

基于科研合作网络的自动审稿人选择研究

万 猛¹ 何良华²

¹(教育部科技发展中心 北京 100080)
²(同济大学计算机科学与技术系 上海 200092)
(helianghua@tongji.edu.cn)

Automatic Selection of Paper Reviewer Based on Scientific Collaboration Network

Wan Meng¹ and He Lianghua²

¹(Center for Science and Technology Development, Ministry of Education, Beijing 100080)
²(Department of Computer Science and Technology, Tongji University, Shanghai 200092)

Abstract In this paper, we study two tightly coupled topics in selecting paper reviewers from authors' scientific collaboration network (SCN): network construction and community detection. Based on the fact that the authors of one journal can be selected as reviewers and the reviewers of one manuscript should come from different research communities, we firstly evaluate the collaboration among all authors according to their signatures and construct the normalized collaboration network. For the second key problem of detecting the communities of one scientific collaboration network, considering it is much sparse and has few connections with inter community for one vertex, we apply the method of orthogonal matching pursuit to calculate compressive collaboration information. We conduct several experiments on simulated and real journal author datasets. Although there is no standard to evaluate different kinds of scientific collaboration network, the community detection accuracy rate and the stability of all authors are used to evaluate the performance of the proposed method. We can see from the vertex linkage matrix that our designed scientific collaboration network has good character of vertex grouping. The extensive study of our detection method in simulated data shows that the proposed method has a great advantage in the detection rate and stability. The significant improvement is about 60% compared with the classic methods.

Key words scientific collaboration network (SCN); sparse representation; orthogonal matching pursuit; paper review; reviewer selection

摘 要 主要研究科研合作网络 (scientific collaboration network, SCN) 中选择审稿人的 2 个核心问题: 网络构建和社区聚类. 基于杂志论文的审稿人主要来自本杂志作者及审稿人应尽量评审与自己无关的科研工作这 2 个事实, 通过论文中作者的排名计算所有作者之间的合作关系, 构建归一化科研合作网络. 考虑到网络中边的稀疏性, 设计了合作压缩感知算法来计算不同作者间的社区类型, 进行科研合作社区聚类. 在模拟数据及 2 个真实期刊作者库上开展了多个实验. 由于没有一个客观标准去评估所选出审稿人的合适性, 通过网络中顶点连接矩阵的自动聚类性评估所构建科研网络的性能, 通过作者合作团体的检测准确性来评价审稿人挑选的有效性. 从实验结果可以看出, 提出的网络构建方法具有较好的顶点聚类性; 和经典算法相比, 合作压缩感知社区检测算法在检测速度和稳定性方面具有很大的优势.

审稿人挑选正确率提高了大约 60%。

关键词 科研合作网络;稀疏表示;正交匹配追踪;论文评审;审稿人选择

中图法分类号 TP391

2 个世纪前,科学期刊就对所提交的稿件进行审查.现在,稿件的评审几乎作为科学论文发表的必经之路,绝大多数期刊中都采用.好的审稿结果不仅能在稿件重要性和学术性方面为期刊编辑部提供一个客观评估,还能对作者反馈同领域专家在稿件结构、逻辑描述等方面的建议,提高了论文和期刊的质量.然而,审稿人的选择非常关键,不同审稿人给出的审稿意见质量也不相同.至今大部分期刊都是定期选择新的审稿人,且这些人的能力只有在经过多次的审稿后才会展现.事实上,审稿人的质量是参差不齐的,涉及到审稿技巧、洞察力、爱岗敬业以及审稿人的公平意识.当然,并非所有的专家都擅长此任务.但是,谁是最好的审稿人?是否存在一种审稿人特征使我们可以预测他们审稿的质量?

事实上,一方面,很好的杂志如《Science》,《Nature》以及 IEEE Transactions,只有很少一部分提交的稿件可以被发表,审稿人的意见对于编辑是否决定发表至关重要;另一方面,对于那些著名的开放获取期刊如《PLoS One》,每年发表数万篇论文,所需审稿人员数量庞大.因此,完全没有差错的评审非常困难.为此,《Nature》曾经做过一期专刊并给出如下结论:“审稿人的选择是建设性和道德性评审的核心”.

《Nature》期刊审稿人的选择标准基于很多因素,包括:专业知识、声誉、建议和以前的审稿经验等.根本上,审稿过程的困难在于它引入了利益冲突或者道德风险.如大多数人都倾向于给他熟悉的人或熟悉的领域的论文好评,因此审稿人不允许评论那些和他们在同一所大学或者在最近 3 年有过合作的人的论文.然而,不可能对每一篇提交上来的稿件都这样做,毕竟有太多的作者和审稿人.事实上,大多数的期刊例如 IEEE Transactions,《Nature》,《Science》,都会从作者的数据库中选择审稿人.确实,能够在那些顶尖的期刊上发表论文的人,必须在自己的领域有很深入且高质量的研究.因此,选择他们作为审稿人是合理的.

