

基于社会计算的科研合作网络信任图研究

徐欣欣, 胡风, 仲明, 王潮

(上海大学特种光纤与光接入网省部共建教育部重点实验室, 上海 200072)

摘要: 基于 SWTrust 架构, 利用复杂加权网络建立科研合作网络模型, 以中国科学引文数据库中合著论文的科研人员为例展开研究, 分析我国复杂网络科研的合作规模、规律等特征, 进一步验证了科研合作网之间的“小世界”特征和弱联结的理论。同时, 首次结合 Gephi 可视化优势构建信任图, 获得科研人员之间的最短信任路径。

关键词: 复杂网络; 科研合作网络; SWTrust 架构; 信任路径; 小世界

中图分类号: TN918

文献标识码: A

doi: 10.11959/j.issn.2096-109x.2017.00142

Trusted graphs of the scientific collaboration network based on social computing

XU Xin-xin, HU Feng, ZHONG Ming, WANG Chao

(Key Laboratory of Special Fiber Optics and Optical Access Networks, Ministry of Education (SCIE), Shanghai University, Shanghai 200072, China)

Abstract: Based on SWTrust architecture, using complex weighted network to establish research cooperation network model, the researchers who has co-authored papers in the chinese science citation database were studied, the scale, law and other characteristics of the complex network research were analyzed, to further validate the "small world" features and weak connection theory between the scientific collaboration. At the same time, for the first time in combination with the advantages of Gephi visualization, a trust graph was built and the shortest trust path between the scientific research was got.

Key words: complex network, scientific research cooperation network, SWTrust architecture, trusted path, small world

1 引言

随着大数据的席卷风潮, 政府、公司等都已积极投入这股势头。信息共享、交换的行为也日益剧增。信息公开意为共享政府和各种组织、机构所拥有的信息, 在正当理由的基础上, 选用简易的方式, 方便公众或其他组织、机构获取上述信息。2014 年, 八国峰会 G8 召开, 信息公开为主要推崇的目标, 并设置了多种原则作为访问、

发布和重复利用八国政府公开数据的基础。

在此大环境下, 中国也开始逐步落实数据开放。2015 年, 国务院印发的《促进大数据发展行动纲要》为开放数据提出了指导意见。目前, 上海市政府数据服务网 (www.datashanghai.gov.cn) 作为首个政府数据的服务网站, 在网页上提供了公共数据和应用的相关注册发布、地图展示等功能; 且涉及政府数据 200 多项相应的数据产品, 涵盖了六大领域。

收稿日期: 2016-11-23; 修回日期: 2017-01-20。通信作者: 王潮, wangchao@staff.shu.edu.cn

基金项目: 国家自然科学基金资助项目 (No.61272096, No.61572304)

Foundation Item: The National Natural Science Foundation of China (No.61272096, No.61572304)

近年来,随着复杂网络理论相关研究的热潮,许多其他相关的分支领域也逐一被开辟。日趋成熟的网络可视化技术可以定性地表示出各种网络的结构特征,从而直观地显示节点之间的连接方式及其特点。同样,对网络的各种属性进行定量的刻画也非常重要,特别是在网络规模庞大、结构复杂时,用可视化技术进行分析就显得比较局限。

随着技术水平的不断提高,需多个学科的研究者进行合作才能解决一些复杂的问题。并且,合作的规模会随着合作的增多而扩大。近年来,虽已有学者对我国科研合作的发展趋势和合作特点进行了研究,但这些研究不能充分展现出中国复杂网络科学合作的情况,不能直观地反映出作者间的合作网络关系,更无法揭示出深层的合作关系情况^[1~3]。2012年,Cho等^[4]定义了信任可用性和路径可靠性,并确定了一个信任链的最佳长度。2014年,Jiang等^[5]为大型社交网络提出SWTrust以生成小型信任图。

针对这些问题,本文在SWTrust架构的基础上,以复杂网络中常见的科研合作网络为研究对象,分别从理论和实际应用这2个方面,对其节点的重要性进行研究;结合复杂加权网络,构建科研合作网络模型,揭示我国复杂网络科研的合作规模、特点、规律及其网络特征等;并对其分析生成信任图,试图获知两者之间的最短信任路径。

2 复杂网络

2.1 复杂网络基础知识

复杂网络是一种基于复杂系统的关于数学的抽象,将系统内的元素以节点的形式在网络中表示,将元素间的交互情况以边的形式在网络中表示。

系统的复杂度、节点和节点之间的相互作用表明,个人之间的相互作用是不同的,且存在于不同网络中的2个共同节点间的相互作用也不尽相同。例如,在社会网络中,一位普通人或一个社会岗位都可以作为一个节点,朋友、同事等关系或业务上的来往,都可以为节点间的连边;在蛋白质相互作用网络中,节点可以代表不同的蛋白质,而它们之间的关系,则可以表示蛋白质之间的相互作用。

2.2 基于复杂网络的社会网络结构特征

近年来,越来越多的研究人员采用复杂网络作为研究社会网络的有力工具,即将社会网络(social network)抽象为复杂系统进行理论分析,然后再将理论分析结果逆向应用于社会网络。

社会网络是由一些个体或群体组成的网络,这些个体或群体因为某种原因有相应的连接关系;这些个体或群体被定义为复杂网络中的节点,它们之间的关系被定义为边^[6]。社会网络除了具有复杂网络的共同特征外,还有自己的特点^[7]。

1) “小世界”特征

小世界(small world)网络、规则网络和随机网络可以用复杂网络的聚类系数和平均最短路径长度来区分。一般而言,规则网络的聚类系数较大,且平均最短路径较长;随机网络的聚类系数较小,且平均最短路径较短。小世界网络则介于规则网络和随机网络之间,具有较大的聚类系数和较短的平均最短路径。

