性较弱的社区结构，因此，在运用表 3 中数据做社区发现前，需先将网络中的重叠社区部分去重叠化。

分析图 7 中的实验结果，TMLCD 的系统内存占用远低于 Graclus 和 PMetis，其原因在于单层粗化算法 TTCA 的粗化缩减比率高于 MMSEA，即在相同粗化程度要求下，TMLCD 所需的迭代粗化次数和每层需要保存的网络信息量要远小于 Graclus 和 PMetis。而 GEM 的系统内存占用不仅低于 Graclus 和 PMetis，甚至低于 TMLCD，主要是因为 GEM 的抽样近似法大大缩减了其所需存储的网络信息量。另外，基于 Amazon 网络的实验结果显示，TMLCD 的内存占用较高，较接近 PMetis，原因是 Amazon 包含的三角形数量较少，可选作粗化对象的三角形数目也相应较少，因此极大地限制了 TMLCD 的粗化性能优势，且系统需要经过更多次的迭代粗化来缩减网络规模和更多的内存空间来存储反粗化时所必需的网络信息。在分析规模为 17M 的 Twitter 网络时，由于 Graclus 采用高计算复杂度的 kernel k-means 算法作反粗化调优，严重阻碍了其分析超大规模网络的可扩展性能，因此图 7 中的 Twitter 网络未显示 Graclus 的内存占用结果。

对比图 8 中的实验结果，TMLCD 的运行时间远小于 Graclus 和 PMetis，原因在于实验预处理中已将所有社会信息网络转变成了不包含重叠社区结构的网络，TMLCD 在分析这些网络时，可直接将初始社区发现结果作反粗化还原而无需任何的调优处理，加之 TMLCD 具有高粗化缩减比率和网络基本拓扑结构的高保持度，因此大大降低了其时间消耗；而 Graclus 和 PMetis 反粗化阶段需要大量的调优工作来重新界定反粗化展开的自由节点的社区归属性，因此其二者的计算时间复杂度较高。图 8 中 Graclus 的运行时间高于 PMetis，其主要原因是前者采用了更加复杂的 kernel k-means 算法分析网络，从而导致了高额的时间开销，且 Twitter 网络仍未能显示 Graclus 的运行时间结果。虽然抽样近似方法加快了 GEM 发现大规模网络社区结构的速度，但图 8 中的 GEM 时间损耗仍高于 TMLCD，这是因为 TMLCD 省去了大量反粗化调优计算过程。另外，基于 Amazon 网络的运行时间再一次验证了 TMLCD 适用于分析富含三角形的大规模网络，对比观察图 8 中实验结果，除了 Amazon 以外，TMLCD 分析其他网络的运行时间均小于 PMetis。