随着信息技术的飞速发展,各大顶级期刊都有自己的网站,所有已经发表过的论文都可在其主页或者相应的数据库上下载,可以构建一个或者多个期刊的作者库.如果可以从期刊作者库中根据特定

标准客观自动选择审稿人,不仅可以减少编辑工作,而且还能提高评审质量.然而,此方面的研究非常少.Newman 是最早研究科研合作网络的学者之一,他提出了最原始的逻辑连接和论文贡献连接方法来构建科研合作网络(scientific collaboration network, SCN)并且作出详细的分析^[1-2].他提出了 2 种描述作者关系的方法:1)简单的逻辑值,即当 2 个作者有至少 1 篇论文,则合作关系为 1,否则为 0;2)加权值,将实际的论文数和作者的贡献整合到一起,例如如果 2 个作者一起写了 3 篇论文,这 3 篇论文又分别有 2,3,4 个合作者,那么这 2 个作者之间的边值为 $1+1/2+1/3=11/6$.近年来,一些很好的研究结果在著名期刊上报告出来,如《Proceedings of the National Academy Sciences》^[3],《Nature》^[4-5],其研究重点集中在不同因素的合作分析上,比如“每篇论文的合作人数”、“每个作者的论文数”、“每个作者的合作人数”等.

有几个在线工具可开展自动选择审稿工作^[6],每个人可以通过网站填写备选审稿人信息,网站系统通过稿件的摘要、关键词和备选审稿人的领域自动选择审稿人,由于它只考虑审稿人的研究领域,只要满足该条件,人人皆可审稿,没有审稿人和作者之间关系约束.实际上,从现有的文献看,基于客观公平和专业的约束自动地选择审稿人的工作还没有发现.本文就是试图探索此类研究,拟从已发表的论文中获取作者、共同作者、通信作者、作者单位信息,构建科研合作网络,研究科研社区的聚类方法,客观公正地选择合适审稿人,避免审稿过程中的非专业和主观倾向问题.

在科研合作网络中,如果将协作的概念进行延伸,就可以构建更全的合作研究社区,在这些合作研究社区中成员之间不仅密切协作,而且也来自于相同的研究领域.因此,如果在一个合作社区中只选择一个成员作为审稿人,一定程度上可避免大多数的非专业和主观倾向的发生.

相比于合作网络构建分析,社区检测研究结果则更加广泛.社区性在很多网络中都有,如生物网络、计算机网络、工程网络、经济网络、政治网络等,其研究已经历很长一段时间.现有的研究主要集中

在复杂网络的拓扑分析、功能分析、识别隐藏模式以及预测网络行为等. 2002 年, Girvan 和 Newman^[7]提出了一种奠基性的模块化算法, 通过重点判别社区之间的边来实现社区识别. 这篇论文的发表引来了很多的跟随者^[8-9], 近几年一系列新的改进方法及应用研究相继发表^[10-11], 如频谱中的修剪法^[12]、网络修正法等^[5]. Ford 和 Fulkerson^[13] 根据最大流最小分割原理, 提出了著名的自上而下的分离算法, 认为在任何 2 个顶点 s 和 t 之间的最小割容量就是从网络中 s 到 t 流量的最大值. 最近 Flake 等人^[8,14] 用最大流方法分析了万维网图中的社区问题. Hastie 等人在 2001 年提出了分等级的聚类算法来揭示图的多等级结构, 分等级的聚类算法在社交网络分析、生物学、工程学、市场营销中都都很常见. 时下很流行的算法由 Girvan 和 Newman^[7] 提出, 该算法试图通过不同的重要性度量来估计边的重要性等级, 进而进行选择、分析社区. 然而, 所有的模块化方法中, 没有一种算法能够证明所产生的模块是网络的最佳分区.

本文首先分析协作的双向性. 在简单逻辑科研合作网络中, 作者和论文是平等对待的. 在加权科研合作网络中, 论文的影响力被归一化, 作者的影响力也被忽略, 因此它们都是无向图. 然而, 人为因素也是描述协作的关键点, 例如如果一个教授有 5 个博士研究生, 在他的指导下每个博士研究生都发表了一篇论文. 对教授来说, 和每个学生的合作仅仅占了他整个合作工作的 20%; 对他的每个学生来说, 和教授的合作是其所有工作, 为 100%. 因此, 合作对象不同, 合作意义不同.

其次, 不同于其他的社交网络, 科研合作网络是连接稀疏并且巨大的. 一般情况下, 一个作者只能和十几个作者有联系, 一个期刊的总作者可以高达成千上万. 因此, 和利用顶点关系列表算法相比, 利用科研合作网络中的边来检测和构建社区将更加稳定和准确. 需要注意的是, 本文中作者之间的合作边不仅是构建科研合作网络的基础, 也是后续检测分析社区的基础. 因此, 本文主要创新如下:

1) 科研合作网络构建的改进. 通过将合作进行方向化, 构建双向科研合作网络; 通过每个作者的归一化计算, 使得网络中每一对连接具有相同合作内涵. 对比于经典的科研合作网络, 本文的改进使得网络中的每条边更具有可比性.

2) 将压缩感知这一线性重构方法引入合作网络的作者社区检测研究, 在大科研合作网络的社区检测计算中, 降低了计算复杂度.

1 双向科研合作网络建模

为了更准确地描述单个研究人员的合作关系, 本文提出 3 个假设定义:

1) 不同作者之间的协作程度是不同的. 本文将第一作者与通信作者之间的合作度定义为 3, 将第一作者或通信作者与其他作者的合作度定义为 2, 非第一或通信作者的作者之间合作度定义为 1. 合作度定义为论文的署名权 W_s .

2) 作者间合作论文数量不同, 合作关系也不同. 本文采用经典思路, 使用累加来描述合作关系. 本文称此项价值为论文的价值权 P .

3) 每个作者的总合作值为 1.