研究表明,现实世界中的社会网络通常具有海量数目的节点和边,但是其最短路径的平均值是很小的,而且聚类系数相对较大,这与复杂网络的“小世界”特征一致。“小世界”在社会网络中的研究源于1967年Stanley的实验^[8]。这一实验初步证实,一个人通过较小数目的连接就能和任何人相连,因此Stanley猜测人类社会实际上就是一个“小世界”。由于实验中的平均最短路径约为6,因此这个现象也被称为“六度分隔理论”。1998年,WS模型由Watts和Strogatz建立,更深入地揭示了“小世界”特征,掀起了小世界网络研究热潮。1999年,Newman^[9,10]在改进WS模型的基础上进一步提出了一个变体模型,这个模型有效避免了形成过程中孤立子图的出现,更易分析。1999年,Kasturirangan^[11]提出一个替代WS的模型。2000年,Dorogovtsev^[12]精确求解了Kasturirangan提出的替代模型,Kleinberg^[13]提出了最短路径长度可调的一般化模型。

2) 无标度的度分布

节点的度分布是复杂网络的一个重要特征,在不同的网络中具有不同的含义。在社交网络中,节点度表示节点的影响程度和重要性。度越大,该节点的影响力越大,对整个网络的影响也越大。

1999年, Barabási 发现实际网络中含有一定比例度值很大的节点, 并提出无标度网络 (scale-free network) 模型, 揭示了复杂网络的无标度特征。现实世界中的网络大部分都不是随机网络, 往往存在少数节点, 拥有大量的连接, 如图 1(a)所示, 合作数学家越多, 在网络中所占的比例越低。在任一网络中, 影响度越大的群或个体所占的比例越低。

无标度网络是符合幂律分布的节点度分布的网络, 它与节点数目固定的随机网络和 WS 小世界的网络模型相比, 能更好地刻画真实网络的结构特征。现实世界中的社会网络极度不均匀, 只有极少数节点的度值很大, 大部分节点的度值都很小。幂律分布意味着存在分形 (fractal) 或自相似性 (self-similarity), 即系统是由很多微小的自我复制或者复制组成^[14]。

3) 传递性或聚类

传递性也被称为聚类, 是衡量网络集团化的一个重要指标。一个网络节点与其邻居节点之间可能也有连接, 传递性用来描述这种可能性的程度。聚类系数则是传递性的量化指标。

集团化是社会网络的一个重要特征, 集团内部成员相互熟悉, 如同现实世界中的朋友圈或熟人

圈。例如, 在社会网络中, 有相同朋友的 2 个人碰面的次数相对较多, 两者之间的信任概率也会较大, 共同的朋友是他们建立朋友关系的桥梁。从网络拓扑来看, 这种关系意味着三角结构。聚类系数 C_i 是节点 v_i 与其邻居节点实际形成三角形数目与最多可能形成数目的比值。聚类系数越高, 节点 v_i 所处的三角形数目就越多。其中, 在网络中的聚类系数通过对总节点的聚类系数进行平均计算而获知。

在随机网络中, $C_i = p$, p 为节点间的连接概率。近年来, 越来越多的研究表明, 许多实际网络虽然都是稀疏图 $M = N$ (M 为社交网络中边的总数, N 为网络中的节点数), 但其聚类系数远高于同规模的随机网络。即这些网络的边分布在全局上是不均匀的, 数量有限的边集中分布在网络中的部分区域。实际网络趋于集团化, 如人的社会关系。

4) 社团结构

社团结构是大多数真实复杂网络都具有的一种重要结构特征^[15,16]。在图论中, 网络中节点的一个子集是一个社团, 社团内的节点具有密集的边连接, 而外部节点具有稀疏的边连接, 即社团内部边的密度远高于社团之间边的密度^[5,17], 如图 2 所示。