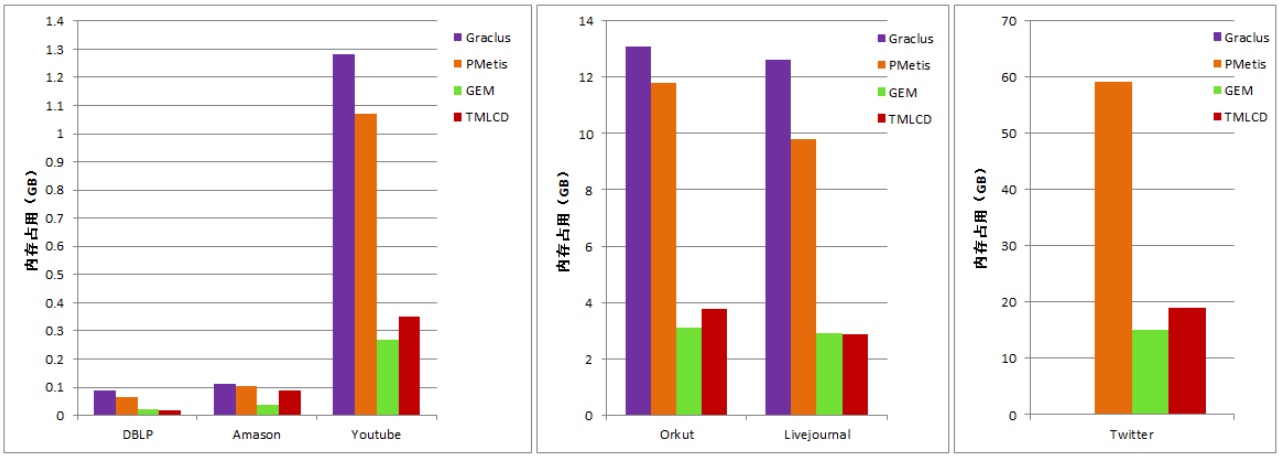


图 7　内存占用的比较

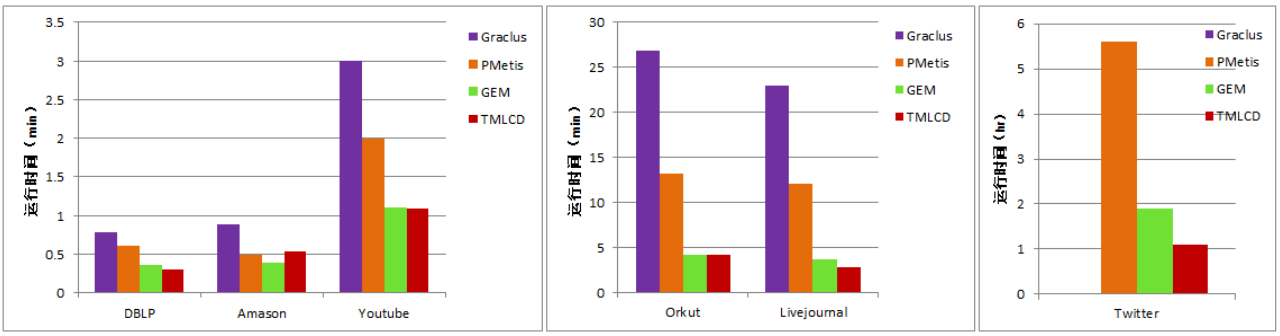
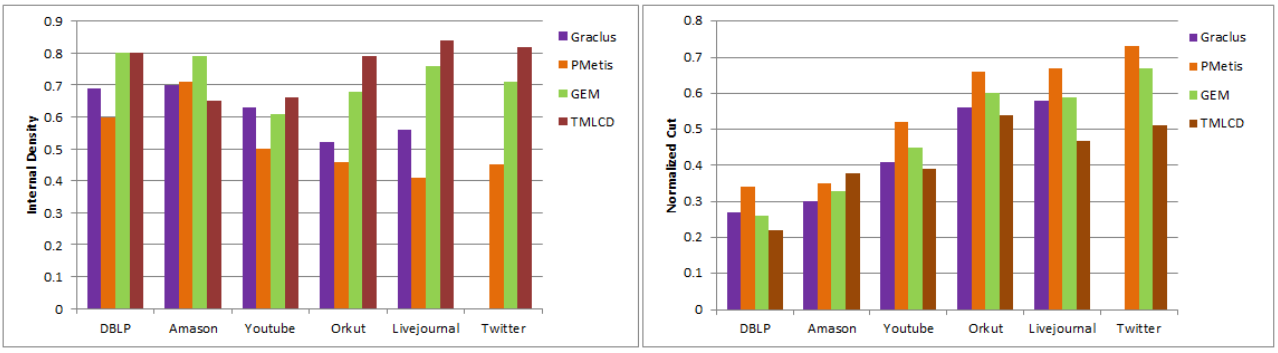


图 8　运行时间的比较



(a) Internal Density Function　　(b) Normalized Cut Function

图 9　社区发现质量的比较

图 9 中对比显示了 Graclus、PMetis、GEM 和 TMLCD 在分析大规模社会信息网络社区结构时的质量评估函数值 IDF 和 NCF 的计算结果。对比分析图 9 中的直方图，除了 Amazon 以外，TMLCD 性能均高于 Graclus 和 PMetis，其原因是 Amazon 包含的可选作粗化对象的三角形数目较少，限制了 TMLCD 的计算性能优势；而 Twitter 上 Graclus 性能结果的再次缺失，仍是因为 kernel k-means 算法过高的计算复杂度。GEM 虽然具有较高的可扩展性，适合于分析大规模甚至超大规模网络，但却是一种随机抽样近似方法，极易丢失网络中的社区结构信息，其社区发现结果精度低于 TMLCD。因此，基于三角形内点同一社区性的粗化策略以对网络基本社区结构的高保持度，极大地提高了 TMLCD 计算大规模社会信息网络社区结构的精度。

