数字人文社会网络分析方法应用与研究*

施晓华　王　昕（上海交通大学图书馆）

摘　要　社会网络分析方法逐渐成为数字人文研究的方法与工具之一，本文以上海交通大学古徽州契约文书数字化资源为研究数据，以MATLAB为数据分析工具，进行了社会网络数据分析与可视化实验。通过数据处理，有效获取契约交易社会网络的关键网络信息，以契约交易者数进行“小世界”幂律分布拟合，并实现有效社区结构划分，揭示数字化资源数字人文分析中的社会网络特性和结构信息。

关键词　社会网络分析　数字人文　特藏资源　网络结构

DOI: 10.13663/j.cnki.lj.2020.05.012

Research and Application of Digital Humanities with Social Network Analysis

Shi Xiaohua, Wang Xin (Library of Shanghai Jiao Tong University)

Abstract　Social network analysis is one of the methods of digital humanistic analysis. In this paper, we take the digital resources of contract documents of Shanghai Jiao Tong University as research data, and MATLAB as research tool, to carry out social network analysis and visualization. Through data processing, we can effectively acquire network nodes and network degree data, fit the power-law distribution based on the number of participants, detect and categorize legit community, and reveal the characteristics and structure of digital humanities and social network in digital resources.

Keywords　Social network analysis, Digital humanities, Special collection resources, Network structure

0　引言

随着时代的变化，各个高校和研究机构都在不断将其特色馆藏进行数字化，由此产生更多数字资源或数据资源，进行展示、利用和研究，特藏资源的概念和内涵也不断地被泛化和扩充。曾蕾等[1]认为：“图书馆、档案馆和博物馆所拥有的数据资源是数据时代各个领域，尤其是数字人文领域的无价之宝。”无论是图书馆馆藏的古籍、民国文献、地方文献、地方戏曲、文史资料等各类资源、档案馆各机构所收藏的档案及私人文书、博物馆的各种馆藏实物，这些资源的数字特藏不仅有利于特色资源的长期保存，加快人文知识的大众普及化，还能为人文学者的研究提供开放获取和研究数据支持。

数字人文（Digital Humanities）[2-3]是一门通过计算机等技术与社会人文研究相结合的一门新学科，它与人文计算、社会计算、媒体研究等相关。具体来说，数字人文学基于原生数字特藏资源，以传统人文学（如历史学、哲学、语言学、文学、艺术、考古学、音乐学以及文化研究）等社会科学为研究目的，以计算机信息技术（如超文本、超媒体、资料可视化、信息检索、统计学、文本检索和数字地图）为研究解决方法，具有综合性、交叉性的新兴研究学科。其关键是研究素

* 本文系2018年上海市社科规划一般课题“数字人文社会网络分析应用与研究”（项目编号：2018BTQ002）的研究成果之一。

材、研究目的和研究方法的有效结合，实现结合图书馆学、社会科学和计算机学科的融合应用，将数字化工具与方法带进社会人文学中。

在计算机学科中的社会网络分析（Social Network Analysis）主要是依靠复杂网络（Complex Network）学科的分析和可视化技术对社会网络节点之间的各类关系数据进行定量或定性分析[4]。其特点在于采用网络视角的独特方法，形成专门的采集数据、量化分析和可视化表示等技术体系。虽然主流社会科学关注的是一元属性（例如，人员的收入、年龄和性别等），但社会网络分析关注的是成对个体的属性，其中二元关系是主要类型，常见的一些二元属性有亲属关系、社会角色、情感、认知、行动、流量、距离和共现等。

在数字人文领域进行社会网络分析可以为人文科学中各种社会关系提供精确的量化分析，从而为构建理论模型和验证命题提供社会化例证，挖掘出更多隐含的社会关系和变化趋势。利用社会网络分析技术对特色资源进行分析和建模，关注的焦点是节点的获取与处理，关系和关系的模式确定，采用的方式和方法从概念上有别于传统的统计分析和数据处理方法，是数字人文技术在数字化资源应用和挖掘上的一种新的尝试。本文在介绍社会网络分析方法，及其在数字人文中的应用现状基础上，以上海交通大学图书馆数字化的古徽州契约文书特藏数据为例进行实践研究，经过网络化处理后，对交易社会网络数据进行了相关网络指标量化计算与可视化分析。