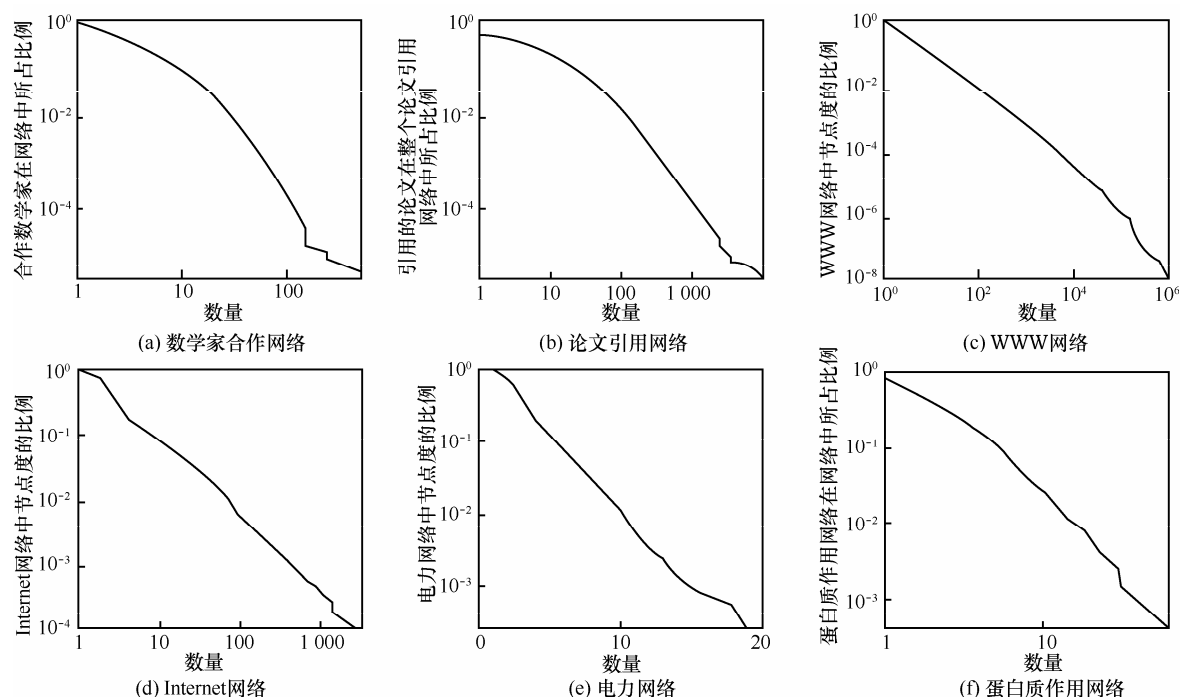


图1 6种真实网络的累积度分布

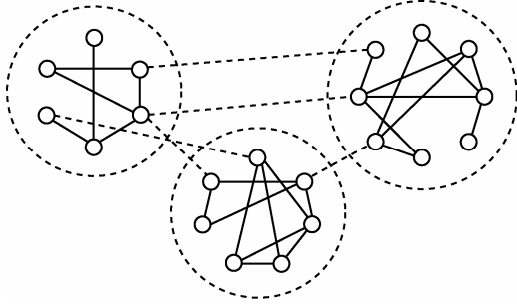


图2 具有3个社团的网络

社团网络对应复杂系统中的某种特定的属性或功能,网络的总属性或功能是各个社团共同作用的结果。在科研合作网络中,网络节点为科研人员,边反映了科研人员论文合作的次数,根据地域或单位划分研究组,从而构成社团^[16]。社团内部科研人员合作较多,而跨社团的合作较少。在爵士乐合作网络中,节点为音乐人,边反映2个音乐人在同一个乐队,社团就是按种族划分的黑人音乐人团体和白人音乐人团体;由于历史原因,每个乐队总是只由黑人或白人构成^[18]。

在惠普实验室的电子邮件网络中,实验室邮件用户是节点,实验室邮件用户之间的邮件收发关系构成边。该网络的社团就是部门成员集合,或某项目组成员集合,或某课题讨论组成员集合^[19,20]。每个社团内部的邮件联系是频繁的,而社团之间的成员邮件联系相对较少。在万维网络中,节点对应网页,边对应网页之间的超链接关系,社团对应一组具有相关或相似主题的网页集合。一个社团内部网页的链接数目远高出其与社团外部网页的链接数目^[21]。在无线通信网络中,以比利时手机通信网络为例(包含204万客户长达6个月的手机通信记录),节点代表手机用户,而加权边表示用户在6个月中的通信次数。社团就是不同语种(法语和荷兰语)手机用户的集合^[22]。

2.3 复杂加权网络

截至当前,大部分的研究都是关于无加权的网络。在网络模型中,只有2种可能性:节点间连接或不连接。若有连接,则其连接强度为固定值1;若不连接,则为0。但在许多实际网络中,网络中的每个节点间都有不同的权值。因此用加权网络模型来描述会更加合适^[23]。

加权网络可由集合 $G=(N,W)$ 描述。其中,包含 N 个节点和一组有权值的边 W 。通常,加权网

络用加权的邻接矩阵 $W=(w_{ij})$ 表示,其中, N 为网络的范围,即节点总数。矩阵元 $w_{ij}(w_{ij}>0)$ 表示节点 i 与 j 间边的权重。

1) 点强度与强度分布

点强度或点权 S_i 为加权网络中 k_i 的自然推广^[24],表示为

$$S_i = \sum_{j \in N_i} w_{ij} \quad (1)$$

其中, N_i 表示所有与节点 i 相连节点的集合。点强度将该节点的近邻数及相关权重都考虑在内,可综合体现出该节点的局域信息。基于点权,引入单位权 U_i ^[25],可以更详细地刻画出节点的连接和权重情况,单位权 U_i 为

$$U_i = \frac{S_i}{k_i} \quad (2)$$

U_i 表示节点连接的平均权重。然而,即使节点具有相同的点强度和度,相同的单位权也可能不同。例如,若单位权相同,则边的权重可能都接近 U_i 。此外,一条或少数边的权重也有可能处于优势。与节点连边的权重差异性可用 Y_i 表示^[26]。

$$Y_i = \sum_{j \in N_i} \left[\frac{w_{ij}}{S_i} \right]^2 \quad (3)$$

2) 加权网络的最短路径

在加权网络分析中,一个重要的问题是每条关联边之间的物理距离。权重为相异权时,可定义2个直接相连节点之间的距离为 $l_{ij} = w_{ij}$;反之,

则令 $l_{ij} = \frac{1}{w_{ij}}$ 。由于 w_{ij} 和 w_{jk} 的边将节点 i 和 k 相

连,基于相异权,可以直接对节点 i 和 k 间的距离取和,即 $l_{ik} = w_{ij} + w_{jk}$;而相似权,需对节点 i 和 k 之间的距离采取调和平均值的措施,即

$$l_{ik} = \frac{w_{ij} w_{jk}}{w_{ij} + w_{jk}}。$$

3) 加权集聚系数

集聚系数能反映出其一级近邻间的集团性质,且随着近邻间的关系越亲密而越高。加权集聚系数为

$$C_H^w = \frac{\sum_{j,k} w_{ij} w_{jk} w_{ki}}{\max_i x_{ij} w_{ij} \sum_{j,k} w_{ij} w_{ki}} \quad (4)$$

3 SWTrust 架构及最短信任路径设计

3.1 SWTrust 架构介绍

在大型社交网络与小型信任图之间有巨大的鸿沟^[27,28], 而 SWTrust 架构基于以下 2 点, 可以跨越这道鸿沟。基于弱联结的理论, 选择尽快连接到目标的长连接; 基于关联主题的度数, 为每一个步骤进行广度优先搜索, 找寻到可给予有效建议的信任邻居。SWTrust 有 3 个核心步骤, 即预处理社交网络 (PSN, preprocessing an online social network)、构建信任网络 (BTN, building the trust network)、生成信任图 (GTG, generating the trusted graph)。该架构适用于多数在线社交网络, 更适合于那些很容易从网络架构中定义域的在线社交网站。其具体流程如图 3 所示。

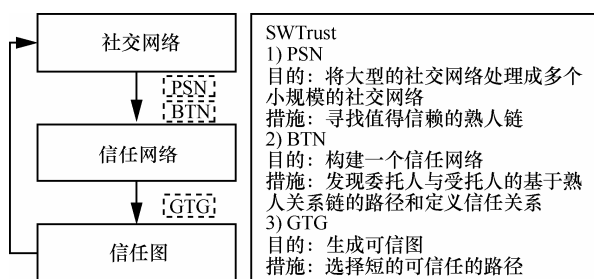


图 3 SWTrust 架构流程

目前, 由于很难为 Twitter 上的内容定义域或类别, 而这是基于用户域信任的熟人发现算法的关键, 因此还没有处理过这类 (如 Twitter) 的社交网络。然而, 使用数据挖掘技术进行预处理, 可以将 Twitter 上的内容分成属于不同主题或不同类目内容的方法, 来预测 Twitter 上用户之间的信任关系。