5　结论与展望

本文提出了一种面向大规模社会信息网络的基于三角形的多层社区发现算法 TMLCD。该算法

采用基于三角形内点同一社区性的粗化策略，将网络上遍历到的三角形三个顶点融合形成一个复合节点来实现粗化，不仅从基本拓扑结构上保持了初始网络的社区效应，提高了社区发现精度，而且以大于 1/2 的粗化缩减比率加快了网络粗化过程，节省了系统的时间与空间开销，同时其反粗化阶段无需大量的调优工作，进一步降低了算法整体的计算时间复杂度。实验结果表明，TMLCD 的计算精度、内存占用以及时间开销均优于 PMetis 和 Graclus，适用于分析富含三角形的大规模社会信息网络。未来我们将在反粗化阶段引入可分离重叠社区结构的调优算法，使 TMLCD 发现不同划分度要求下的自然社区，拓展其应用需求；亦或基于 MapReduce 实现 TMLCD 的并行化计算，进一步提升其扩展性以分析更大规模的社会信息网络社区结构。

参 考 文 献

[1] Ng A Y, Jordan M I, Weiss Y. On spectral clustering: analysis and an algorithm. Advances in Neural Information Processing Systems, 2002, 2: 849-856.

[2] Ravasz E, Barabsi A L. Hierarchical organization in complex networks. Physical Review E, 2003, 67(2): 026112.

[3] Girvan M, Newman M E J. Community structure in social and biological networks. Proceedings of the National Academy of Sciences, 2002, 99(12): 7821-7826.

[4] Schuetz P, Caflisch A. Efficient modularity optimization by multistep greedy algorithm and vertex mover refinement. Physical Review E, 2008, 77(4): 046112.

[5] Handcock M S, Raftery A E, Tantrum J M. Model based clustering for social networks. Journal of the Royal Statistical Society: Series A (Statistics in Society), 2007, 170(2): 301-354.

[6] Lipczak M, Milios E. Agglomerative genetic algorithm for clustering in social networks.//Proceedings of the 11th Annual conference on Genetic and evolutionary computation. Montreal, Canada, 2009: 1243-1250.

[7] Ziv E, Middendorf M, Wiggins C H. Information-theoretic approach to network modularity. Physical Review E, 2005, 71(4): 046117.

[8] Charikar M, O'Callaghan L, Panigrahy R. Better streaming algorithms for clustering problems.//Proceedings of the Thirty-fifth Annual ACM Symposium on Theory of Computing. San Diego, USA, 2003: 30-39.

[9] Satuluri V, Parthasarathy S, Ruan Y. Local graph sparsification for scalable clustering.//Proceedings of the 2011 ACM SIGMOD International Conference on Management of data. Athens, Greece, 2011: 721-732.

[10] Karypis G, Kumar V. Parallel multilevel k-way partitioning scheme for irregular graphs. SIAM Review, 1999, 41(2): 278-300.

[11] Karypis G, Kumar V. A fast and high quality multilevel scheme for partitioning irregular graphs. SIAM Journal on Scientific Computing, 1998, 20(1): 359-392.

[12] Dhillon I S, Guan Y, Kulis B. Weighted graph cuts without eigenvectors: a multilevel approach. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(11): 1944-1957.

[13] Abou-Rjeili A, Karypis G. Multilevel algorithms for partitioning power-law graphs. IEEE International Parallel and Distributed Processing Symposium. Rhodes Island, Greece, 2006.

[14] Kumpula J M, Kivelä M, Kaski K, Saramäki J. Sequential algorithm for fast clique percolation. Physical Review E, 2008, 78 (2):026109.

[15] Sales-Pardo M, Guimerà R, Moreira A A, Amaral L A N. Extracting the hierarchical organization of complex systems. Proceedings of the National Academy of Sciences, 2007, 104: 15224-15229.

[16] Schank T, Wagner D. Finding, counting and listing all triangles in large graphs.//Proceedings of the 4th International Conference on Experimental and Efficient Algorithm. Santorini Island, Greece, 2005, 606-609.

[17] Pizzuti C. GA-Net: A genetic algorithm for community detection in social networks.//Proceedings of the 10th International Conference on Parallel Problem Solving from Nature X. Dortmund, Germany, 2008, 5199: 1081-1090.

