# 一种改进的加权复杂网络聚类方法

郭 陶 张 琨 郭文娟 庄克琛 贺定龙 李配配
(南京理工大学计算机科学与技术学院 南京 210094)

**摘 要** 在复杂网络聚类中,为了克服聚类结果局部收敛和对多维数据聚类效果差的缺点,通过对复杂网络聚类方法的应用分析,将NJW算法和粒子群聚类算法应用到加权复杂网络簇结构的探测中,设计和实现了一种改进的加权复杂网络聚类方法。实验验证了该方法在簇结构较复杂的网络中具有较高的执行效率和较好的执行效果。

**关键词** 复杂网络,聚类算法,网络簇结构,NJW算法,PSO聚类
**中图法分类号** TP393 **文献标识码** A

## Improved Method of Weighted Complex Networks Clustering

GUO Tao ZHANG Kun GUO Wen-juan ZHUANG Ke-chen HE Ding-long LI Pei-pei
(School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094, China)

**Abstract** In order to overcome the shortcomings of the local convergence and poor results for multidimensional data clustering in complex networks clustering, by applying and analyzing complex networks clustering method, applied NJW algorithm and PSO clustering algorithm to the detection of the cluster structure of the complex networks, so designed and implemented an improved method of weighted complex networks clustering. Experiment demonstrates that the proposed method has high efficiency and good result in the implementation of larger and structure of more complex networks.

**Keywords** Complex networks, Clustering algorithm, Cluster structure of networks, NJW algorithm, PSO clustering

## 1 引言

复杂网络聚类方法的研究不仅对分析复杂网络的拓扑结构、理解复杂网络的功能、发现复杂网络中的隐藏规律以及预测复杂网络的行为具有十分重要的理论意义,而且具有广泛的应用前景。目前聚类算法在复杂生物网络中的应用主要集中在蛋白质交互网络分析预测未知蛋白质功能、新陈代谢网络分析、蛋白质相似性网络分析和基因调控网络分析等领域[1]。复杂网络自20世纪末逐渐兴起以来,正迅速地在深度和广度上与其它学科进行交叉[2,3]。一方面,从现实世界存在的网络中不断地发现新结构与新现象,大量重要的应用问题涌现出来[4,5];另一方面,以研究复杂网络一般规律为目标的理论研究工作也迅速发展,不断提出新的理论模型和新的分析方法[6]。

## 2 预备知识

### 2.1 复杂网络的聚类算法概况

近年来,关于复杂网络的研究正方兴未艾,成为国际上的一个研究热点[7-9]。网络簇结构(network cluster structure)是复杂网络最普遍和最重要的拓扑结构属性之一。复杂网络具有同簇节点相互连接密集、异簇节点相互连接稀疏的特点。复杂网络聚类方法[10,11]旨在揭示复杂网络中真实存在的网络簇结构。

目前已存在多种复杂网络聚类算法[12-14],可以将它们中的大多数归纳为两大类:基于优化的方法(optimization based method)和启发式方法(heuristic method)[13]。前者将复杂网络聚类问题转化为优化问题,通过最优化预定义的目标函数来计算复杂网络的簇结构。后者将复杂网络聚类问题转化为预定义启发式规则的设计问题。

在基于优化的方法中,谱方法采用二次型优化技术最小化预定义的"截"函数。当一个网络被划分为两个子网络时,"截"即指子网间的连接密度。具有最小"截"的划分被认为是最优的网络划分。采用矩阵分析技术,谱方法将求解最小"截"问题转化为求解带约束的二次型优化问题:

$$\min\{(X^{T}MX)/(X^{T}X)\}$$

式中,向量 $X$ 表示网络划分,$M$ 表示对称半正定矩阵。对于平均"截",$M=D-A$ 表示网络的拉普拉斯矩阵(Laplacian matrix),其中 $D$ 表示由节点度构成的对角矩阵,$A$ 为网络的邻接矩阵;对于其他截函数,$M$ 是拉普拉斯矩阵的不同变体。由拉格朗日方法,上述约束二次型的近似最优解(即网络的近似最优划分)可以通过计算 $M$ 的第2小特征向量求得。谱方法本质上是一种二分法,在每次二分过程中,网络被分割成两

本文受国家自然科学基金(61003210),江苏省自然科学基金(BK2010491,BK2011023),南京理工大学"卓越"计划,"紫金之星"项目(20100601)资助。