3.2 基于 SWTrust 的最短信任路径设计

设 A 为一个社交网络, 它可被看作一个大型图。首先, 由 A 中的 2 个选定节点 (委托人和受托人), 尽可能地找寻两者之间的短信任路径来有效地构建一个子图, 即从大型社交网络中生成一个信任图^[29]; 然后, 生成一个有质的信任图。

该架构适用于多数在线社交网络, 更适合那些很容易从网络架构中定义域的在线社交网站。SWTrust 的步骤如下。

3.2.1 PSN

PSN 的过程, 通常称之为基于用户域的可信熟人链发现。根据用户 i 与他们之间的社交距

离, 即用户之间的亲密度, 将用户 i 的邻居分为 3 类。

第 1 类: 本地邻居 (local neighbors), $d(i, ij)=1$ 。在这个类别中, $D_{ij} \subseteq D_i$, 表明这 2 个用户具有相当密切的联系。

第 2 类: 较长时间的联系 (longer contact), $d(i, ij) > 1$ 且 $d(i, ij) < D_{ij}$ 。在这个类别中, 用户 i 和其邻居 n_{ij} 之间有多于一个 (但不是全部) 的共同基本域, 表明这 2 个用户之间确实有一个连接, 但不可以不具有非常密切的联系。因此, n_{ij} 是用户 i 花费较长时间的联系人。

第 3 类: 最长时间的联系 (longest contact), $d(i, ij) > 1$ 且 $d(i, ij) < D_{ij}$ 。在这个类别中, 用户 i 和他的邻居 n_{ij} 仅有一个共同域。因此, n_{ij} 是用户 i 花费最长时间的联系人。

在每一个类别中, 以优先级降序排列邻居。通过以下方式, 从这 3 个类别中, 选择一致的下一级邻居节点。首先, 在第 3 类中, 选择一个优先级最高的节点; 然后, 在第 2 类中, 选择一个优先级最高的节点; 最后, 在第 1 类中, 选择一个优先级最高的节点。如果有必要选择更多的节点, 则在剩余的节点中同样迭代地执行以上步骤。

注意, 相同社交距离的邻居可能会被分成不同的类目。区别在于, 在第 3 类中一个邻居与用户 i 仅有一个公共域, 而在第 2 类中, 公共域至少有 2 个。公共域越多, 它们之间的相似处就越多; 相应地, 它们之间的差异更小。此处, 首次利用 Gephi 的可视化优势, 可直观地获知以上 3 个类目的区别, 更便于分析。

3.2.2 BTN

从源到头, 构建一个信任网络, 需要执行以下 2 个操作。1) 为 2 个选定的节点尽可能地找到多的短路径, 正如前所述, 这是一种典型的广度优先搜索。对于每个步骤, 使用 PSN 来发现可信赖的熟人链。2) 在直接连接的节点之间, 添加信任信息。

关于广度优先搜索, 有以下 2 种方法: 一种是集中式的方法, 更适合小规模在线社交网络, 如一个大学或一个公司的社交网络等; 另一种是分布式的方法, 该方法更适合运用在大型的在线社交网络, 它可以减轻服务器的负担。

在以下算法 1~算法 3 中, G 为在 PSN 处理后的社交网络, L 为路径的最大长度, c 用于控制

路径长度。令 $L_+(u)$ 表示由 PSN 过程中选出 u 的可信熟人集，并在最长时间的联系、较长时间的联系和本地邻居类目中，通过优先级递减顺序进行排序。设 R_{SOURCE} 为未被访问的节点，算法 1~ 算法 3 步骤如下。

算法 1 集中式广度优先搜索 CBFS

输入: $G, SOURCE, SINK$

输出: D (从 $SOURCE$ 到 $SINK$ 的路径集合)

- 1) 定义 $SOURCE$ 为当前节点，其中路径长度 $c = L - 1$ 。
- 2) 对于当前未被访问的任何邻居节点 u 执行以下操作
- 3) if 该邻居节点 u 为目标节点 $SINK$ then
- 4) 回溯到路径 P ，并添加到集合 D 中
- 5) else
- 6) if $c > 0$ then
- 7) $L_+(u)$ 中的所有节点添加到未被访问的节点集合中， $c = c - 1$ 。
- 8) 邻居节点 u 设为已访问状态。
- 9) end if
- 10) end if
- 11) end for

算法 2 分布式广度优先搜索 DBFS

输入: $G, SOURCE, SINK$ 。

输出: D

- 1) $c \leftarrow L - 1$ 。定义 $SOURCE$ 为当前节点
 - 2) 对于当前每个未被访问的邻居节点 u 执行以下操作。
 - 3) if 该邻居节点 u 为目标节点 $SINK$ ，then
 - 4) 邻居节点 u 向他的请求者发送响应，将 u 添加到回应者中；当响应发送到起始节点， $SOURCE$ 停止
 - 反向应答者得到路径 P ，并添加到集合 D 中
 - 5) else
 - 6) if $c > 0$ then
 - 7) $c = c - 1$ ，邻居节点 u 发送请求到可信熟人集 $L_+(u)$ 。
 - 8) 将邻居节点 u 设置为已访问
 - 9) end if
 - 10) end if
 - 11) end for
- 集中式广度优先搜索算法和分布式广度优先搜索算法描述如下。

1) 集中式广度优先搜索

CBFS 在存储网络信息的中央服务器中被执行。因此，服务器必须提供存储和计算。如果许多用户都要求某人的信任评估，则服务器会很忙碌。

2) 分布式广度优先搜索

在 DBFS 中，下一级邻居节点的选择由当前用户决定。之后，他将请求发送到所选择的邻居节点继续执行这一过程，并等待他们的响应。以下是用于描述这个过程的准备工作。

① 使请求者成为想要知道问题（根据主题，受托人的声誉如何）答案的用户，且委托人是第一位请求者。

② 使请求包含请求者、主题、目标和值等信息。

③ 使响应包含响应方的信息。

④ 使响应者成为已发出一个响应节点的序列。从受托人到委托人，反向应答将会得到一个委托人—受托人的路径。

如图 4(a)所示，在图 4(a)的基础上执行 CBFS 或 DBFS，得到路径 SOURCE-1-SINK、SOURCE-2-6-9-SINK、SOURCE-3-7-SINK、SOURCE-4-SINK。由于边和节点 8 未到达 SINK，因此，它们将被删除。由于节点 5 被访问时，它的邻居节点 9 已经被节点 6（由于 $P(2,6)=0.8$ 及 $P(2,5)=0.7$ ，因此，节点 6 比节点 5 有更高的优先级）访问过，并且节点 5 没有其他可用的邻居节点到达 SINK。因此，SOURCE-2-5-9-SINK 这条路径也被剔除。

