doi:10.3969/j.issn.1000-436x.2014.10.019

基于强度排序的通信社区检测算法

卫红权，陈鸿昶，刘力雄，兰巨龙
（国家数字交换系统工程技术研究中心 河南 郑州 450002）

摘 要：针对当前电信网中如何有效刻画含权网络的真实特征，完善和发展相关复杂网络模型的难题，特别是对通信社区检测结果层次结构不清晰及运算复杂度高的问题，从复杂网络特征分析入手，设计了一种新的通信社区检测算法。该算法基于通信强度排序方法实现通信社区的有效检出，基于通信密度分布生成高分辨率层次嵌套树，通过距离矢量修剪嵌套树，实现社区稳定检测和层次结构分析同时降低计算复杂度。该算法使用真实网络数据进行了有效验证。

关键词：复杂网络；电信网；通信强度；层次结构；通信社区
中图分类号：TN915.0 **文献标识码**：A **文章编号**：1000-436X(2014)10-0165-06

Communication community detection algorithm based on ranking of strength

WEI Hong-quan, CHEN Hong-chang, LIU Li-xiong, LAN Ju-long
（National Digital Switching System Engineering & Technological Research Center, Zhengzhou 450002, China）

Abstract: According to the characteristics of how to effectively describe real weighted network of the current telecom network problems, improvement and development of related models of complex networks, especially for communication community detection results hierarchy was not clear and the problem of high complexity, from the analysis of the characteristics of complex network, a new algorithm for community detection design communication. The algorithm to achieve effective communication strength ranking method based on community detection in communication, communication density distribution of generating high resolution based on hierarchical nesting tree, the distance vector pruning nested tree, the level of analysis and structure of community stability and reduce the computational complexity. The algorithm is verified using real network data.

Key words: complex network; telecommunications network; communication strength; hierarchy; communication community

1 引言

电信网是人们沟通信息的重要渠道之一，发现电信网中的社区结构有助于理解真实社会关系的多种特征，该领域的研究具有重要的理论意义和应用价值。电信网具有很高的复杂性，是一种“复杂网络（complex network）”[1~3]，分析电信网的社区结构本质上是对复杂网络进行聚类及簇结构检测[4,5]。近年来在该方面进行了许多研究，但目前尚没有一种具有普适意义的方法能揭示各种复杂网络呈现出簇结构的多样性。此外，针对复杂网络社区检测的大部分研究集中在无权网络上，Barabaci 和 Albert 提出的无标度网络[6]和 Watts 和 Strogatz 提出的小世界网络[1]及其各种变种是最具代表性的无权网络，但此类研究只能反映出网络节点间简单连接方式下的相互作用、簇结构等特性。而在电信网中，人与人的通信频度反映了其联系的强度，是一个含权网络，无权网络模型不能反映实际网络节点间相互作用的强度和连边的差异性。并且电信网中真实社区成员之间往往呈现出某种层次化关系。简而言

收稿日期：2014-01-10；修回日期：2014-04-15
基金项目：国家重点基础研究发展计划（“973”计划）基金资助项目（2012CB315905）；国家自然科学基金资助项目（61171108）
Foundation Item: The National Basic Research Program of China(973 Program) (2012CB315905);The National Natural Science Foundation of China(61171108)

之，就是直接上下级之间会产生直接通信关系，跨层级之间的成员存在着间接通信关系，这种间接通信关系对于揭示成员的层次化结构具有相当意义。如何能有效地刻画含权网络的真实特征，完善和发展相关复杂网络模型，是个极富挑战性的课题。

2 复杂网络社区发现算法研究现状

2002 年，Girvan 和 Newman 发表了一篇具有开创意义的论文，设计了一种启发式的社区发现算法（GN 算法）。GN 算法利用不断移除簇间的边从而识别复杂网络中的社区结构，其启发式规则为：社区间边的边介数应大于社区内部边的边介数。所谓边介数，就是经过网络中某条边的任意两节点间最短路径数[7]，GN 算法的提出掀起了复杂网络社区发现研究的热潮。