1 社会网络分析方法简介

社会网络分析主要研究社会实体间的关系，这些实体可以是人、组织、国家或机构，或者人类活动或认知的产物，如网站、语义概念等。它与社会学中的结构主义有关，强调社会参与者之间关系的重要性，以及他们的行为、观点和态度，网络中的关系模式所反映出的现象或数据是网络分析的焦点。社会网络分析被认为适合分析社会凝聚力、交流和演化趋势，以及社会群体内部或社会群体的社会地位的重要应用方法。

在计算机网络科学中，已经有较多专用分析软件和工具，Ucinet、Pajek、Cytoscape 和 MATLAB 等软件都是可以专门处理社会网络分析、进行可视化的工具。一般网络按照形态可以分为随机网络、规则网络和小世界网络[5]等，最值得社会学研究是社会网络存在高度的小世界特性[6]、无标度（幂律分布）特性[7]和相互协作的网络社区结构，这些特性都对应着不同的网络参数和结构：

（1）小世界（small world）特性

在人际关系网络中，人和人的距离到底有多“远”，历来是社会科学所关注的问题。匈牙利 Karinthy 于 1929 年提出，世界上随机选择的两个人都可以通过 6 个熟人而互相连接起来的著名“小世界理论”构想，并通过实验进行了证实，因此这些网络也被称为小世界网络，我们一般会使用特征路径长度和网络平均聚合系数两种特征来衡量小世界网络。

任选一个联通网络的两个节点，连接这两个节点对应的最少边数，被定义为两个节点的路径长度。网络中所有节点对的路径长度平均值，定义为网络的特征路径长度，这是一个网络的全局特征。如公式 1，n 为网络节点数，$d(i,j)$ 为节点 i 与节点 j 的最短距离：

$$L=\frac{1}{n*(n-1)}\sum_{i\neq j}d(i,j) \quad \cdots\cdots \quad (1)$$

我们将网络中实际存在的边数除以最多可能存在的边数得到的数值，定义为这个节点的聚合系数，而所有节点聚合系数的平均值定义为网络平均聚合系数。聚合系数是网络的局部特征，在社会网络中反映了相邻两个人之间关系网络的重合度，即某节点人员的联系人之间也相互联系的程度。如公式 2，$|E_i|$、$|V_i|$ 分别为第 i 个节点连接的节点数和边数：

$$CC(G)=\frac{1}{n}\sum_{i}\frac{2*|E_i|}{|V_i|*(|V_i|-1)} \quad \cdots\cdots \quad (2)$$

如果一个网络的平均集聚系数明显高于相同结点集生成的随机图，而且平均最短距离与相应随机生成的随机图接近，一般认为这个网络是“小世界”的。

（2）幂律（Power law）分布

在现实世界的很多社会网络中，网络中少量节点却往往会拥有大量的连接，而其余大部分节点的连接却很少，我们认为，这些节点符合网络的无标度特性（Scale-free），其度数分布一般都符合幂率分布（见公式3），这些网络常被称为无标度网络：

$$P(z) \sim z^{-\tau} \qquad \cdots\cdots \quad (3)$$

其中 τ 是常数。

社会网络的无标度性是描述大量复杂实际网络系统整体上严重不均匀分布的一种内在异质特性。

（3）社区化（Community）

社会网络中的节点往往也呈现出社区集群化特点，社会网络中总是存在每个成员都相互认识、相互熟悉的朋友圈或关系圈。社区化的意义是网络聚集化的程度，这是一种网络的内聚倾向。通过社区发现算法，分解获取的各个社区反映了一个大网络中各集聚的小网络分布和相互联系[8]，它是用来揭示网络聚集行为的一种技术。在大型社会网络上进行社区发现的操作过程，同时就是按照某些标准进行了分类划分，由此可以进一步分析探索每个社区内在特性；而从数据分析与计算复杂度来看，社区发现等同于对整个大型网络进行分类任务，在降低计算复杂性方面起到了一定的作用。

2　数字人文中社会网络分析应用与发展

数字人文资料数据中存在了大量结构化网络信息，由数字化资源内容中的不同元素之间的关系创建，其中人与人之间的关系即成为数字人文领域的社会网络[9]。斯坦福大学图书馆公布了一个Kindred Britain[10]数字人文网络，由近3万名个人组成，其中多数是英国文化中的偶像人物，他们之间通过血缘、婚姻或亲属关系相互连接起来；用户可以基于此网络，分析英国社会中家庭关系、商业和政治环境及其相互之间的联系。