郭 陶(1988—),男,硕士生,主要研究方向为复杂网络理论及应用,E-mail:guotao715@yahoo.cn;张 琨(1977—),女,副教授,主要研究方向为复杂网络、网络安全。

个近似平衡的子网络。当网络中含有多个簇时，谱方法递归地分割现存的子网络，直到满足预先定义的停止条件为止[13]。与其他方法相比，谱方法具有明显的优势，该方法不仅思想简单、易于实现、不易陷入局部最优解，而且具有识别非凸分布的聚类能力，非常适用于解决许多实际应用问题。

目前将谱方法应用于复杂网络聚类的成果中，有学者利用传统的 K-means 算法对谱方法产生的特征向量进行聚类[15]，而传统的 K-means 算法对初始聚类中心敏感、易于陷于局部优化；有学者采用的复杂网络聚类方法未将谱方法产生的特征向量进行必要的处理[16]，导致其方法的使用范围具有一定的局限性。

在此客观现实下，本文利用谱方法具有的优势，结合现有学者的研究成果[15,16]，以现在最常用、最经典的谱聚类算法——NJW 算法[17]为基础，使用鲁棒性较强的基本 PSO 聚类算法[16]，设计实现了一种改进的加权复杂网络聚类方法——ICMWCN 方法。该方法主要是对加权网络进行聚类，其优点在于谱聚类算法是一种配对聚类方法，仅与数据点的数目有关，而与维数无关，因此可以避免由于特征向量的过高维数所造成的奇异性问题[15,17]，而后选取基本 PSO 聚类算法对数据进行聚类，这是因为基于社会生态学的 PSO 算法主要是在群体的集群行为和自组织原则指导下的随机搜索和优化技术，它强调分布式、相对简单主体之间直接或间接的交互作用，具有很强的适应性和鲁棒性[16]。

### 2.2 社团模块度函数

为定量分析算法的性能，本文引用社团模块度的概念来衡量网络划分质量。社团模块度是 Newman[18] 等人引进的一个衡量网络划分质量的度量标准。假设有某种划分形式，将网络划分为 $k$ 个社团。定义一个 $k\times k$ 维的对称矩阵 $E=(e_{ij})$，其中 $e_{ij}$ 表示网络中连接两个不同社团的节点的边在网络所有边中所占的比例，这两个节点分别位于第 $i$ 和第 $j$ 个社团。模块度衡量标准是利用完整的网络来计算的。模块度 $Q$ 表示为：

$$Q=\sum_i(e_{ii}-a_i^2)=Tre-\|e^2\| \tag{1}$$

式中，$Tre=\sum_i e_{ii}$ 为矩阵 $E$ 对角线上各元素之和，它表示网络中连接某一个社团内部各节点的边在所有边的数目中所占比例；$a_i=\sum_j e_{ij}$ 为每行中各元素之和，它表示与第 $i$ 个社团中的节点相连的边在所有边中所占的比例；$Q$ 值越大说明社团结构越明显。实际网络中，该值通常位于 0.3～0.7 之间。

## 3 基于 NJW 算法的改进加权复杂网络聚类方法

对于某加权复杂网络 $G(V,E,W)$，该网络的节点数为 $n$，给定一个簇的数目 $k$，本文的 ICMWCN 方法的基本步骤如下：

a)由连接权矩阵 $W=[w_{ij}]_{n\times n}$ 得出网络节点间的相似度矩阵 $A=[a_{ij}]_{n\times n}$。若节点 $i$ 和节点 $j$ 之间没有连接或者 $i=j$，则 $a_{ij}=0$，否则 $a_{ij}=1/w_{ij}$。相似度矩阵 $A$ 的度矩阵 $D=[d_i]_{n\times n}$。其中 $D$ 为对角矩阵，$d_i=\sum_j a_{ij}$。注意，这里考虑的均为无向网络。

b)用谱方法将寻找复杂网络的簇结构问题转换为数据的聚类问题，即 NJW 算法的第 1 步。本文采用的是平均截，使用非规范拉普拉斯矩阵，即 $M=D-A$。具体就是计算 $M=D-A$ 的前 $k-1$ 个最小特征值对应的非平凡特征向量 $e_1,e_2,\cdots,e_{k-1}(e_l\in R^{n\times 1},l=1,\cdots,k-1)$ 并组成矩阵 $E=[e_1,\cdots,e_{k-1}]\in R^{n\times(k-1)}$。这里选取前 $k-1$ 个最小特征值对应的非平凡特征向量。这里的拉普拉斯矩阵是对称半正定矩阵[21]，它的特征值是实非负的，并且只存在一个特征值为 0 的特征向量，但存在 $k-1$ 个特征值接近 0，相应的特征向量可以近似构造 NJW 算法第 1 步中的矩阵。