3.2.3 GTG

该 GTG 过程的目标是从信任网络中选取短的信任路径。由于信任网络是相当小的处理，因此，可以以一个集中式的方法来执行这一操作。集中式生成信任图 (CGTG) 可在一个集中式服务器中生成信任图。设 D 代表经 BTN 过程后的结果路径集， $RT(i, j)$ 为从节点 i 到节点 j 的推荐信任， $e(i, j)$ 为路径的边。

算法 3 集中式生成信任图

输入: $D, SOURCE, SINK$

输出: D

- 1) 对于每一个 D 集合中的路径 P 执行以下操作
- 2) if $P > L$
- 删除路径 P
- 3) if 路径 P 的长度低于信任的临界值
- 删除路径 P
- 4) end for

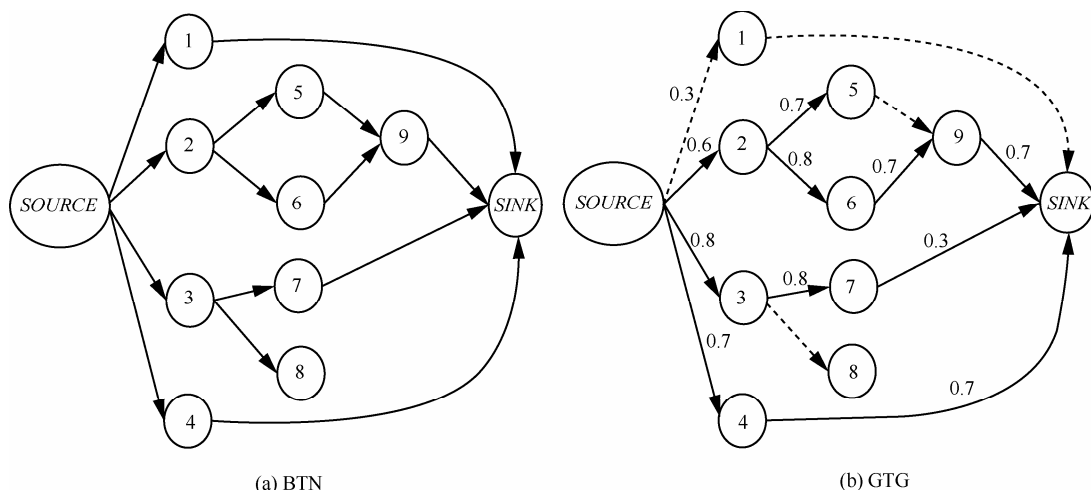


图4 一个经历 BTN 和 GTG 过程的案例

如图 4(b)所示, 假设信任值被标记在图 4(b)的边上, 且信任阈值 $th=0.5$ 。经 GTG 过程后, 由于从 SOURCE 到节点 1 的信任值低于 th , 路径 SOURCE-1-SINK 将被排除。截至当前, 从 SOURCE 到 SINK 的信任图生成。

4 验证 Gephi 与 SWTrust 架构的实际应用

本节在 Gephi 和 SWTrust 架构结合的基础上, 从 Web of Science 中获取中国科学引文数据库中的数据, 结合复杂加权网络, 构建出科研合作网络的模型, 并对其信任图进行分析。

4.1 科研合作网络数据的收集与整理

1) 数据来源

本文研究所用的论文数据来源于 Web of Science 中的中国科学引文数据库, 基于基本检索, 搜索主题为复杂网络, 搜索时间为 1989~2015 年。实际搜索结果显示共得到 2501 篇论文, 去除 1994 年、1995 年、1999 年、2000 年的冗余项, 剩下 2002~2015 年的数据, 共 2 496 篇论文。

2) 数据整理

统计每篇论文有几个作者可得该数据库的作者基本信息, 如表 1 和图 5 所示。

从表 1 和图 5 中可以看出, 独著论文的作者数目相对较少, 为 155 篇, 在发表论文中所占比例为 6.21%。以 2~6 人的小团队形式合作发表所占比例为 92.39%, 这些研究团队对新知识、新思想的传播具有较大的推动作用。这说明该合作网络合作规模较大, 作者间合作效果良好, 但大的团

队规模数量还是较少。构建网络模型时所考虑的数据来源还需同时满足: 不计独著论文; 若多人合著, 则作者必须至少有 2 人。符合这 2 个条件的论文共计 2 341 篇, 满足条件的作者共计 5 486 名。

表 1 Web of Science 中国科学引文数据库的论文作者数统计

作者数	论文数量	作者数所占论文比例
1	155	6.21%
2	617	24.72%
3	807	32.33%
4	581	23.28%
5	214	8.57%
6	87	3.49%
≥ 7	35	1.40%

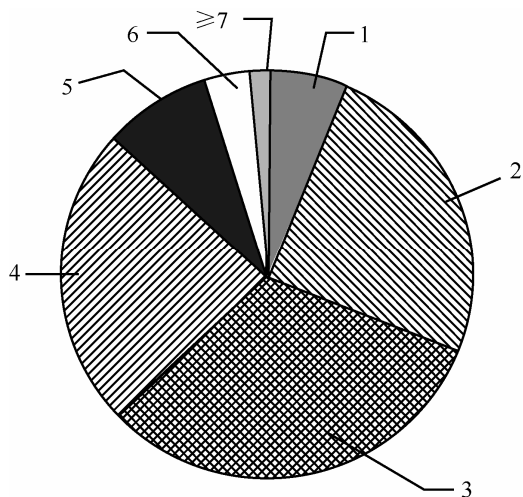


图5 作者数所占论文的比例 (数字代表作者数)