复杂网络研究至今已经出现了很多社区发现方法，这些方法大致可分为以下几大类：第 1 类是基于图分割的算法，代表算法有 K-L 算法[8]、派系过滤算法[9]和随机游走算法[10]等；第 2 类是基于层次聚类的算法，代表算法有基于边介数度量的分裂算法[11]和基于相似度测量的凝聚算法[12]等；第 3 类是谱方法[13,14]，代表算法有谱平分法[15]；后来，学者们结合其他学科角度也提出了一些算法，如基于 PCA 的算法[16]、基于电阻网络特性的算法[17]、基于信息论的算法[18]和最大化模块度[19]的算法等。现有很多社区发现算法，一般都基于节点之间的聚集度、相似度。比如，基于边聚集度检测社区（EBC）[7]是广泛采用的算法，但该算法只考虑相邻节点相互之间的影响，比如在传销类网络中，只要是一条“联络线”上的节点，即便没有直接通信关系，它们之间都应该存在着影响力。文献[20]提出了一种基于拓扑势的社区发现算法(CDOTP, community discovery method in networks based on topological potential)，该方法引入拓扑势描述网络节点间的相互作用，将每个社区视为拓扑势场的局部高势区，通过寻找被低势区域所分割的连通高势区域实现网络的社区划分，但该方法没有考虑节点间的联络强度，也没有给出分析社区内组织结构的方法。

这些方法存在 2 个问题：一是如何因顶点在社区之间移动导致快速增长的分区数量引入的大量噪声，模糊了稳定分区的明显特征，导致评估大型网络的稳定性困难；二是这些方法多具有很高的时间复杂性 $O(m^2n)$，即便是采用模拟退火或者极值最优化等优化方法辅助计算，其收敛速度通常也很缓慢。

3 OPFICS 算法设计与分析

本文针对以上社区发现方法所存在的主要问题，在研究电信网用户行为的特征和电信网作为复杂网络所呈现出来的“聚类”特性基础上，设计了一种基于通信强度排序的社区检测与结构分析 OPFICS（ordering points by communication frequecy to identify communities structure）算法。OPFICS 算法从网络含权节点度属性分析出发，以节点（用户）间的通信强度作为度量指标，定义了网络的可达通信距离和核心通信距离，通过对节点（用户）的可达通信距离排序给出具有高分辨率的社区检测结果。其中节点密度高的社区揭示了核心成员的位置，该算法所构造的不同分辨率嵌套树则给出了社区的层次关系，对层次树的修剪算法可以清晰揭示有代表性的通信社区。

3.1 通信社区检测

电信网由通信个体和通信关系组成，分析其通信社区结构是一个典型的复杂网络社区挖掘问题，2 个通信个体之间的通信次数决定了它们的联络强度，所以在建立数学模型时，通信个体之间的边是带权边，其权值与通信次数相关。作为复杂网络的一个实例，通信社区可以建立如下模型。

电信网中用户构成图 $G=(V,E,W)$，其中，V 表示电信网中节点（用户）集合，E 表示节点（用户）的连接关系集合，W 为权重矩阵向量，表示用户间通信频度。为简化分析流程，本文不考虑通话是由那个节点（用户）发起的，即图 G 是无向图。通信社区定义为图 G 的稠密连通分支，该分支具有社区间连接稀疏，社区内连接稠密的特点。

算法中定义电信网中通信距离为通信节点之间连接边的权值，此权值与两节点间联络强度（所有通信方式的通信次数之和）负相关。这里所讨论的基于通信强度通信社区，就是由存在通信联系节点所构成的集合，在这些集合之外的节点是孤立点，不予考虑。OPFICS 算法是基于核心成员的通信近邻和节点间的通信联系来构建通信社区模型，通过聚类算法找到一个未归入任何社区的核心成员，以该成员为初始对象建立通信社区；然后经过迭代运算将其通信成员归入社区中，算法执行过程中，对间接通信成员进行合并处理，当没有新的成员可以加到该社区时，一次归并过程结束，算法继

续寻找下一个没有被归入任何社区的核心成员，重复该过程直至遍历所有节点，最后对没有归入任何类的非核心节点归入孤立点集合。

需要注意的是，最核心的成员不一定是拥有最多邻居的成员，而是在整个社区影响力最大的成员，为了在通信社区中找出影响力最大的节点作为为该社区中最核心的成员，这里引入核心通信距离和可达通信距离，相关定义如下。