[18] Hendrickson B, Leland R. A multilevel algorithm for partitioning graphs.//Proceedings of the 1995 ACM/IEEE on Supercomputing. San Diego, USA, 1995: 28-es.

[19] Yang J, Leskovec J. Defining and evaluating network communities based on ground-truth.//Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics. Beijing, China, 2012: 1-8.

[20] Danon L, Duch J, Diaz-Guilera A, Arenas A. Comparing community structure identification. Journal of Statistical Mechanics: Theory and Experiment, 2005: P09008.

[21] Whang J J, Sui X, Dhillon I S. Scalable and memory-efficient clustering of large-scale social networks.//Proceedings of the 12th IEEE International Conference on Data Mining. Brussels, Belgium, 2012, 705-714.

KANG Ying, born in 1984, Ph.D. candidate, E-mail: kangying@iie.ac.cn. Her current research interests include data mining, community detection etc.

YU Bo, born in 1985, Ph.D., assistant researcher, E-mail: yubo@iie.ac.cn. His current research interests include wireless networks, wireless sensor networks, distributed computing, query processing etc.

LIN Zheng, born in 1984, Ph.D., assistant researcher, E-mail: linzheng@iie.ac.cn. Her current research interests include nature language processing, sentiment analysis etc.

WANG Wei-Ping, born in 1975, Ph.D., professor and doctoral supervisor, E-mail: wangweiping@iie.ac.cn. His main research interests include data stream, high performance database and parallel processing etc.

ZHOU Jiang, born in 1980, Ph.D., assistant researcher, E-mail: zhoujiang@iie.ac.cn. His current research interests include mass data storage, distributed file system and high available system etc.

MENG Dan, born in 1965, Ph.D., professor and doctoral supervisor, E-mail: mengdan@iie.ac.cn. His main research interests include high performance computer architecture, distributed file system and system security etc.

Background

Along with the expansion of the core concept of Web 2.0, a number of social media like Weibo, Social News, WikiPedia etc. emerge. Based on these interactive social media, a new kind of networks come up which are called Social Information Networks(SINs). At present, the analysis of SINs has been applied in many domains such as friend recommendation, personalized information navigation, resource classification, economic trend forecasting, electronic commerce, marketing management, network security and public opinion monitoring. However, the unparalleled explosion and increasing complexity of SINs make it infeasible to analyze SINs from the viewpoint of node or whole network. SINs have some notable properties like small world, free scale etc., and the most important one is community structure. Taking advantage of community structure, we can study SINs from the perspective of mesoscale, reducing the complication of the network modeling and analysis.

Community has become a breakthrough point for network structure and function analysis. In the past ten years, community detection, as a basic research topic, has attracted plenty of attention from related fields. So many kinds of community detection algorithms are proposed one after another, for example k-means, spectral algorithms, hierarchical clustering, divisive algorithms, modularity-based methods, dynamic algorithms etc. In spite of the efficient application to small networks, these traditional algorithms are unsuitable to analyze large-scale networks due to their computing complexity of $\Theta(n^2)$ or greater.

Facing up to tackle large-scale data networks, approximate computing is necessary. Multilevel community detection is one of the most popular methods to discover the community structure of large-scale networks recently. Although the current multilevel community detection algorithms have the capability to analyze networks involving millions of nodes, the so-called maximum matching edge selecting coarsening policy makes the coarsening shrink rate less than 2 without exception, restricting the scalability of algorithms severely. In comparison, we adopt a remarkable characteristic of triangle, which is that the inner vertices belong to the same community, to design a new coarsening policy and propose a triangle-based multilevel community detection algorithm called TMLCD. During the coarsening phase, TMLCD merges three vertices when traversing a triangle in the network to promote the coarsening shrink rate more than 2 and accelerate the computing speed. Moreover, the performance of community detection has been improved, because TMLCD can keep the basic community effect of the initial network. Experimental results of real networks indicate that, TMLCD outperforms the currently classical multilevel community detection algorithms in terms of computing precision, memory occupation and running time, when analyzing the SINs that are rich in triangles.

This work is supported by the National Science and Technology Support Project of China under grant number 2012BAH46B03, National HeGaoJi Key Project of China under grant number 2013ZX01039-002-001-001, "Strategic Priority Research Program" of the Chinese Academy of Sciences under grant number XDA06030200, and National High Technology Research and Development Program of China (863 Program) under grant number 2012AA01A401.