科研合作网络是一个有向图 $G=(V, E)$, 其中 V 是顶点集合, 代表各个作者; E 是网络的边, 代表各个作者的合作关系. 假设作者 i 的合作者共有 N 位, 其中和作者 j 合作了 P 篇论文, 在所有的 P 篇论文中, 作者 i 署名权分别为 $W_{ij}=(w(1), w(2), \dots, w(P))$, 则作者 i 与作者 j 的总论文署名权为

$$W_{ij} = \sum_{k=1}^P w(k), \text{ 其中: } w(k) = \begin{cases} 3, & \text{如果 } i \text{ 为第一作者, } j \text{ 为通信作者;} \\ 2, & \text{如果 } i \text{ 为第一作者或通信作者,} \\ & j \text{ 为其它作者;} \\ 1, & \text{如果 } i, j \text{ 都为普通作者.} \end{cases}$$

3 种科研合作网络的构建方法如下:

1) 逻辑科研合作网络. 其边值为逻辑数值 0 或者 1, 则 $G_L=(V, E)$, 其中, E 中元素 e_L 为

$$e_L = \begin{cases} 1, & \text{if } W_{ij} > 0; \\ 0, & \text{otherwise.} \end{cases}$$

2) 权重科研合作网络. 其边值引入权重变量, 则 $G_W=(V, E)$, 其中, E 中元素 e_W 为 $e_W=W_{ij}$.

3) 双向科研合作网络. 其边值引入合作方向变量, 则 $G_B=(V, E)$, 其中, E 中元素 e_B 为 $e_B(i, j) = W_{ij} / \sum_{k=1}^N W_{ik}$. 因为 $W_{ij} \neq W_{ji}$, 所以 $e_B(i, j) \neq e_B(j, i)$.

从上述网络构建方法看, 第 1、第 2 种网络可以由如下公式统一给出:

$E=W \otimes E_T$, 其中 W 为 i 与 j 的论文署名权重; $e_{ij} \in E_T$ 且

$$e_{ij} = \begin{cases} 0, & \text{如果 } i \text{ 与 } j \text{ 合作过至少 1 篇论文;} \\ 1, & i \text{ 与 } j \text{ 没有合作任何论文.} \end{cases}$$

如果不考虑合作论文的多少及署名位置, 则 W

为 1 个矩阵. 因此, 第 1、第 2 种网络为同类网络; 第 3 种网络因为考虑到合作的方向因素, 是双向网络.

2 合作压缩感知社区聚类算法

网络构建后, 对网络中不同作者的聚类是本文第 2 项研究内容. 本文的聚类目标是合作关系, 传统处理思路是社区聚类与分块. 网络模块的定义取决于特定的系统和应用领域, 广泛被接受的定义是社区中的节点, 其到社区内的边比到社区外的边要多. 有的定义虽然有所不同, 但相互兼容; 也有的定义有着严格的数学公式. 总之, 网络顶点间的连接是绝大多数合作网络的社区定义基础, 但并不是唯一标准, 也许 2 个顶点间有连接但并不属于同一个社区.

本文中, 作者社区是评审人员选取的基础, 根据选取中的利益冲突避免规则, 作者与审稿人越不熟悉越好, 网络的边是本文的唯一评价标准. 为了设计本文的社区检测算法, 提出如下 5 个假设:

- 1) 合作社区定义非常广泛, 只要有边, 即可看作是一个社区;
- 2) 2 个作者如果和同一个作者有过合作, 那么这 2 个作者属于同一个社区;
- 3) 只要社区内有一人与作者有关系, 则整个社区都不能审稿;
- 4) 网络中 2 个没有路径到达的作者可以互相审稿;
- 5) 孤立的作者不能审稿.

假设共有 M 个作者, 根据科研合作网络的定义, 其边矩阵 E 第 i 个向量 $e_i = \{e_{i,1}, e_{i,2}, \dots, e_{i,M}\}$ 表示第 i 个作者与其他作者的合作关系. 如果第 j 个作者与其属于同一个社区, 则根据上述假设可知, $e_j = \{e_{j1}, e_{j2}, \dots, e_{jM}\}$ 和 e_i 至少有一个相同元素. 经典的谱方法就是根据此相同元素进行计算, 然而, 它对相同合作者的数量、连接矩阵 E 中相同作者的顺序等因素比较敏感.

如上一段所述, 有 M 个作者的合作网络边矩阵为 $E = \{e_1, e_2, \dots, e_M\}$, e_i 是一个长度为 M 的向量, 如果作者 i 与作者 j 具有相同的合作者 k , 那么 e_{ik} 与 e_{jk} 将同时不为 0, 则它们的投影 $p_{ij} = \langle e_i, e_j \rangle$ 必将大于 0; 同理, 如果作者 i 的合作者集与作者 j, k 同时有交集, 则可先计算投影 $p_{ij} = \langle e_i, e_j \rangle$, 然后计算向量 e_k 与 $e_i - p_{ij}e_i$ 的投影, 以此类推. 上述的计算思想与信号处理中著名的稀疏重构理论非常相似. 因此, 本文引入正交匹配追踪算法^[15]快速计算每个

顶点合作边向量和和其他顶点合作边向量的重构系数, 批量计算合作网络中与此作者的可能合作社区对象.

在压缩感知 (compressive sensing, CS) 理论中, 如果知道信号 x 的稀疏度 k , 则可根据最优 l_0 范数问题的求解思路从测量信号 y 中重构原始信号 x :

$$\begin{aligned} \hat{x} &= \min \|x\|_0, \\ \text{s. t. } y &= \Phi x. \end{aligned} \tag{1}$$

其中, $\| \cdot \|_0$ 为向量的 0 范数, 记为向量中非零元素的个数. 如果矩阵 Φ 满足约束等距性, 且稀疏度 k 、测量信号维数 M 、原始信号维数 N 满足 $M \geq k \lg N$, 则 x 可被重构.