4.2 科研合作网络规模统计

本节主要对该合作网络的总体情况进行统计分析。首先,分析论文的合作概况,再统计其合作的规模,如表2所示。

由表2可以看出,论文数与节点数并不一定成正比关系,且呈下降趋势。这是因为后面年份的作者有很大一部分是前面出现过的,这说明作者主要以合作形式重复出现。

下面介绍几种常用的统计方法:合著率、合作指数与合作系数。

合著率是指合著的论文数与总论文数的比率^[1,30],用 DC 表示,为

$$DC = 1 - \frac{f_i}{N} \quad (5)$$

其中,用 f_i 表示合著者人数为 i 的论文; N 表示论文总数。

合作指数为个人、学科、机构、地域或由期刊发表的篇均作者数,以 CI 表示,为

$$CI = \sum_{j=1}^k \frac{jf_j}{N} \quad (6)$$

其中, k 表示合著者人数的最大值。

合作系数用 CC 表示,为

$$CC = \frac{\sum_{j=1}^k \frac{f_j}{j}}{N} \quad (7)$$

表3分别为该合作网络的合著率、合作指数与合作系数。

表3 2002~2015年合作网络的合著率、合作指数与合作系数

年份	合著率	合作指数	合作系数
2002	100%	2.4	0.43
2003	100%	3	0.35
2004	92.3%	3.12	0.39
2005	93.2%	2.86	0.41
2006	92.3%	3.05	0.39
2007	92%	3.06	0.40
2008	91.7%	3	0.40
2009	94.6%	3.1	0.38
2010	92.3%	3.07	0.39
2011	94.8%	3.15	0.38
2012	94.8%	3.27	0.37
2013	95%	3.32	0.36
2014	93.4%	3.42	0.36
2015	94%	3.39	0.36

表2

2002~2015年合作网络规模

年份	独著	合著	总论文数	节点数	平均节点数/总论文数
2002	—	10	10	24	2.4
2003	—	13	13	40	3.08
2004	2	24	26	82	3.15
2005	3	41	44	111	2.52
2006	6	72	78	212	2.72
2007	10	115	125	320	2.56
2008	14	154	168	414	2.46
2009	12	209	221	560	2.53
2010	22	264	286	715	2.5
2011	14	255	269	728	2.7
2012	15	271	286	775	2.71
2013	17	321	338	971	2.87
2014	20	281	301	923	3.07
2015	20	311	331	1 001	3.02

从合著率来看,该合作网络各年的合著率呈逐年增长的趋势,90%多的论文都是合作论文,这说明该合作网络的作者合作性比较强,作者间的关系较为紧密;从合作指数看,该合作网络平均每篇论文的作者数为2~3人,这与前文表1中作者数为3的论文数为最大的情况是相符合的。

分别将合作系数、合作率与合作指数做成线形图,由图6可得合著率曲线成稳定下降平缓趋势,而合著指数与合作系数更容易受关系、规模等的影响,不如合著率稳定。

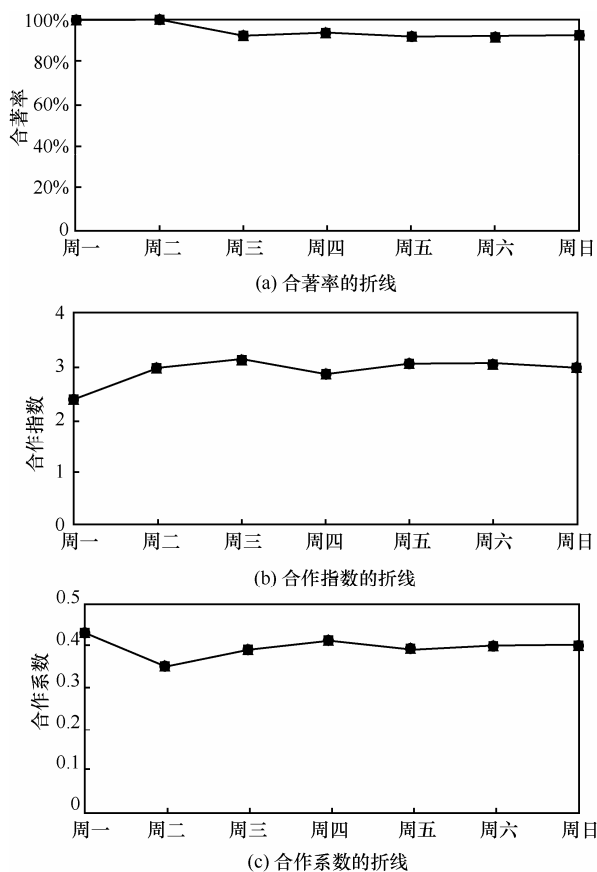


图6 合著率、合作指数与合作系数的折线图对比

4.3 科研合作网络的小世界网络特征分析

本文研究构建合作网络的方法是把每个作者作为网络中的节点,如果2个作者之间有合作关系,即共同发表一篇科研论文,就连一条边,将合作的次数记录为边的权重。

首先,从数据库中取出作者字段列的内容,对5486个作者进行编号,然后建立“作者—作者”邻接矩阵,有多少次合作关系即对应矩阵

的值,这样就得到邻接矩阵文件。样例如表4所示。

表4 邻接矩阵样例

Source	Target	Weight
汪秉宏	李平	2
汪秉宏	李明	1
汪秉宏	张文耀	1
汪秉宏	孙桂全	1
刘之景	周涛	1
刘之景	柏文洁	1
刘之景	汪秉宏	1
刘之景	严钢	1

利用 Gephi 的社交网络图形优势,可充分解说其社交网络内涵。利用 Gephi 这款可视化工具,可获得图7,此为吴杰教授 SWTrust 架构中的第一步 PSN。

从图7中可看出,中间聚集了部分作者,且作者与作者之间的联系十分紧密,体现出了“小世界”特征,充分验证了基于 Gephi 的 PSN 的可用性。其直观性与信息获取的容易程度,更凸显出结合 Gephi 的 PSN 的优势。