c)将 $E$ 的行向量转变为单位向量，得到矩阵 $F$，即：

$$f_{ij}=\frac{e_{ij}}{\sqrt{\sum_j e_{ij}^2}}$$

得到 $n$ 个数据向量形式的待聚类数据样本集。

d) 将矩阵 $F$ 的每一行看作是 $R_{k-1}$ 空间中的一个点，用基本 PSO 聚类算法，对上面得到的数据样本集 $F$ 进行聚类。

e) 将数据点 $f_i$ 划分到聚类 $j$ 中，当且仅当 $F$ 的第 $i$ 行被划分到聚类 $j$ 中。

该算法的流程图如图 1 所示。

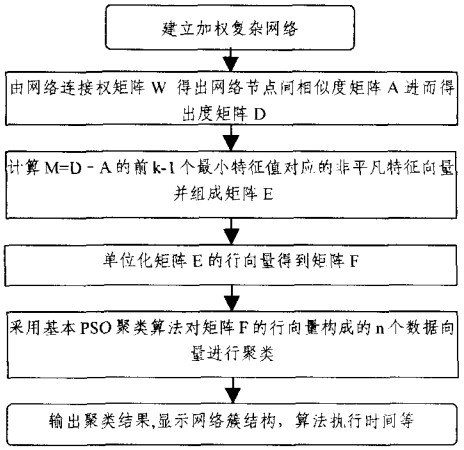
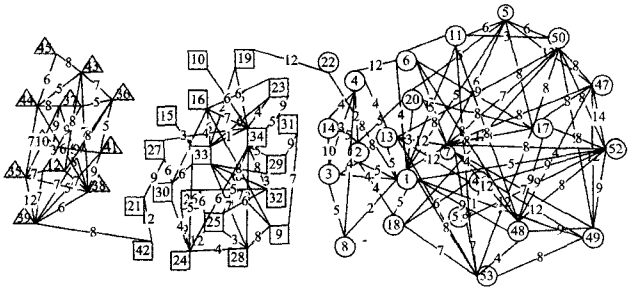
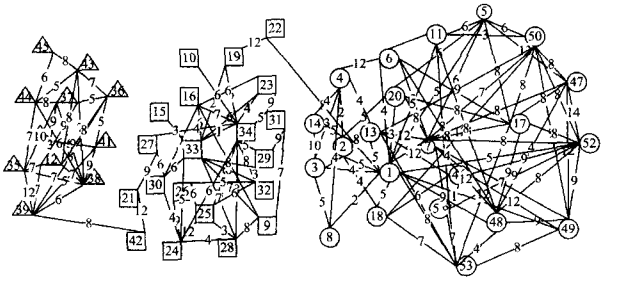



图 1 ICMWCN 方法流程图

## 4 算法仿真与实验



(a) 基于 K-means 聚类算法的复杂网络社团发现新方法对该网络的聚类效果



(b) ICMWCN 方法对该网络的聚类效果

图 2

本文方法采用 VC++6.0 实现，在仿真实验平台上对大量结构复杂程度不同的网络进行测试后发现，ICMWCN 方法与作为对比的基于 K-means 聚类算法的复杂网络社团发现新方法[22]都具有较高的运行效率和精确度。基于 K-means 聚类算法的复杂网络社团发现新方法是将分簇后模块度最大的簇划分结果输出，适用于网络结构较为复杂的网络，且能执行出较好的结果。但是在一些连接紧密、结构较复杂的网络中，本文的 ICMWCN 方法具有更好的执行效果和更高的执行效率。

下面选取一个节点数为 53 的加权复杂网络，对本文的方法进行对比分析。如图 2 所示为运行基于 K-means 聚类算法的复杂网络社团发现新方法和本文的 ICMWCN 方法的聚类效果图，图中圆形顶点表示被划分到簇 1 中，方形顶点表示被划分到簇 2 中，三角形顶点表示被划分到簇 3 中。

基于 K-means 聚类算法的复杂网络社团发现新方法对该网络分簇后，其社团模块度大小为 0.590，整个方法执行时间为 2.129s，分簇结果如表 1(a)所列。