数字人文社会网络分析不同于一般的微博、微信或网站服务的社会网络分析应用，其社会网络数据主要都是从已有资源数据中处理获取，需要通过网络化分析与加工处理，其中网络节点与节点的关系属性会根据不同人文社科需求而异。在数字人文社会网络中，参与人员成为网络中的每一个节点，人员 u 和 v 共同参与的某一社会活动成为之间的链接（如共同参加讨论、交易或文件签署等）。对应的链接权重 $w(u, v)$ 等于 u 和 v 共同参与的事件数量。在社会网络中，还可以包括节点 u 和 v 之间的关联方向（如 u 向 v 购买货物）和全局权重（u 和 v 之间的交易次数或交易额等）。

目前国内外一些人文学研究者，已经开始通过数字人文社会网络方法进行了一些实践研究。胡静[11]以朝鲜时期科举考试榜目数字档案为研究资料，建立技术中人的社会网络，以探索技术中人阶级在朝鲜初期和中期的阶级发展过程。通过分析社会网络中代际考试科目的变化情况，有效相关流动的变化过程。严承希等[12]引入基本的网络分析指标对宋代政治社会真实网络进行整体性网络结构分析，用k-core分析对全宋政治网络进行分解和可视化应用，力图进一步揭示和解释这种逐步强化的相权政治的网络特征和空间属性。

Newman[13]对科学合作网络的结构进行了研究，并揭示其中的“小世界”效果。Jackson[14]通过社会网络方法进行数字人文历史研究，基于中世纪苏格兰人物数据库，使用网络密度模型研究该时期的关键人物，丰富了现有历史研究方法的研究结果。Quan-Haase等[15]通过研究数字人文学者使用社交工具，应用社会网络分析揭示了学者在社交活动中的网络效应。Algee-Hewitt[16]使用网络分析方法对1550-1900年英国舞台喜剧进行了量化模型分析，获取网络的不同分布特征，如特征值中心度和介度中心度等。其中演员作为网络节点，演员们通过共同进行舞台表演组成戏剧社会网络。Graham[17]使用网络科学工具对考古学或碑文数据集进行了分析利用，并提出网络分析工具将成为考古学数据工具系列中的常规工具。数字人文的社会网络研究主要为通过研究特色资源体系中人物之间的关联特性，分析其相互关联性和演化趋势。Robert[18]通过社会网络分析方法，对殖民地密西西比河谷的一个法裔印第安

人社区的研究中所提供的东西进行研究。

一般数字人文资源数据中并不存在社会网络信息，在进行社会网络分析前，需要专业技术人员在相关资源数字化过程中或数字化以后进行网络数据的处理和获取，主要信息包括节点、边、边方向和边权重。在获取网络数据之后，还需数字人文研究框架与人文学科领域研究目标，合理有效应用不同的社会网络分析方法，针对不同机构的社会网络数据（人物网络、交易网络等）进行不同的定量与定性分析，获取隐含知识信息，支持可视化展示。

3　上海交通大学古徽州契约文书社会网络分析实践

上海交通大学图书馆于2013年开始对本校历史系采购、收藏的古徽州府内6县的契约文书进行数字化加工与编目工作，截至2017年底已完成5.6万件数字化特藏资源[19]。古徽州所属区域主要包括目前安徽歙县、休宁、祁门、黟县、绩溪和江西婺源6个县。很早以来，中国人就通过订立契约来处理各种社会经济关系，契约在人们的日常生活中扮演着重要的角色，每个契约文书在交易发生时都会有多人参与其中，且身份不同，如买方、卖方、中间人和代笔人等。通过契约文书交易，两个人共同参与契约合同，即为产生社会关系，进一步形成古徽州各个地区的契约文书交易社会网络，通过对网络特性和结构进行分析将有助于从总体上把握古徽州契约交易的社会结构与属性。

在近代中国历史上古徽州地区有着相当繁荣的商业和社会经济活动，选取这一地区的数字化资源数据作为研究对象既有学术价值又具备典型推广意义。希望通过本节实践，为今后人文学科在计量化研究的基础上，进一步通过人文数据洞察新的研究特性，基于数据开展全新研究应用，并能引入新的视角和应用工具。

3.1　数据处理

上海交通大学图书馆对特藏古徽州地区文献中这些文献完成了契约的归户性、事主、事由、标的物和地域信息的基本编目标引；这批文献的前期数字化相关工作为后续的实例分析打下了扎实的研究基础。本文将对徽州府内6县相关的49 717件徽州契约文书数据进行社会网络化数据处理与分析。