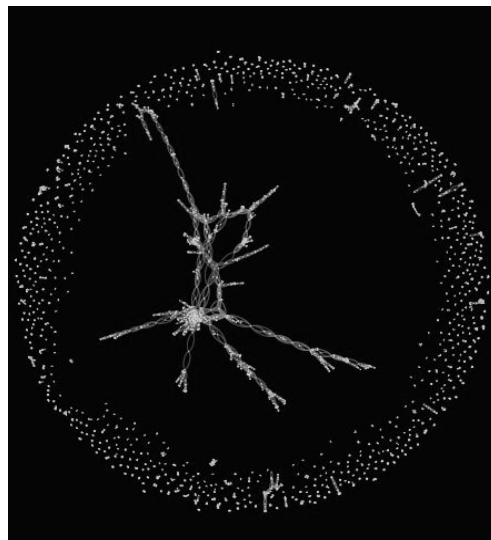


图7 科研合作网络权重

4.4 科研合作网络最短信任路径的可视化研究

由于数据间体现出“小世界”特征,因此,选取文章发表数在15篇以上(包含15篇)的作者,如表5所示,对其进行初步信任网络图的构

建。此处，由于该网络是合作网络，因此作者之间都被预设信任关系。

表5 发表15篇以上的论文作者

人名	度数	文章数 X	引用总次数	标准化影响因子
赵海	41	44	28	2.33
杨建梅	20	25	29	0.47
汪秉宏	49	22	220	6.28
方锦清	33	24	177	2.32
汪小帆	39	21	235	4.02
周涛	57	21	262	7.36
荣莉莉	10	17	56	1.28
张嗣瀛	15	18	32	1.41
姚灿中	9	17	21	0.66
井元伟	21	16	16	2.92
樊瑛	42	15	25	3.08

通过对比相关作者的标准化影响因子，可获知其度数、引用总次数与该标准化影响因子呈正相关的关系。其中，周涛为影响力最大人员，如图8所示，且查询该人在复杂领域的成果确实引人关注，SCI引用上千次，H指数目前已55。

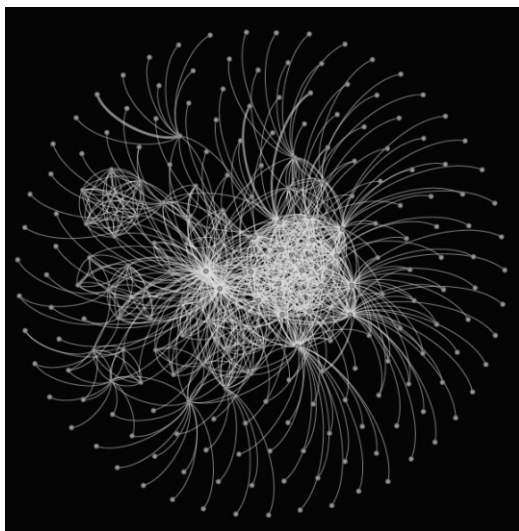


图8 周涛个人关系

构建 SWTrust 架构中的第二步 BTN，步骤如下。

1) 利用典型的广度优先搜索，为2个选定节

点找到足够多的短路径；在直接连接的节点中添加信任信息。

这里，在赵海关系网中选取连接最弱中的一员，此处选取贾思媛与赵海构建信任图。查找这两者之间的短信任路径是本次研究重点，即为吴杰教授 SWTrust 架构中的第三步 GTG^[31]。图9为赵海与贾思媛的 BTN 信任图。

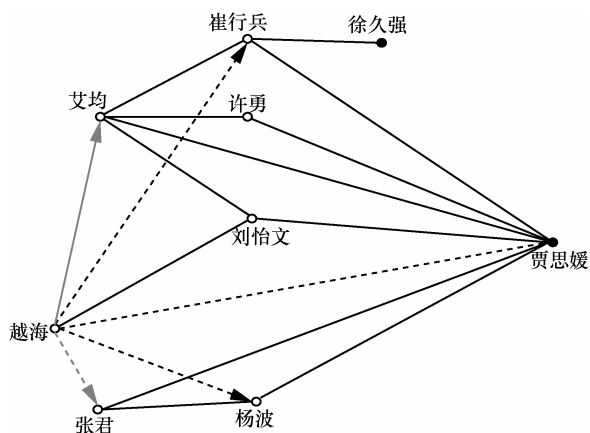


图9 赵海与贾思媛的 BTN 信任图

2) 根据权重，首先排除刘怡文。其次，由于徐久强未连接到贾思媛，因此，也排除。短信任路径可获知如下4条。

赵海，艾均，贾思媛

赵海，张君，贾思媛

赵海，杨波，贾思媛

赵海，崔行兵，贾思媛

3) 考虑权重比例，最短信任路径可选取路径1。通过短信任路径，可以获知两者之间的联系人，由可视化图可更直观地体现双方的关系，验证弱联结的理论。

基于 Gephi 的 BTN 信任图，可直观地获知两者之间的亲疏关系程度，更有利于 GTG 的分析。且 GTG 在该图的基础上，更易于对其进行最短信任路径的分析判断。这充分验证了结合 Gephi 的 SWTrust 架构更易于科研合作网络的研究，即该新方法可被采纳。

5 结束语

复杂网络理论广泛应用于社会的各个领域。以特定的方式构建出切合实际的演化模型，使其具有重要的理论意义和潜在的应用价值。本文选

取科研合作这一特定领域, 结合复杂网络理论和 SWTrust 架构, 利用可视化的优势, 探讨了该网络的特性。本文的主要贡献总结如下。

1) 通过深入分析节点和边的权重对节点重要性的影响, 从网络加权的角度, 以科研人员为节点, 合著关系为边, 将合作次数定义为边的权值, 提出了构建科研合作的加权网络模型。该方法综合考虑了节点间的边权值以及相互关系的信息, 能更全面地反映出节点在网络中的重要性程度。

2) 进一步研究了由吴杰教授提出的 SWTrust 架构^[5], 通过 PSN、BTN、GTG 等步骤筛选得到短信任路径。本文首次将 SWTrust 架构与 Gephi 结合, 并对科研合作网络进行短信任路径分析。

3) 通过收集、处理 Web of Science 中的中国科学引文数据库中的合著论文数据, 构建一个实际的科研合作网络。在 SWTrust 架构的基础上, 结合 Gephi 的可视化优势, 进一步协助其信任图的构建与分析, 可获知两者之间的最短信任路径, 验证了科研合作网之间的“小世界”特征和弱联结的理论。