定义 1　通信强度。某段时间内用户 p 与 q 的通信强度为 p 与 q 之间所有通信时长之和。

$$Comm\_Count(p,q)=\sum CDR_{p,q}.length \tag{1}$$

定义 2　直接通信距离。若通信强度 $Comm\_Count(p,q)>0$，表明节点 p,q 之间存在直接通信关系，定义其通信距离为通信次数的倒数，联系越紧密，距离越小，这符合对关系距离的认识。

$$Comm\_Distance(p,q)=\begin{cases}0, p==q\\ 1/Comm\_Count(p,q), p!=q\end{cases} \tag{2}$$

定义 3　通信圈。所有与节点 p 的通信距离小于 ε 的节点构成节点 p 的通信圈。

$$Neigbour(p)=\{v_m, d(p,v_m)<\varepsilon\} \tag{3}$$

定义 4　核心节点。如果节点 p 的通信圈中至少包含了 $MinPts$ 个成员，即 $|Neigbour(p)|>=MinPts$，则称节点 p 为核心接点。

定义 5　核心通信距离。节点 p 的核心通信距离是使 p 成为核心通信节点的最小 ε'。如果 p 不是核心成员，则 p 的核心通信距离没有定义。

$$Core\text{-}Distancee_{\varepsilon,MinPts}(p)=\begin{cases}\infty, |Neigbour(p)|\geqslant MinPts\\ \min\{\varepsilon\in R^{+}:|\{q\in Neigbour(p):\\ Comm\_Distance(p,q)\leqslant\varepsilon\}|\geqslant MinPts\},\\ |Neigbour(p)|\geqslant MinPts\end{cases} \tag{4}$$

定义 6　可达通信距离。节点 q 到节点 p 的可达通信距离是 p 的核心通信距离和 $Comm\_Distance(p,q)$ 的较大值，如果 p 不是核心节点，则 p 和 q 之间的可达通信距离没有定义。

$$Relation\_Distance(p,q)=\max(Core\text{-}Distancee_{\varepsilon,MinPts}(p), Comm\_Distance(p,q)) \tag{5}$$

算法在每增加一个核心节点及其通信圈成员时，根据通信圈成员相对于该核心节点的可达通信距离进行排序，该排序结果以数组方式存储，算法结束后对该数组进行可视化显示，结果如图 1 所示。

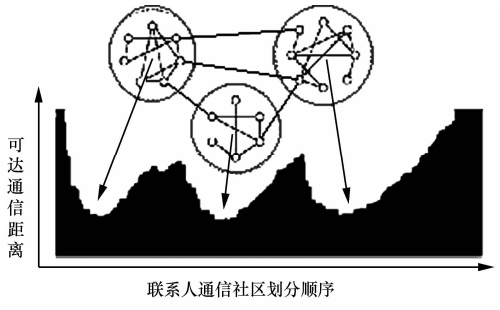


图 1　可达通信距离排序社区划分

图 1 根据通信节点各自的可达通信距离按社区顺序绘出，图中的高斯波“谷”表示某通信社区，其谷底节点表示该社区中的核心成员。

3.2　组织结构生成

上述算法生成通信社区的节点（通信成员）序列组，算法执行过程中同步记录了可达通信距离和核心通信距离，只要通过算法输入一个足够大且有意义的 ε，则最后的结果中就会包含所有可能的 ε' 通信社区 $(\varepsilon'<\varepsilon)$，这些通信社区构成相应的嵌套层次树。该层次树揭示了通信社区中的层次化组织结构关系。但对于一系列非常接近 ε' 对应的通信社区来说，其层间距离非常近，嵌套的层次特征不明显（如图 2 中处于中间位置的 2 个通信社区）。

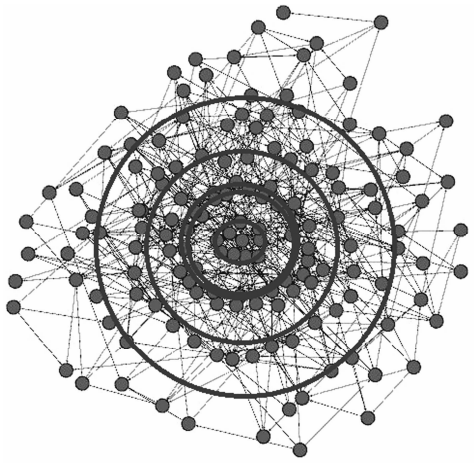


图 2　不同 ε' 得到的社区嵌套关系

为了形成清晰的社区嵌套层次结构，本文对结果集中接近的 ε' 进行修剪，只保留具有代表性的通信社区。下面给出社区嵌套层次修剪方法。

1) 初始嵌套层次树构建

通过前述算法可知，每一个 ε' 社区中的通信成员是按可达通信距离排序的，其父社区（阈值为 ε_0）包含了 $n(n>0)$ 个子社区（阈值为 $\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_m$），则