表1　契约交易中的主要人员身份信息

序号	身份	次数
1	中見人	17 640
2	凭中	13 376
3	買人	8 444
4	業戶	8 110
5	代筆（人）	6 214
6	花戶	3 457
7	受稅人	2 703

由表1可见，在古徽州契约文书交易中，中间人是一个非常重要的角色，中间人（中見人或凭中）的斡旋给契约达成提供了一个谈判缓冲带，架起了一个辅助沟通的桥梁，具有积极意义和作用，也是契约文书交易中最多的角色，占了交易人总数的17%多。

3.2　网络参数分析

针对获取的数据，表2以MATLAB工具进行交易网络构建和主要网络参数分析。

特藏数：为每个地区的数字化契约文书数量，其中约92%的特藏为单页的文件。

表2　古徽州6个地区契约文书契约交易网络数据

项目 \ 地区	歙县	休宁	祁门	黟县	绩溪	婺源
特藏数	28 837	2 922	450	780	2 466	5 982
参与人数	101 242	7 914	1 259	2 521	7 492	17 480
去重后	59 174	4 426	969	1 635	5 185	12 341
网络节点数	55 678	4 032	879	1 397	4 185	11 534
网络边数	208 358	9 858	2 511	3 948	15 327	35 868
每个人的协作数	7.5	4.9	5.7	5.6	6.4	6.2

（续表）

项目 \ 地区	歙县	休宁	祁门	黟县	绩溪	婺源
无标度拟合 t	3.5	2.53	2.7	3.36	2.74	3.5
最大联通组	39 754	596	80	182	1 204	3 274
第二联通组	136	363	44	178	156	139
特征路径长度	8.7	7.0	3.6	3.9	12.2	12.9
聚类系数	0.123 8	0.184 5	0.356 6	0.309	0.117 9	0.097 1

参与人：表 2 中分别列出每个地区契约特藏的参与人数和按照姓名进行去重后的参与人数，去重后的每一个参与人将成为构建的交易网络的一个节点。

网络节点（边）数：本文对参与契约的人员进一步进行了数据处理，剔除了一些无法表示参与人员姓名的名称，如张氏、王氏等，将留下的姓名区分为不同节点。每一件数字化的契约文书数据中，都包含 a 个交易人员，这些人相互发生一次交易关系，即为一条交易网络的边，共生成 $a*(a-1)/2$ 条边；此处如果 $a=1$，即为一个孤立节点，没有形成交易关系。经统计，“歙县”契约社会网络中的丁能樞（縣知事）节点，参与了 104 篇契约交易，在网络中与 384 个节点连接，是该地区交易网络的最大中心节点。

平均协作数：表示特藏交易中，平均每个人和其他人发生交易关系的数值。

最大联通组：在一个大型网络中，可以分为不同的联通组（子网）。每个联通组具有不同的特征路径长度和聚类系数。

特征路径长度：祁门与黟县地区的交易人员特征路径长度较短，因此对应的整体聚类系数也更高。

3.3　幂律分布拟合

本节使用 MATLAB 直方图模拟工具（http://tuvalu.santafe.edu/~aaronc/powerlaws/bins/）对古徽州“歙县”契约社会网络中每个人的协作数进行直方图曲线拟合，拟合出公式 1 中对应的参数。

如图 1，蓝色为古徽州“歙县”契约社会网络中每个人的协作数量直方图，该分布满足幂律定律（参数为 3.5），红色为对应幂律分布曲线。由图 1 可见，大量的契约网络交易的协作者保持在 1–20 之内，整个直方图与小世界幂律分布非常拟合。