参考文献:

- [1] DE-SOLLA- PRICE D J. Little science, big science[M]. New York: Columbia University Press, 1963: 38-45.
- [2] NEWMAN M E J. Scientific collaboration networks. I. network construction and fundamental results[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2001, 64(2).
- [3] NEWMAN M E J. The structure of scientific collaboration networks[C]//The National Academy of Sciences of the United States of America. 2001:404-409.
- [4] CHO J H, SWAMI A, CHEN I R. Modeling and analysis of trust management with trust chain optimization in mobile and hoc networks[J]. Journal of Network and Computer Applications, 2012, 35(3): 1001-1012.
- [5] JIANG W, WANG G, WU J. Generating trusted graphs for trust evaluation in online social networks[J]. Future Generation Computer Systems, 2014, 31: 48-58.
- [6] NEWMAN M E J. Who is the best connected scientist? a study of scientific coauthorship networks[J]. Complex Networks, 2004, 6(4): 337-370.
- [7] 高霖. 社会网络动态性及网络环境中的分布式搜索策略研究[D]. 合肥: 中国科学技术大学, 2009.
- GAO L. Research on social network dynamics and distributed search strategy in network environment[D]. Hefei: University of Science and Technology of China, 2009.
- [8] TRAVERS J, MILGRAM S. An experimental study of the small world problem[J]. Sociometry, 1969: 425-443.
- [9] NEWMAN M E, WATTS D J. Renormalization group analysis of the small-world network model[J]. Physics Letters A, 1999, 263(4): 341-346.
- [10] NEWMAN M E, WATTS D J. Scaling and percolation in the small-world network model[J]. Physical Review E, 1999, 60(6): 7332-7338.
- [11] KASTURIRANGAN R. Multiple scales in small-world graphs[J]. Disordered System and Neural Networks, 1999, 9(3): 56-71.
- [12] DOROGOVTSSEV S N, MENDES J F F. Exactly solvable small-world network[J]. Euro Physics Letters(EPL), 2000, 50(1): 161-167.
- [13] KLEINBERG J M. Navigation in a small world[J]. Nature, 2000, 406(6798): 845.
- [14] SCHROEDER M R. Fractals, chaos, power laws: minutes from an infinite paradise[M]. Courier: Dover Publications, 2009, 98-123.
- [15] GIRVAN M, NEWMAN M E J. Community structure in social and biological networks[J]. The National Academy of Sciences of the United States, 2002, 99(12): 78-21.
- [16] NEWMAN M E J, GIRVAN M. Find and evaluating community structure in networks[J]. Physical Review E, 2004, 69(2).
- [17] SCOTT J. Social network analysis[J]. Sociology, 1988, 22(1): 109-127.
- [18] GLEISER P M, DANON L. Community structure in jazz[J]. Advances in Complex Systems, 2003, 6(4): 565-573.
- [19] TYLER J R, WILKINSON D M, HUBERMAN B A. E-mail as spectroscopy: automated discovery of community structure within organizations[J]. The Information Society, 2005, 21(2): 143-153.
- [20] 赖大荣. 复杂网络社团结构分析方法研究[D]. 上海: 上海交通大学, 2011.
- LAI D R. Research on complex network community structure analysis method[D]. Shanghai: Shanghai Jiao Tong University, 2011.
- [21] FLAKE G W, LAWRENCE S, GILES C L, et al. Self-organization and identification of Web communities[J]. Computer, 2002, 35(3): 66-70.
- [22] BLONDEL V D, GUILLAUME J L, AMBIOTTER L, et al. Fast unfolding of communities in large networks[J]. Journal of Statistical Mechanics: Theory and Experiment, 2008(10): 108-130.
- [23] 吴金闪, 狄增如. 从统计物理学看复杂网络研究[J]. 物理学进展, 2004, 24(1): 18-46.
- WU J S, DI Z R. Research on complex networks from statistical physics [J]. Progress in Physics, 2004, 24(1): 18-46.
- [24] BARRAT A, BARTHELEMY M, VESPIGNANI A. Weighted evolving networks: coupling topology and weighted dynamics[J]. Phys Rev Lett, 2004,(92).
- [25] FAN Y, LI M, CHEN L, et al. Network of econophysicists: a weighted network to investigate the development of econophysics[J]. International Journal of Modern Physics B, 2004, 18(17): 2505-2511.
- [26] BARTHELEMY M, GONDRAN B, GUICHARD E. Spatial struc-

ture of the internet traffic[J]. Physica A Statistical Mechanics & Its Applications, 2003, 319(7):633-642.

- [27] WANG G, WU J. FlowTrust: trust inference with network flows[J]. Frontiers of Computer Science, 2011, 5(2):181-194.
- [28] WANG G, WU J. Multi-dimensional evidence-based trust management with multi-trusted paths[J]. Future Generation Computer Systems, 2011, 27(5), 529-538.
- [29] JOSANG A, HAYWARD R, POPE S. Trust network analysis with subjective logic[C]//The 29th Australasian Computer Science Conference(ACSC2006). 2006 :85-94.
- [30] SABRAMANYAM K, WU R Z. Study on the measurement of cooperation in scientific research [J]. Information Science, 1986(1): 94.
- [31] JIANG W, WU J, LI F, et al. Trust evaluation in online social networks using generalized flow[J]. IEEE Transactions on Computers, 2016, 65(3):952-963.

作者简介:



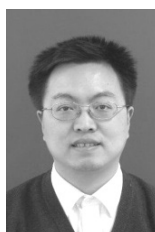
徐欣欣(1993-), 女, 江苏盐城人, 上海大学硕士生, 主要研究方向为信息安全、社会网络。



胡风(1991-), 男, 浙江温州人, 上海大学博士生, 主要研究方向为信息安全、量子计算密码、社会网络。



仲明(1990-), 女, 江苏无锡人, 上海大学硕士生, 主要研究方向为网络信息安全、数据可视化。



王潮(1971-), 男, 江苏镇江人, 上海大学教授、博士生导师, 主要研究方向为网络信息安全与椭圆曲线密码学、量子计算密码、社会网络。