因此, 本文引入压缩感知算法, 提出合作压缩感知社区检测算法 (collaboration compressive sensor, CCS). 其中, 观察信号是每个作者的合作向量 $e_i, i=1, 2, \dots, M$; 原始信号是所有可能的其他作者合作向量 $e_j, j=1, 2, \dots, i-1, i+1, \dots, M$. 由于无法确认每个作者的稀疏度 k_i , 因此本文采用控制重构误差来约束重构向量个数, 即如果重构信号与观察信号的误差在限定阈值 ϵ 内: $\|e_i - \Phi \hat{x}\| \leq \epsilon$, 则停止分解.

合作压缩感知社区检测算法伪代码如下:
算法 1. 合作压缩感知社区检测算法伪代码.
输入:

- 1) 网络顶点数 N , 网络边 $E = (e_i)_{i=1}^N \in \mathbb{R}^{N \times N}$, 社区 $C = (C_i)_{i=1}^F$ 且 $C_i = (e_i)_{i=1}^M \in \mathbb{R}^{N \times N}$;
 - 2) 错误阈值 ϵ .
- 计算:
- 1) 计算矩阵 A
 - ① 初始化: $F=0, C=\emptyset, i=0$;
 - ② 在边权矩阵 E 中搜索 C_i , 其中 $e_i \in C_i$ 或 $F=F+1, C_i=C_F=\emptyset$;
 - ③ 计算 $A=E \cap C_i$, 重复上述工作;
 - 2) 同一社区成员计算
 - ① 初始化: $x^0=0, r^0=e_i-Ax^0, S^0=\sup x^0=\emptyset, k=0$;
 - ② $k=k+1$;
 - ③ 选择 j_0 使得 $\min_c \|ce_{j_0}-r^{k-1}\|_2 \leq \min_c \|ce_j-r^{k-1}\|_2$;
 - ④ 计算 $S^k=S^{k-1} \cup \{j_0\}$;
 - ⑤ 计算 $x^k=\operatorname{argmin}_x \|Ax-e_i\|_2, x=S^k$;
 - ⑥ 计算 $r^k=e_i-Ax^k$;
 - ⑦ 如果 $\|r^k\|_2 \leq \epsilon$, 重复计算;

⑧ 更新 $C_i = C_i \cup \{a_p\}$, 如果 $x_p^k > 0$;

⑨ 更新 $C = (C_1, C_2, \dots, C_F)$.

输出:

每个作者 v 的可能同一社区成员集合.

本文提出的合作压缩感知检测方法为线性分析, 对一个网络节点为 N 的网络来说, 其计算复杂度为 $O(\lambda(N-1)^2)$, λ 为迭代计算次数. 从文献[16]中可知, 特征值计算类算法的复杂度为 $O(N^3)$, 常用的 Kernighan-Lin 算法的复杂度为 $O(N^2)$. 可以看出, 本文的算法复杂度具有一定的优势, 且 N 越大优势越明显; 另一方面, 由于算法原理是统计计算, 所以初始状态的不同不会对该方法有影响, 后面相关实验也证明了本算法的鲁棒性.

3 实验及结果

3.1 数据介绍

本文所用数据集来自于 2 个部分: 模拟数据和真实数据.

模拟数据用来评价不同网络构建算法所构建的逻辑合作网络的自动聚类性能, 采用文献[16]中的复杂网络模型, 主要有 5 个参数: $G = (N_v, N_e, D_v, P_e, P_o)$. 其中, N_v 是社区数量; N_e 是社区大小; D_v 是顶点度数, 是网络中每个顶点与其他顶点连接数目的度量; P_e 是区内连接概率, 是区内一个顶点有 D_v 连接的概率; P_o 是区间连接概率. 在所有的参数中, 如果 D_v 和 P_e 越大, 则检测出社区的概率就越大.

真实数据来自如下 2 个数据库: 1) 著名 ArXiv 数据库^①, 时间是从 1993-01—2003-04, 共 124 月; 2) 中国科技论文在线^②, 该网站由中国教育主办、教育部科技发展中心承办, 共有 39 个专业, 发布的论文数超过 100 万. 本文使用的数据来自数学、计算机、管理、经济、电子学等 5 大学科, 包括作者姓名、第一作者、通信作者、作者单位等.

3.2 合作网络构建实验

第 1 个实验研究不同科研合作网络构建方法. 为了清楚地描述构建过程, 首先利用 8 篇论文的信息进行展示. 假设这 8 篇论文中共有作者 Q, R, S, T, U 五个作者, 每篇论文的作者信息如表 1 所示:

表 1 作者 Q, R, S, T, U 的论文合作情况
Table 1 Paper Collaboration Among Author Q, R, S, T and U

Collaboration Author	Paper List	Author Number	First Author	Corresponding Author
Q and R	1	4	Q	
	2	2	Q	R
	3	3		
Q and S	4	3	Q	
	5	4	S	Q
Q and T	6	2	T	
R and U	7	4	U	R
	8	2	R	

根据表 1 中显示的数据构建的 4 种合作网络如图 1 所示:

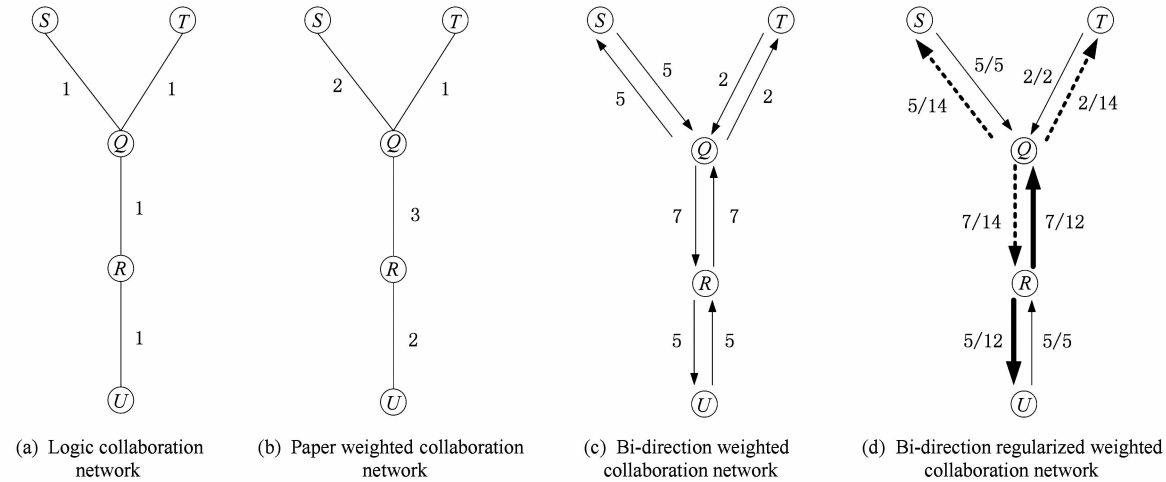


Fig. 1 Four kinds of graphs of table 1.

图 1 表 1 数据的 4 种科研合作网络图示

① <http://snap.stanford.edu/data/>

② <http://www.paper.edu.cn>

1) 从图 1 中我们可以看到,图 1(a)和图 1(b)显示的科研合作网络是无向图,两者的区别是论文贡献权重不同;图 1(c)及图 1(d)显示的网络是有向图,且对图 1(d)的网络论文和人的因素进行了归一化.

2) 图 2 的 4 幅图是 ArXiv 数据中连接矩阵前 100 个元素构成的 4 种科研合作网络矩阵图示,其中,横、纵坐标是作者序号,竖向坐标为合作数值,0 为没有合作,1 为有合作. 因为顶点(作者)矩阵是按照论文的顺序排列,相邻顶点构成社区的可能性最

大,主要体现为图 2 对角线上数值常为 1. 图 2(a)(b)显示了简单逻辑的科研合作网络和权重科研合作网络;图 2(c)(d)显示了第 2 类加权合作网络和归一化合作网络. 从图 2 可以看出,简单逻辑的合作网络和加权网络相比更加无序,加上论文权重和作者签名权重后,网络构成有了较大的变化. 但是,最为明显的是图 2(d)的归一化网络,其对角数值有效地表达为 1,社区的特征被最显著地显示出来,完全符合真实情况.

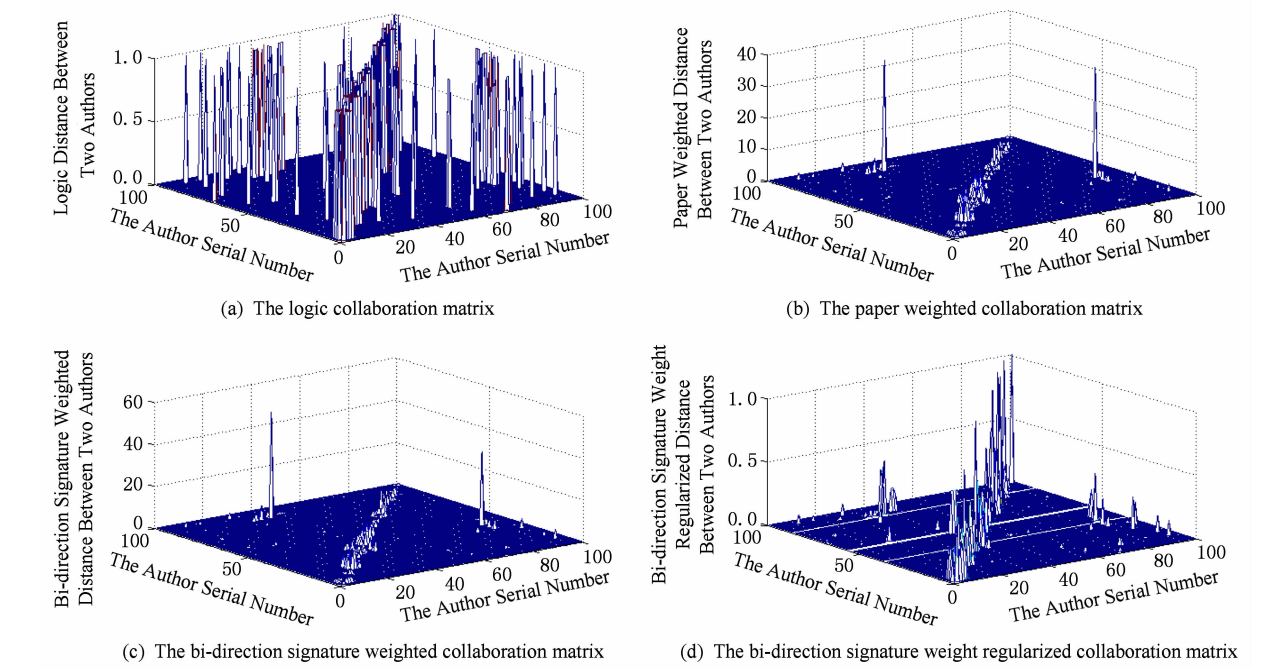


Fig. 2 Four kinds of graphs of 100 authors from ArXiv database.
图 2 基于 ArXiv 数据库中前 100 个作者的 4 种科研合作网络展示