3.4　社区发现实验

结合生成的社会网络，作者对“歙县”契

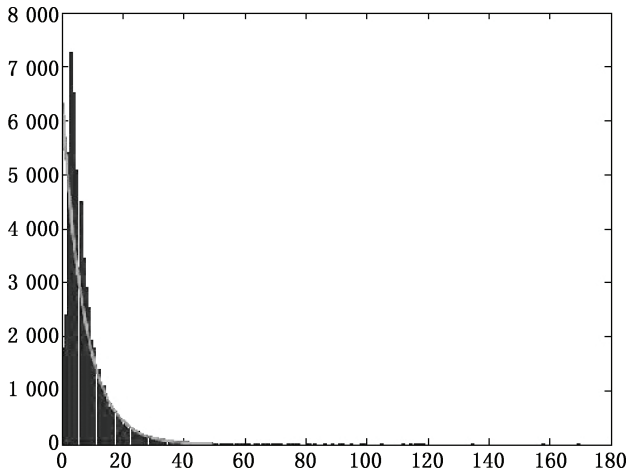

图 1　古徽州“歙县”协作人数分布与幂律分布曲线拟合图

约社会网络中的最大联通组的 39 754 个节点组成的子网络，进行了社区发现实验，网络节点度总和为 224 631。本文使用 Shi 等人[20]提出的基于对称贝叶斯非负矩阵分解框架的一种机器学习社区发现算法，简称 BSNMF。BSNMF 直观有效，运行速度快，能根据社区数据自动获取目标社区数量。由 BSNMF 方法对网络进行社区检测处理，获取的社区结果数 1 124，模块度达到 0.9，取得了较好的社区发现效果[21]。

表 3　古徽州“歙县”网络中发现的几个主要社区及节点信息

社区	节点数	节点度和	节点平均度	最大姓氏姓名	最大姓氏占比（%）
1	529	6 121	11.57	江（276）	52.2
2	482	3 445	7.15	胡（180）	37.3
3	418	2 461	5.89	方（101）	24.2
4	404	3 132	7.75	詹（88）	21.8
5	399	3 283	8.23	方（125）	31.3
6	390	1 786	4.58	方（276）	70.8
7	377	2 105	5.58	畢（92）	24.4
8	358	3 383	9.45	潘（186）	52.0

在发现的社区结果中，节点数排名前 8 的子社区在表 3 中显示，在这些社区中的节点人员通过共同参与契约交易，形成了内部更加紧密、与外部相对独立的“歙县”地区交易网络社区，在这些社区中的人员之间具有更加紧密的熟悉度关系，如亲属或属于相同氏族。由表 3 中看到发现的前 8 个社区中，7 个社区的节点度数都大于整个网络的平均节点度数 5.11。但是非常有意思的是，社区 6 的平均节点度数虽然小于整体网络平均度数，但是网络节点中的有 70.8% 都是姓方的人员，其同姓程度异常高。

针对发现的社区 1，我们通过 Cytoscape 可视化工具，进行网络社区可视化应用，网络图遵循力导向图布局[22]。

如图 2，在发现最大的一个社区中，江姓（黄色）有 276 位，占 52.2%；程姓位 52，占 9.8%，其余姓氏均不到 3%。由此，可以认为此网络为以江姓氏族为主形成的古徽州“歙县”契约交易紧密社区，根据地址分析，主要发生在“安徽省徽州歙縣二十六都”地区。

4　总结

社会网络分析方法可以有效在大量社会关系资源中，使用独特的网络化视角，发现潜在的社会特性和关联社区，对基于数字人文研究有着很好的拓展应用作用。数字人文社会网络分析的意义在于，它可以为人文科学中各种关系提供精确的量化分析，从而为构建理论模型和验证命题提供社会化例证，挖掘出更多隐含的社会关系和变化趋势，如小世界原理、无标度网络和内部关联

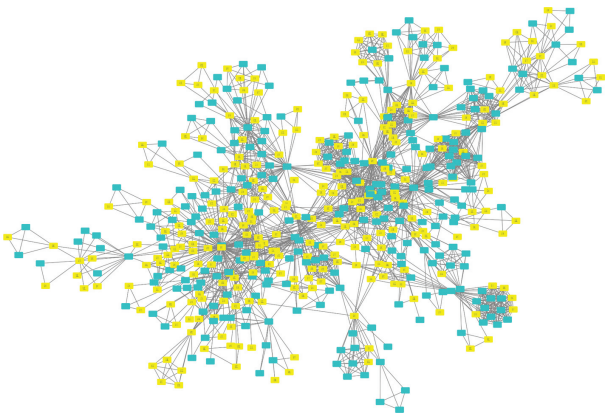

图 2　古徽州“歙县”网络社区结构图，黄色为江姓的交易人

社区等相关信息。同时，在以人物为关键要素的数字人文社会网络分析中，如何自动有效清洗、关联和甄别其中参与人员信息，是网络数据处理和清洗的关键，需要在数据处理中加入更多专家经验和智能化手段以取得新的突破。