$$\varepsilon_0 > \varepsilon_i, i=0,1,\cdots,m \tag{6}$$

算法在遍历结果序列过程中，当从子社区 i 转移到子社区 $i+1$ 时，其通信成员的可达通信距离排序关系会发生变化，由此特性来设计初始嵌套层次树构建算法。其中层次树的节点（代表了可能的社区）排序关系结果可以用堆结构存放。

当算法执行结束时，堆中保留的通信成员指示出所有可能的 ε' 通信社区在结果序列中的起始点，且该点的可达通信距离代表了社区的密度 ε'，所有可能的社区嵌套关系可以通过遍历堆结构树得到，堆顶为树根节点对应的通信成员。

2) 嵌套层次树的修剪优化

在修剪初始嵌套层次树时，首先对其数据进行扩展，每个节点用于表示可能的社区，节点中对应的可达通信距离即代表了社区的密度 ε'，修剪的目的就是清除树中与父节点的可达通信距离差距很小的节点。考察初始嵌套层次树中各节点可达通信距离（即社区的密度 ε'）的变化情况，记节点 i 与其父节点 $i-1$ 间可达通信距离的差为

$$Delta_{rd}(i) = RD(i) - RD(i-1) \tag{7}$$

$Delta_{rd}(i)$ 变化比较平稳时，意味着这一段节点对应的这些可能的 ε' 社区差别不大，当 $Delta_{rd}(i)$ 出现突变时，意味着一个差别较大的社区出现了，即：$Delta_{rd}(p\_s)$ 为函数 $Delta_{rd}(i)$ 的局部极大值时，p_s 为修剪后社区的起点。如图 3 所示。

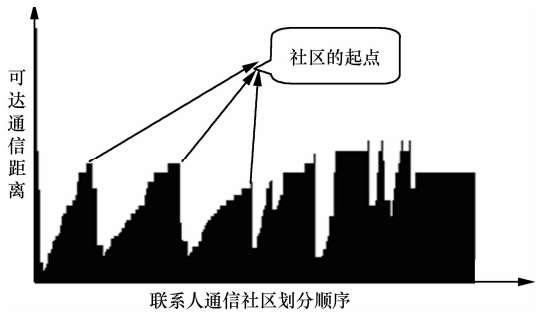


图 3 修剪优化目标

另一方面，将树中节点 i 的权值 $Weight(i)$ 定义为该节点对应的 ε' 社区中通信成员数量，子节点 i 包含的通信成员数量与父节点 $i-1$ 的比值 $Ratio_{Weight}(i)$ 如果趋于 1（肯定小于 1），即

$$(Ratio(i)_{Weight} = \frac{Weight(i)}{Weight(i-1)}) \longrightarrow 1 \tag{8}$$

则表明这 2 个社区之间的差别较小，节点 i 有可能需要被修剪，而如果远小于 1，则表明这 2 个社区之间的差别较大，节点 i 应该予以保留，换句话说，$Ratio_{Weight}(i)$ 接近 1 时，可以认为 $Ratio_{Weight}(i)$ 对判断 i 是否需要修剪的影响较大，反之亦然，因此可以将 $Ratio_{Weight}(i)$ 作为判断 $Delta_{rd}(i)$ 出现极大值的影响因子。

4 OPFICS 算法实证分析

基于信息理论，本文将标准共同信息（NMI, normalized mutual information）的衡量引入到社区结构的比较中，这种方法比简单评价社区数目的正确率更有意义。标准共同信息评价法的过程是：定义矩阵 $\boldsymbol{N}$，行表示为真实的社区编号，列表示检测结果中的社区编号，$N_{i,j}$ 表示在真实社区 i 中出现又在检测结果社区 j 中出现的顶点个数。基于信息理论得到的真实社区结构和检测结果社区结构的 NMI 为

$$NMI = \frac{-2\sum_{i=1}^{C_A}\sum_{j=1}^{C_B} N_{i,j}\log\left(\frac{N_{i,j}\boldsymbol{N}}{N_{i.}N_{.j}}\right)}{\sum_{i=1}^{C_A} N_i \log\left(\frac{N_i}{\boldsymbol{N}}\right) + \sum_{j=1}^{C_B} N_j \log\left(\frac{N_j}{\boldsymbol{N}}\right)} \tag{9}$$