3) 本文计算了 6 个学科中最短路径这一网络参数,结果如图 3 所示. 其中,横坐标为学科,纵坐标

为最短路径参数数值. 从图 3 可以看出,最短路径随着不同的学科和不同的方法而不同,尤其是 HEP-TH 数据集;然而,和其他 3 种科研合作网络相比,本文设计的归一化合作网络的距离是最稳定的.

3.3 基于模拟数据的社区聚类检测实验

根据 Yang 等人^[16]的研究结果,和其他算法相比,FEC(finding and extracting communities)的方法有着更加全面的性能. 因此,为了评价提出方法的社区检测能力,FEC 方法用来作参照,共开展了 2 组实验:1)测试检测精确度;2)评价敏感性和稳定性.

需要说明 2 点:1)在 2 组实验中,参数 N_c 及 P_c 设为常数,分别为 80 和 0,因为它们相比于其它 3 个参数影响更小. 2)如果检测结果 C'_i 和真实的 C_i 中有最多的共同顶点数 N'_v ,则认为检测结果为 C_i , N'_v 为有效检测顶点. 检测精度计算公式如下:

$$R = N'_v/N, \tag{2}$$

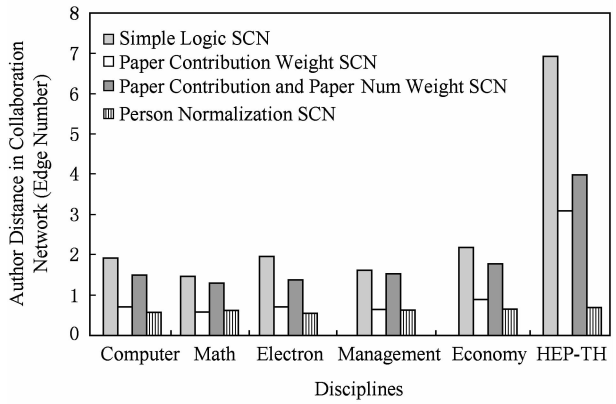


Fig. 3 The shortest distance of four kinds networks of six disciplines.
图 3 6 个学科中 4 种网络的最短距离

其中, $N' = \sum_{i=1}^{N_c} N_v^i, N = \sum_{i=1}^{N_c} N_v$.

图 4 显示了一些参数下模拟数据网络图示. 图 5 显示了相应的结果, 其中横坐标是 P_e 数值, 纵坐标是根据式(2)计算的检测精度. 为了简洁地显示结果, 本文将 N_v 和 D_v 的比值作为一个变量 $P_v = D_v/N_v$.

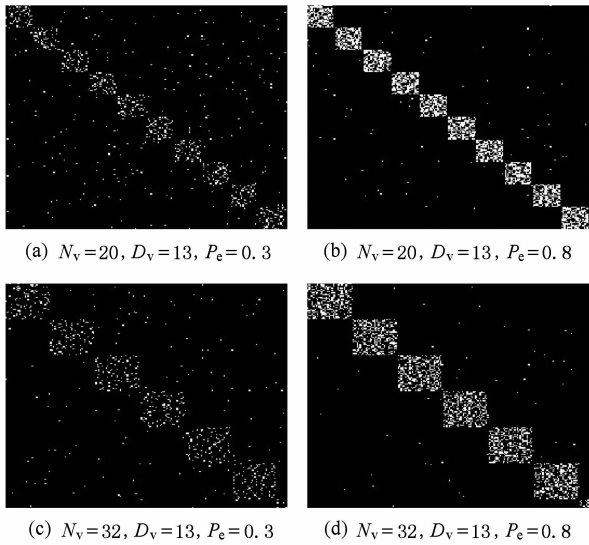


Fig. 4 Some simulation results.

图 4 部分模拟数据图

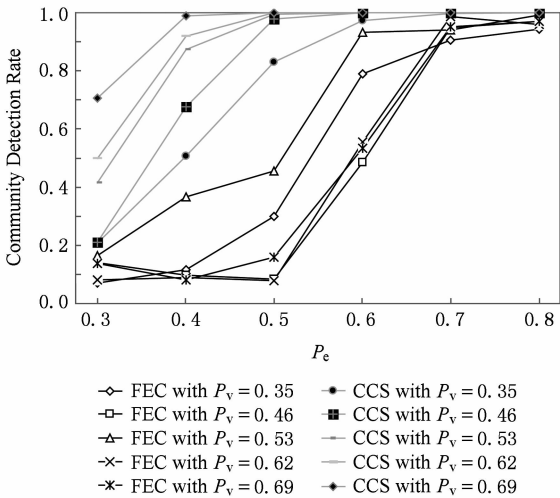


Fig. 5 Community detection rate of FEC and CCS method under different parameters.