经过多年来学者们在计算机人工智能学科研究的共同努力，利用社会网络分析进行数据应用创新的方法已经取得了丰硕的成果，但是在数字人文分析环境下与数字化特色资源和人文科学研究者结合应用的案例还相对偏少。期望通过本文介绍，能有更多计算机智能应用和方法能结合进入数字人文领域中来，充分利用前沿大数据处理技术与人工智能方法，来进行特色资源隐藏知识挖掘的分析和展示，推动数字人文实践与应用不断发展。

（本文数据链接地址：http://hdl.handle.net/20.500.12304/10242.）

参考文献

[1] 曾蕾, 王晓光, 范炜.图档博领域的智慧数据及其在数字人文研究中的角色[J]. 中国图书馆学报, 2018, 44(1): 17–34.
[2] Burdick A, Drucker J, Lunenfeld P, et al. Digital Humanities[M]. Cambridge: Mit Press, 2012.
[3] Drucker J. Introduction to digital humanities: course book: concepts, methods, and tutorials for students and instructors[M]. Los Angeles: UCLA, 2014.
[4] Charu C Aggarwal. Social Network Data Analytics[M]. New York: Springer Publishing Company, Incorporated, 2011.
[5] Steven H Strogatz. Exploring Complex Networks[J]. Nature, 2001, 410(6825): 268–276.
[6] D Watts, S Strogatz. Collective dynamics of “small world networks” [J]. Nature, 1998, 393, 440–442.
[7] Barabási A L, Albert R, Jeong H. Mean–field theory for scale–free random networks[J]. Physical A: Statistical Mechanics and its Applications, 1999, 272(1–2): 173–187.
[8] 丁连红, 时鹏. 网络社区发现[M]. 北京: 化学工业出版社, 2008.
[9] P J Carrington, J Scott S. Wasserman. Models and Methods in Social Network Analysis (Structural Analysis in the Social Sciences) [M]. Cambridge University Press, 2005.
[10] Kindred Britain[EB/OL]. [2019–05–10]. http://kindred.stanford.edu/.
[11] 胡静.数字人文在韩国史研究的应用探索——以杂科中人社会网络分析为中心[J]. 韩国研究论丛, 2018(2): 214–233.
[12] 严承希, 王军.数字人文视角: 基于符号分析法的宋代政治网络可视化研究[J]. 中国图书馆学报, 2018, 44(5): 87–103.
[13] Newman M E J. The structure of scientific collaboration networks[J]. Proceedings of the national academy of sciences, 2001, 98(2): 404–409.
[14] Jackson C. Using social network analysis to reveal unseen relationships in Medieval Scotland[J]. Digital Scholarship in the Humanities, 2017, 32(2): 336–343.
[15] Quan–Haase A, Martin K, McCay–Peet L. Networks of digital humanities scholars: The informational and social uses and gratifications of Twitter[J]. Big Data & Society, 2015, 2(1): 1–12.
[16] Algee–Hewitt M. Distributed Character: Quantitative Models of the English Stage, 1550–1900[J]. New Literary History, 2017, 48(4): 751–782.
[17] Graham S. On Connecting Stamps–Network Analysis and Epigraphy[J]. Les nouvelles de l’archéologie, 2014 (135): 39–44.
[18] Robert Michael Morrissey. Archives of Connection, Historical Methods: A Journal of Quantitative and Interdisciplinary History[J], 2015, 48(2): 67–79.
[19] 王昕, 张洁, 汤萌. 徽州契约文书地域信息组织与揭示的路径探究[J]. 新世纪图书馆, 2018(4): 55–59.
[20] Shi X, Lu H. Community detection in scientific collaborative network with bayesian matrix learning[J]. Frontiers of Computer Science, 2019, 13(1): 212–214.
[21] Newman M E J. Modularity and community structure in networks[J]. Proceedings of the national academy of sciences, 2006, 103(23): 8577–8582.
[22] Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks[J]. Genome research, 2003, 13(11): 2498–2504.

施晓华　上海交通大学图书馆，副研究馆员。研究方向：数字图书馆、社会网络分析与机器学习方法。作者贡献：整体设计、数据分析与论文撰写。E-mail：xhshi@sjtu.edu.cn　上海　200240
王　昕　上海交通大学图书馆，副研究馆员。研究方向：数字图书馆。作者贡献：数据校对、论文修订。　上海　200240

（收稿日期：2019–05–21　修回日期：2019–08–28）