表 1(a)　基于 K-means 聚类算法的复杂网络社团发现新方法对该网络的分簇结果

| 簇号 | 簇 1 | 簇 2 | 簇 3 |
|---|---|---|---|
| 簇中的节点号 | 1-2-3-4-5-6-7-8-11-13-14-17-18-20-22-46-47-48-49-50-51-52-53 | 9-10-15-16-19-21-23-24-25-26-27-28-29-30-31-32-33-34-42 | 12-35-36-37-38-39-40-41-43-44-45 |

本文的 ICMWCN 方法对该网络分簇后，其社团模块度大小为 0.599，整个方法执行时间为 0.602s，分簇结果如表 1(b)所列。

表 1(b)　ICMWCN 方法对该网络的分簇结果

| 簇号 | 簇 1 | 簇 2 | 簇 3 |
|---|---|---|---|
| 簇中的节点号 | 1-2-3-4-5-6-7-8-11-13-14-17-18-20-46-47-48-49-50-51-52-53 | 9-10-15-16-19-21-22-23-24-25-26-27-28-29-30-31-32-33-34-42 | 12-35-36-37-38-39-40-41-43-44-45 |

结果显示，两方法对该网络的 22 号节点的划分产生了差异，其余节点的划分结果完全一致。本文的 ICMWCN 方法执行后，其模块度要高出前者的 1.4%，但执行时间仅为前者的 28.3%。该网络节点间连接紧密、簇结构较为复杂，根据实验结果，与对比的算法相比，本文的方法具有较好的执行效果和更高的执行效率。

再次选取一个节点数为 41 的加权复杂网络，对本文的方法进行对比分析。如图 3 所示为运行基于 K-means 聚类算法的复杂网络社团发现新方法和本文的 ICMWCN 方法的聚类效果图，图中圆形顶点表示被划分到簇 1 中，方形顶点表示被划分到簇 2 中，三角形顶点表示被划分到簇 3 中。

基于 K-means 聚类算法的复杂网络社团发现新方法对该网络分簇后，其社团模块度大小为 0.576，整个方法执行时间为 0.468s，分簇结果如表 2(a)所列。

表 2(a)　基于 K-means 聚类算法的复杂网络社团发现新方法对该网络的分簇结果

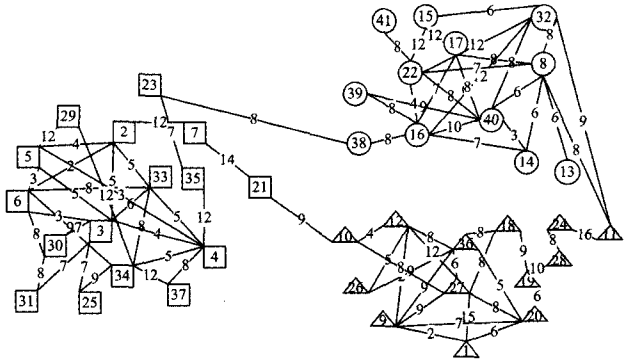
| 簇号 | 簇 1 | 簇 2 | 簇 3 |
|---|---|---|---|
| 簇中的节点号 | 8-13-14-15-16-17-22-32-38-39-40-41 | 2-3-4-5-6-7-21-23-25-29-30-31-33-34-35-37 | 1-9-10-11-12-18-19-20-24-26-27-28-36 |

本文的 ICMWCN 方法对该网络分簇后，其社团模块度大小为 0.598，整个方法执行时间为 0.421s，分簇结果如表 2(b)所列。

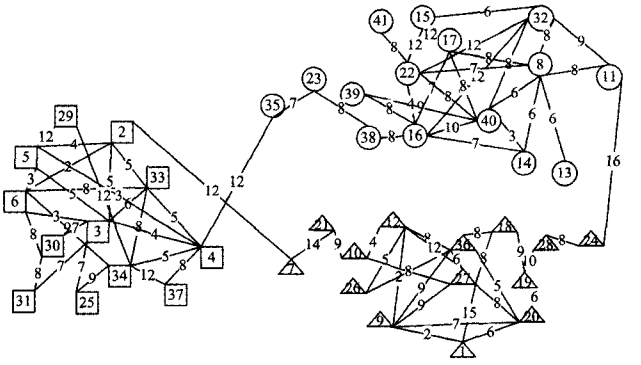
表 2(b)　ICMWCN 方法对该网络的分簇结果

| 簇号 | 簇 1 | 簇 2 | 簇 3 |
|---|---|---|---|
| 簇中的节点号 | 8-11-13-14-15-16-17-22-23-32-35-38-39-40-41 | 2-3-4-5-6-25-29-30-31-33-34-37 | 1-7-9-10-12-18-19-20-21-24-26-27-28-36 |