其中，C_A 表示真实社区的个数，C_B 表示检测结果中的社区个数，$N_{i.}$ 表示 $\boldsymbol{N}$ 中第 i 行的和，$N_{.j}$ 表示 $\boldsymbol{N}$ 中第 j 列的和。

当检测结果与真实网络完全一致时，NMI 达到最大值 1；当检测结果与真实网络完全不一致时，NMI 达到最小值 0。

OPFICS 算法的实证采用某大学 PBX（用户交换机）2012 年 1 月 1 日至 2014 年 1 月 1 日 2 年的实际通话详细记录（CDR）作为实证数据，数据集中包含了 1 852 个电话号码（联系人），共 135 万条通话记录，CDR 数据存入数据库，利用数据库统计出联系人两两之间的通信次数，从而得出通信网络图，顶点表示联系人，2 个联系人之间有通话记录，就在对应顶点之间建立一条无向边，其权值为他们之间的通话次数。

测试的过程如下：首先计算出联系人之间的通话次数，根据本文通信距离定义，计算出联系人之间边的权重，然后采用基于 Fibonacci Heap 的

Dijkstra 算法求出联系人两两间的最短路径，最后应用本文的算法进行分析。

实证过程中显示，在计算出联系人直接通信情况及其两两之间的最短路径后，数据集显示出节点服从幂率分布，是典型的无标度网络，同时两两节点间的平均路径长度较短，具有小世界网络的特征。由于其规模和社会网络的特点，适合采用通信强度来进行通信社区分析，如图 4 所示。

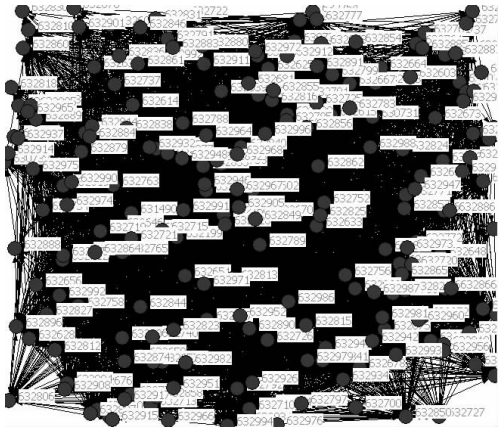
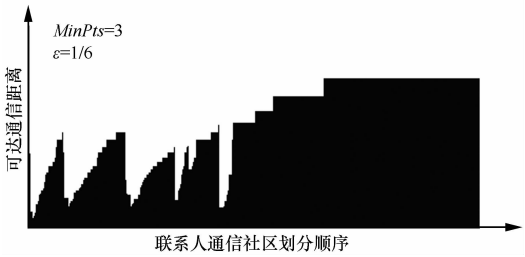
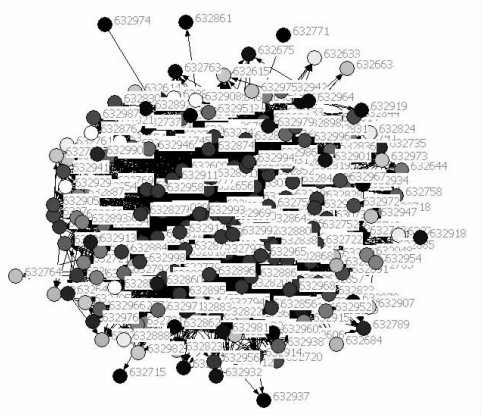