图 5 2 种算法在不同参数下的社区检测结果

从图 5 可以看出, 检测精度随着 P_v 和 P_e 的增大而增加. 当 $P_e < 0.6$ 时, FEC 算法的检测精度低于 50%, 而本文提出的算法检测精度显然更高; 当 $P_e = 0.5$ 时, 本文提出的算法均超过 80%, 远远高于 FEC 算法的最好结果 40%; 当 $P_e < 0.8$ 时, 本文

提出的算法完胜 FEC; 当 $P_e > 0.8$ 时, 2 种算法的差别不是很明显, 检测精度均高于 90%.

在真实的科研合作网络中, 一般而言每个作者的合作人数大概在 20~30, 即社区内节点数大约为 20 个或 30 个, 这比社区数目一半还要小, 即 $P_v < 0.5$. 另外, 从文献[1]中聚类系数等参数可以估计出 $P_e < 0.5$, 因此可以得出这样的结论: 和大多数传统算法相比, 本文提出的算法更适合真实低概率连接的科研合作网络中的社区检测.

为了测试社区检测的稳定性和鲁棒性, 本文打乱连接矩阵的顺序. 事实上, 这样的操作对很多检测方法影响较大, 比如神经网络、决策树等. 一般减少此类影响的做法是重新安排数据之后作分析. 本文则不作任何处理, 直接检测.

图 6(a)显示的是生成的原始模拟网络; 图 6(b)显示的是打乱顺序之后的网络. 从图 6(b)可以看出, 此时的连接矩阵更随机、更无序.

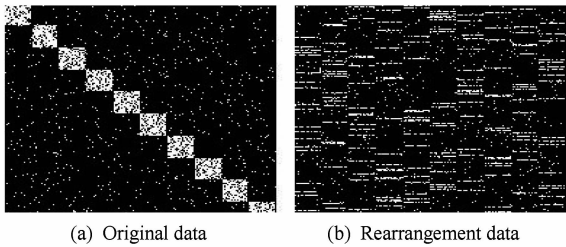


Fig. 6 Vertex link graph.

图 6 顶点连接矩阵图

图 7 显示了 2 种算法在乱序情况下的检测结果, 其中横坐标是 P_e 数值, 纵坐标是根据式(2)计算

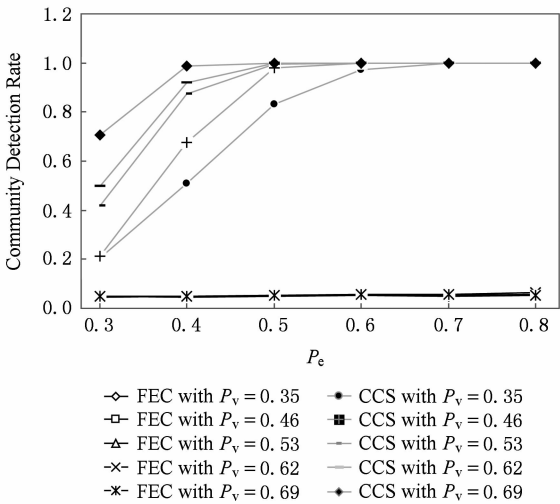


Fig. 7 The performance of two methods on random arrangement of original data.

图 7 2 种算法在随机排列矩阵上的性能

的检测精度. 从图 7 可以看出, FEC 算法对于数据的重排非常敏感, 并且它在不同情况下的精度大约只有 10%, 非常不好; 然而, 本文提出的算法效果要好很多, 和打乱之前相比没有太多的变化. 图 8 所示的是 2 种算法在随机排列矩阵上的检测效果, 可以看出本文提出的 CCS 算法成功地进行了区域检测, 这表明 CCS 算法相比于 FEC 算法性能更好.

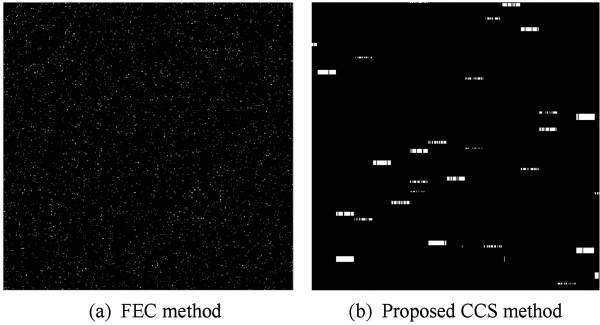


Fig. 8 Two methods on random arrangement of original data.

图 8 在随机排列矩阵上 2 种算法的检测效果

3.4 基于真实数据的评审搜索实验

最后一组实验是测试本文提出算法在真实作者库上的检测性能.

为了测试搜索评审人的结果, 本文分析了中国科技论文在线网站中 5 个学科的社区内作者间距离 D_e 和社区间作者间距离 D_o . 首先, 对每个单位设置一个数值, 如北京大学为 1, 清华大学为 2 等. 因为每个作者有一个单位, 则对第 i 个社区有一个单位向量 $\mathbf{V}^i = (v_1^i, v_2^i, \dots, v_{N_v^i}^i)$, 其中 N_v^i 是第 i 个社区的顶点个数. 因此, 第 i 个社区内距离 D_e 以及第 i 和第 j 社区之间的距离 D_o 的计算公式如下:

$$D_e(i, i) = \frac{1}{N_v^i} \sum_{j=1}^{N_v^i} (v_j^i - m^i)^T (v_j^i - m^i). \quad (3)$$

$$D_o(i, j) = \frac{1}{N_v^i N_v^j} \sum_{j=1}^{N_v^i} \sum_{l=1}^{N_v^j} (v_k^i - v_l^j)^T (v_k^i - v_l^j). \quad (4)$$

其中 m^i 是 v^i 的均值. 本文仅计算最大的 100 个社区中的距离, 结果在图 9 中显示, 其中横坐标代表的是学科, 纵坐标是根据式(3)、式(4)计算的距离.