结果显示，对比方法将 11 号节点划分到簇 3 中，而本文的 ICMWCN 方法将 11 号节点划分到簇 1 中。观察 11 号节点在网络中的位置，它与 8 号、24 号、32 号节点是邻居，8 号和 32 号节点相连且它们的度都很大，而 24 号节点度为 2，与周围节点连接稀疏，因此将 11 号节点划分到包含 8 号和 32 号节点的簇 1 中明显要比将其划分到只包含 24 号节点的簇 3 中更加合理，更加满足簇内节点间连接紧密、簇间节点间连接稀疏的定义。经过计算，本文的 ICMWCN 方法执行后的模块度要高出前者的 3.8%，执行时间为前者的 90.0%，显示出 ICMWCN 方法对该网络分簇有较好的执行效果和较高的运行效率。



(a)　基于 K-means 聚类算法的复杂网络社团发现新方法对该网络的聚类效果



(b) ICMWCN 方法对该网络的聚类效果

图 3

对另外 3 个加权网络进行聚类，对比算法和 ICMWCN 方法的执行效果如表 3 所列。

表 3　基于 K-means 聚类算法的复杂网络社团发现新方法和 ICMWCN 方法对 3 个不同网络的执行效果

| 网络序号 | 节点数 | 边数 | 基于 K-means 聚类算法的复杂网络社团发现方法 | | ICMWCN 方法 | |
|---|---|---|---|---|---|---|
| | | | 执行时间 (s) | 模块度 Q | 执行时间 (s) | 模块度 Q |
| 1 | 40 | 83 | 0.562 | 0.546 | 0.437 | 0.574 |
| 2 | 44 | 90 | 0.832 | 0.643 | 0.490 | 0.650 |
| 3 | 45 | 88 | 0.562 | 0.510 | 0.359 | 0.594 |

根据表3所列的实验结果发现，相比基于K-means聚类算法的复杂网络社团发现新方法，本文的ICMWCN方法在执行时间和聚类精确度上都具有一定的优势，这种优势尤其表现在算法的执行时间和效率上。

本文的ICMWCN方法是将PSO聚类算法和谱方法中的平均截方法相结合的加权复杂网络聚类方法。该方法的算法复杂度为$O(\max\{n^2, p\times t\times n\})$（其中$n$为复杂网络的节点数，$p$为基本PSO聚类算法中初始设置的粒子数，$t$为基本PSO聚类算法的迭代次数），相比基于K-means聚类算法的复杂网络社团发现新方法的算法复杂度$O(m\times n^3)$（其中$m$为复杂网络的边数，$n$为复杂网络的节点数），其算法复杂度低，实现步骤简单。该方法使用基本PSO聚类算法克服了传统K-means聚类算法对初始聚类中心敏感以及易于陷于局部优化等缺点，在结构复杂、连接紧密的网络中有更快的执行速度、更好的聚类效果和精准度。

**结束语** 首先对现有的复杂网络聚类算法进行了简单的分析，然后结合现有的知识设计和实现了基于NJW算法的改进加权复杂网络聚类方法（ICMWCN方法）。通过仿真实验发现，该方法在连接紧密、簇结构较复杂的复杂网络中具有较高的执行效率和较好的执行效果，但是该方法有待改善，以适用于大中型加权复杂网络，并投入到实际工程应用当中。目前复杂网络聚类问题还远未被很好地解决[13]，未来工作可集中在：从网络的“内在”属性出发，给出一种“客观”的理论模型去理解、刻画和计算复杂网络簇结构；设计出快速、高精度和无监督的复杂网络聚类方法；针对特殊类型的复杂网络设计出新型的复杂网络聚类方法。