图 4　原始网络连接拓扑

采用本文设计算法，选择不同的 *MinPts* 和 ε 情况下得出的社区检测结果如图 5 和图 6 所示。

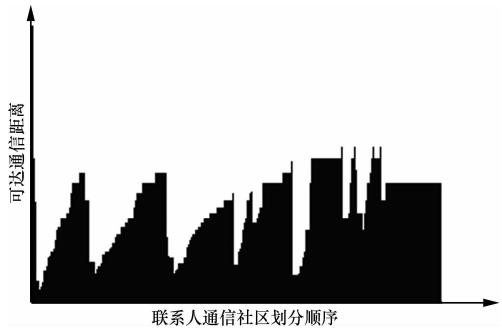


(a) 社区顺序划分结果

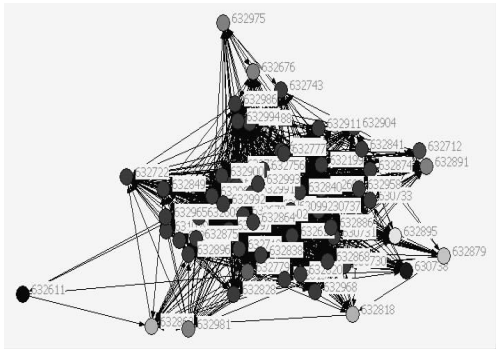


(b) 各社区包含节点构成结果

图 5　结果示意 1（在 $MinPts=3,\varepsilon=1/6$ 条件下）



(a) 社区顺序划分结果



(b) 各社区包含节点构成结果

图 6　结果示意 2（在 $MinPts=3,\varepsilon=1/10$ 条件下）

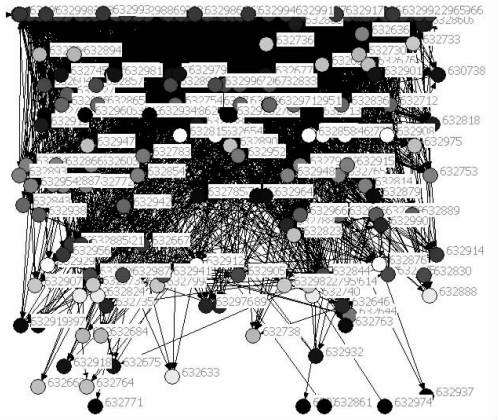


图 7　社区层次结构

如表 1 所示，当 *MinPts* 设为 3，ε 设为 1/6 时，大部分联系人都属于一个大的社区，当 *MinPts* 设为 3，ε 设为 1/10 时，则可以得出所有联系人中最核心的通信成员。对比 EBC 的结果，EBC 显示出 120 多个社区，而其中孤立点社区占 76%，很多 EBC 社区其实只包含了一个成员，而本文的算法倾向于将小的子社区合并成大社区，检测出一个相对较大的社区和几个小的社区，更符合“集团”的通信关系特点。而 CDOTP 能给出和本文算法差不多的社区结构，但其计算量较大，且没有给出社区内层次组织结构。这 3 种方法都能分析出通信社区结构，但本文的算法选用

较小的ε时能分析出小的社区粒度，揭示出小而密集的群体，这代表了通信社区中的核心成员。另外，由于对可达通信距离进行排序时采用了空间索引技术，其时间复杂度能达到（$n\lg n$），本文所设计的 OPFICS 算法优于 2 种对比参考算法。

表 1　不同 *MinPts* 和 ε 的社区划分结果

社区成员	参数			
	MinPts=3 ε = 1/6	*MinPts*=6 ε = 1/10	*MinPts*=3 ε = 1/10	*MinPts*=3 ε = 1/20
1	207	30	27	41
2	12	25	16	23
3	66	14	23	41
4	7	23	9	28
5	9	77	77	14
6	7	2	1	7
7	7	1	无	无

5　结束语

本文在电信网中引入通信社区的概念，并给出了一种检测社区并分析社区核心成员及层次结构的算法 OPFICS。该算法不仅考虑了节点间通信关系的有无，还考虑了其间的通信强度，采用加权的节点度作为度量特征，更好地反映了电信网节点间的亲密程度，由此分析出的社区结构也较好地映射了真实的社会关系网络。同时，该算法还给出社区的层次化结构结果，并由此分析出社区中的核心成员。使用人工和现网数据测试的结果表明，OPFICS 算法能有效检测出电信网的通信社区并准确给出社区的层次关系及核心成员节点。本文所做的工作为进一步对通信融合网络中的社区发现及其多维多层结构研究分析奠定基础。

参考文献：

[1] WSTTS D J, STROGAT S H. Collective dynamics of small-world networks[J]. Nature, 1998,393(6638):440-442.

[2] BARABÁSI A L, ALBERT R. Emergence of scaling in random networks[J]. Science, 1999,286(5439):509-512.