从图 9 可以看出, 在所有学科上, 区内作者间距离 D_e 都小于区间作者间距 D_o , 这表明同一社区作者的单位集中度比社区间作者的单位集中度要高, 符合单位内合作比单位间合作要频繁这一常识, 这说明采用本文提出的算法算出的社区和现实比较相符.

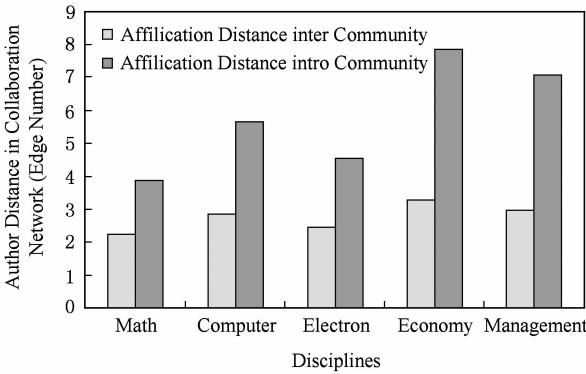


Fig. 9 Two kinds of distance between authors of five disciplines from database of www.paper.edu.cn.

图 9 中国科技论文在线数据库上 5 大学科中的 2 种作者距离图示

4 结 论

本文针对公平选择审稿人问题, 对传统的科研合作网络进行改进, 引入合作的双向性构建科研合作网络. 在此基础上, 提出了合作网络社区的压缩感知检测法, 对科研合作网络中的作者快速稳定地进行社区分类, 由于不同社区作者之间没有合作, 因此来自不同社区的审稿人能提供更加客观的审稿意见. 仿真数据的实验表明, 本文的改进网络具有更好的社区聚类性; 本文的压缩传感社区检测算法具有更加准确和稳定的检测效果, 同时在真实作者库的结果也符合真实情况. 更重要的是, 就作者所知, 本文的通过社交网络选择审稿人研究尚属首次, 而论文评审是科研成果评价的核心, 因此本文工作的社会意义较大.

参 考 文 献

[1] Newman M E. Scientific collaboration networks. I. Network construction and fundamental results [J]. Physical Review E, 2001, 64(1): (06131) 1-8

[2] Newman M E. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality [J]. Physical Review E, 2001, 64(1): (06132) 1-7

[3] Mason W, Watts D J. Collaborative learning in networks [J]. Proceedings of the National Academy Sciences, 2012, 109(3): 764-769

[4] Watts D J, Strogatz S J. Collective dynamics of ‘small-world’ networks [J]. Nature, 1998, 393(6684): 440-442

[5] Guimera R, Nunes Amaral L A. Functional cartography of complex metabolic networks [J]. Nature, 2005, 433(7028): 895-900

- [6] Andrade-Navarro M A, Palidwor G A, Perez-Iratxeta C. Peer2ref: A peer reviewer finding Web tool that uses author disambiguation [J]. *BioData Mining*, 2012, 5(14): 1-8
- [7] Girvan M, Newman M E J. Community structure in social and biological networks [J]. *Proceedings of the National Academy Sciences*, 2002, 99(12): 6-15
- [8] Flake S L G W, Giles C L, Coetzee F M. Self-organization and identification of Web communities [J]. *IEEE Computer*, 2002, 35(3): 6-14
- [9] Newman M E. Fast algorithm for detecting community structure in networks [J]. *Physical Review E*, 2004, 69(2): 66-133
- [10] Guimera R, Sales-Pardo M, Amaral L A. Modularity from fluctuations in random graphs and complex networks [J]. *Physical Review E*, 2004, 70(2): 7821-7826
- [11] Danon L, Duch J, Guilera A, et al. Comparing community structure identification [J]. *Journal of Statistical Mechanics*, 2005, 9(9): 1-10
- [12] Tao W, Jin H, Zhang Y. Color image segmentation based on mean shift and normalized cuts [J]. *IEEE Trans System Man Cybernetics—Part B*, 2007, 37(5): 1382-1389
- [13] Ford L R, Fulkerson D R. Maximal flow through a network [J]. *Canadian Journal of Mathematics*, 1956, 8: 399-404
- [14] Flake G W, Lee Giles C. Efficient identification of Web communities [C] //Proc of the 6th Int Conf on Knowledge

Discovery and Data Mining. New York: ACM, 2000: 150-160

- [15] Tropp J A, Gilbert A C. Signal recovery from random measurements via orthogonal matching pursuit [J]. *IEEE Trans on Information Theory*, 2007, 53(12): 11-21
- [16] Yang B L, Da You, Liu Jiming. *Discovering Communities from Social Networks: Methodologies and Applications* [G] //Handbook of Social Network Technologies and Applications. Berlin: Springer, 2010: 1-16



Wan Meng, born in 1975. PhD and associate professor at Science and Technology Development Center, Ministry of Education. Member of IEEE and ACM. His main research interests include computer network architecture, network and systems management, science and technology management, system engineering etc.



He Lianghua, born in 1977. PhD and associate professor at Tongji University. His main research interests include signal processing, pattern analysis, and machine learning.