## 参考文献


[1] 田野，刘大有，杨博. 复杂网络聚类算法在生物网络中的应用[J]. 计算机科学与探索，2010，4(4)：330-337

[2] 周涛，柏文洁，汪秉宏，等. 复杂网络研究概述[J]. 物理，2005，34(1)：31-36

[3] 汪秉宏，周涛，何大韧. 统计物理学与复杂系统研究最新发展趋势分析[J]. 中国基础科学，2005，7(3)：37-43

[4] 汪晓帆. 复杂网络理论及其应用[M]. 北京：清华大学出版社，2006

[5] 谭跃进，吴俊，等. 复杂网络抗毁性研究综述[J]. 系统工程，2006，24(10)

[6] 车宏安，顾基发. 无标度网络及其系统科学意义[J]. 系统工程理论与实践，2004，24(4)：11-16

[7] Watts D J. Networks, dynamics, and the small world phenomenon[J]. AM J Sociol, 1999, 105: 493-592

[8] Albert R, Barabási A L. Statistical mechanics of complex networks [J]. Rev. Mod. Phys., 2002, 74: 47-97

[9] Hemant B, Narsingh D. Discovering communities in complex networks[C] // ACMSE 2006: Proceedings of the 44th ACM Southeast Conference. Florida, 2006: 280-285

[10] 李孔文，顾庆，张尧，等. 一种基于聚集系数的局部社团划分算法[J]. 计算机科学，2010，37(7)：46-53

[11] 刘美玲. 基于复杂网络的社团发现算法研究[D]. 济南：山东师范大学，2010

[12] 汪小帆，刘亚冰. 复杂网络中的社团结构算法综述[J]. 电子科技大学学报，2009，38(5)

[13] 杨博，刘大有，Liu Ji-ming 等. 复杂网络聚类方法[J]. 软件学报，2009，20(1)：54-66

[14] 刘婷，胡宝清. 基于聚类分析的复杂网络中的社团探测[J]. 复杂系统与复杂性科学，2007，4(1)

[15] 蔡晓妍，戴冠中，杨黎斌. 基于谱聚类的复杂网络社团发现算法[J]. 计算机科学，2009，36(9)：49-50

[16] 李峻金，向阳，牛鹏，等. 一种新的复杂网络聚类算法[J]. 计算机应用研究，2010，27(6)

[17] 高倩. 基于模糊理论的谱聚类算法研究与应用[D]. 南京：江南大学，2009

[18] Newman MEJ. The structure and function of complex networks [J]. SIAM Review, 2003, 45: 167-256

[19] 刘靖明，韩丽川，侯立文. 基于粒子群的K均值聚类算法[J]. 系统工程理论与实践，2005，6：54-58

[20] 冯征，阎敏，张智峰. 一种基于PSO的模糊聚类算法[J]. 计算机工程与应用，2006，27：150-165

[21] 王林，戴冠中. 复杂网络的Scale-free性、Scale-free现象及其控制[M]. 北京：科学出版社，2009

[22] 赵凤霞，谢福鼎. 基于k-means聚类算法的复杂网络社团发现新方法[J]. 计算机应用研究，2009，26(6)


（上接第92页）


[6] Bohte S M, Gerding E H, La Poutré J A, et al. Competitive Market-Based Allocation of Consumer Attention Space[C] // EC'01 Proceedings of the 3rd ACM conference on Electronic Commerce. The ACM Press, 2001: 202-205

[7] Buyya R, Abramson D, Giddy J, et al. Economic models for resource management and scheduling in Grid computing[J]. Concurrency and Computation, 2002, 14(13/15)

[8] Buyya R, Stockinger H, Giddy J, et al. Economic Models for Management of Resources in Peer-to-Peer and Grid Computing [Z]. 2001. www.buyya.com/papers/economicmodels.pdf

[9] Feldman M, Chuang J C C H. Overcoming free-riding behavior in peer-to-peer systems[J]. SIGecom Exch, 2005, 5(4): 41-50

[10] Sun H, Huai J. Combining Reliability and Economic Incentives in Peer-to-Peer Grids[C] // ICN '09 Proceedings of the 2009 Eighth International Conference on Networks. 2009

[11] Marx L M, Sun P. Bidder Collusion at First-Price Auctions[M]. Springer, 2009

[12] Jansen B J, Mullen T. Sponsored search: an overview of the concept, history, and technology [J]. Int. J. Electronic Business, 2008: 6, 114-131

[13] Lai K, Rasmusson L, Adar E, et al. Tycoon: An implementation of a distributed, market-based resource allocation system[J]. Multiagent and Grid Systems, 2005, 1(3): 169-182

[14] 龚纯，王正林. 精通MATLAB最优化计算[M]. 北京：电子工业出版社，2009：401