[3] BARABÁSI A L, ALBERT R, JEONG H, BIANCONI G. Power-law distribution of the world wide web[J]. Science, 2000, 287(5461): 2115a.

[4] 陈国强, 王宇平. 分解多目标优化揭示复杂网络社区层次结构[J]. 西安电子科技大学学报, 2013,40(3):205-211.
CHEN G Q,WANG Y P. Revealing of the hierarchy community of the complex network by decomposition multi-objective optimization[J]. Journal of Xidian University, 2013,40(3):205-211.

[5] YANG J, LESKOVEC J. Community-affiliation graph model for overlapping network community detection[A]. Data Mining (ICDM), 2012 IEEE 12th International Conference on Digital Object Identifier[C]. 2012.170-1175

[6] BARABÁSI A, BONABEAU E. Scale-free networks[J]. Scientific American, 2003, 288(5):60-69.

[7] GIRVAN M, NEWMAN M E J. Community structurein social and biological networks[J]. PNAS, 2002, 99(12):7821-7826.

[8] GREGORY S. Fuzzy overlapping communities in networks[J].Journal of Statistical Mechanics: Theory and Experiment, 2011(2):2-17.

[9] PALLA G, FARKAS I, POLLNER P, *et al*. Directed network modules[J]. Phys New J, 2007,186.

[10] NEWMAN M E J. Fast algorithm for detecting community structure in networks[J]. Physical Review E, 2004, 69(6): 066133.

[11] AHN, YY, BAGROW JP, LEHMANN S. Link communities reveal multi-scale complexity in networks[J]. Nature, 2010, 466: 761-764.

[12] MEJ Ne. Modularity and communities structure in networks[J]. Proc of the National Academy of Science, 2006,103(23):8577-8582.

[13] ZHANG T T, WU B. A method for local community detection by finding core nodes, advances in social networks analysis and mining (ASONAM)[A]. 2012 IEEE/ACM International Conference on Digital Object Identifier[C]. 2012.1171-1176.

[14] QI G J, AGGARWAL C C, HUANG T. Community detection with edge content in social media networks ,data engineering (ICDE)[A]. 2012 IEEE 28th International Conference on Digital Object Identifier[C]. 2012.534-545.

[15] 黄亮. 社会网络中的社区发现与链接预测算法研究[D]. 武汉: 华中科技大学, 2012.
HUANG L. Algorithms of Community Detection and Link Prediction in Social Networks[D]. Wuhan: Huazhong University of Science and Technology, 2012.

[16] 张珊. 复杂网络的节点重要性及社区结构研究[D]. 西安：西安电子科技大学, 2013.
ZHANG S. Research on Data Vital Nodes and Community Structure in Complex Network[D]. Xi'an: Xidian University, 2013.

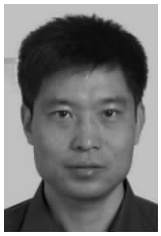
[17] GUO C H, ZHANG L. An analysis method based on PCA for the community structure in complex networks[J]. Operations Research and Management Science, 2008, 17(6):144-149.

[18] NEWMAN M E J, GIRVAN M. Finding and evaluating community structure in networks[J]. Physical Review E, 2004, 69(2):56-58.

[19] LU Z B, WANG J, LI Y Z. An overview on overlapping community detection, Computer Science & Education (ICCSE)[A]. 2012 7th International Conference on Digital Object[C]. 2012.486-490.

[20] 淦文燕, 赫南, 李德毅等. 一种基于拓扑势的网络社区发现方法[J]. 软件学报, 2009, 20(8): 2241-2254.
GAN W Y, HE N, LI D Y, *et al*. Community discovery method in networks based on topological potential[J]. Journal of Software, 2009, 20(8):2241-2254.

作者简介：



卫红权（1971-），男，河南唐河人，国家数字交换系统工程技术研究中心副研究员，主要研究方向为融合网络安全、可重构网络理论与技术。

陈鸿昶（1964-），男，河南新密人，国家数字交换系统工程技术研究中心教授、博士生导师，主要研究方向为通信与信息系统、融合网络安全。

刘力雄（1974-），男，湖南邵阳人，国家数字交换系统工程技术研究中心副教授，主要研究方向为电信网信息安全。

兰巨龙（1962-），男，河北张家口人，国家数字交换系统工程技术研究中心教授、博士生导师，主要研究方向为可重构网络理论与技术。