基于弱连接理论的GitHub网络的分形特征分析

匡 立[1] 易云飞[1,2,3] 李元香[1]
(武汉大学软件工程国家重点实验室 武汉 430072)[1] (河池学院计算机与信息工程学院 宜州 546300)[2]
(广西混杂计算与集成电路设计分析重点实验室 南宁 530006)[3]

摘 要 研究者发现许多真实网络中存在分形特征,并广泛认为网络的异配性导致了分形性。因此,具有同配性的社交网络的分形特征很少被研究。针对开源合作平台中存在许多大型项目的开发者之间未必有实际的合作关系的现象,引入了边的权重来移除弱连接的边。通过重整化群分析,发现GitHub网络在移除弱连接的边之后,其网络结构从小世界变化为分形网络。此外,对网络的Pearson相关系数和邻居相关度进行分析后,发现网络具有很强的同配性,验证了之前对分形起源的理论分析。

关键词 开源社区,合作网络,分形特征,复杂网络,弱连接
中图法分类号 TP393.01 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.7.032

Analysis of Fractal Property on GitHub Network Based on Weak Ties Theory

KUANG Li[1] YI Yun-fei[1,2,3] LI Yuan-xiang[1]
(State Key Laboratory of Software Engineering, Wuhan University, Wuhan 430072, China)[1]
(College of Computer and Information Engineering, Hechi University, Yizhou 546300, China)[2]
(Guangxi Key Laboratory of Hybrid Computation and IC Design Analysis, Nanning 530006, China)[3]

Abstract Researchers have found that many real-world networks have fractal properties. It is widely believed that the fractal property origins from disassortative mixing. Thus, the fractal property of social networks, which are mostly assortative mixing, is rarely investigated. As in the open-source collaboration platform, there are many large projects, where many developers do not have real collaboration activities. In the paper, we introduced the power of links to remove the weak links in the network. Based on the renormalization group analysis, we found the network transfers from small-world to fractal network when we removed the weak links. Furthermore, by analyzing the Pearson correlation coefficient and neighbor connectivity, we found the fractal networks formed by strong links are assortative mixing. It enhances our former conclusion on the origin of fractality.

Keywords Open-source community, Collaboration network, Fractal property, Complex network, Weak link

1 引言

开源社区为软件开发提供了一个高效的合作平台。对开源社区进行社交网络建模分析,可以得到网络的统计特性及结构特征。对网络结构的特征分析,对于网络中的信息传播及知识积累有重要的意义。GitHub作为新兴的开源服务平台,受到了广泛关注。研究表明,GitHub的网络结构表现出无标度特性和小世界特征[1,2]。

由于大多数社交网络的同配性质(节点倾向于与自己度数相似的节点相连)和小世界特征,通常认为社交网络不具有分形特征。但是在社交网络的构建中往往会生成许多多余的连接,比如在GitHub的许多大型项目中(有的项目有上千人合作),项目中任意两个开发者之间都会建立连接,但是许多开发者之间并未产生实际的合作关系。因此,Gallos等人提出通过边的权重,即边两端的人的合作次数,将边分为弱连接和强连接[3],并发现去掉弱连接的边之后,IMDB演员合作网络由小世界网络变成分形网络。由此说明,重整化群分析[4]可以应用到类似的社交网络中。

本文通过GitHub的项目和开发者数据构建网络G,引入边权重的阈值w_T来消除网络中的弱连接,对消除弱连接后的网络Gw_T进行重整化群分析。结果表明,在不同阈值w_T下,由强连接构成的网络都显示出明显的分形特性。由于弱连接在网络的结构中起到重要作用,本文分析了边权重在网络中的分布,在整个网络中,边权重的分布表现出近似于重尾的幂律分布。即整个网络中弱连接占多数,强连接相对较少。进一步比较了在不同尺度的重整化方法下,盒子内部的边与

到稿日期:2014-11-20 返修日期:2015-02-09 本文受广西自然科学基金项目(2013GXNSFBA019282),广西高等学校科研项目(KY2015YB254),国家级大学生创新创业训练计划项目(201410605055,201510605024,201510605025),广西混杂计算与集成电路设计分析重点实验室开放基金课题(HCIC201411),广西自治区级大学生创新创业项目(201410605055,201410605056,201410605057)资助。

匡 立(1986-),男,博士,CCF会员,主要研究方向为复杂网络和演化计算;易云飞(1981-),男,博士,副教授,CCF会员,主要研究方向为复杂网络、智能计算等,E-mail:gxyiyf@163.com(通信作者);李元香(1962-),男,博士,教授,博士生导师,CCF高级会员,主要研究方向为演化计算和并行计算。

盒子外部的边的平均边权重。实验表明，盒子之间的捷径边的平均权重小于盒子内部的边。该结果进一步验证了网络中的强连接组成了分形结构，而主要由弱连接组成的盒子之间的捷径边导致了小世界现象[5]。

另外，研究者普遍认为异配性导致了分形网络的产生[6,7]。由于社交网络普遍具有的同配性，本文还分析了具有分形特征的 Gw_T 网络的邻居连接度及 Pearson 相关系数，结果表明 Gw_T 网络具有明显的同配性。这个结果印证了我们的前期结论[8]：异配性不是导致分形网络产生的原因。

2 复杂网络中的分形

分形是指在不同尺度上具有自相似性的图样。分形几何理论首先由 Mandelbrot 提出，用来解释一些具有超出自身拓扑维度的分形维度的数学集合，如 Cantor 集、Koch 曲线等[9]。传统意义上的分形都是指欧氏空间中的几何图形，而复杂网络的拓扑空间是否存在分形性成为研究者关注的问题。因此，宋等人提出一个重整化群分析法来度量复杂网络的分形性[4]。网络的分形维度由盒子覆盖法来确定，该方法类似于传统分形几何中的盒子计数法。盒子覆盖法 ℓ_B 的定义为使用最少的大小相同的盒子来覆盖整个网络，其中，盒子大小 ℓ_B 的定义为，盒子中任意两个节点的最短距离必须小于 ℓ_B。通过重整化群分析法，他们发现许多真实的网络具有分形性和自相似性，如万维网、蛋白质交互网络等。其中，分形性是指覆盖一个网络所需的最小盒子数量 $N_B(\ell_B)$ 与盒子的大小呈幂律关系，定义为：

$$N_B(\ell_B)\sim\ell_B^{-d_B} \tag{1}$$

其中，d_B 定义为分形的维度。而自相似性，又称为度分布的尺度不边性，是指重整化之后的网络的节点度，仍然服从与原网络相同幂指数的幂律分布，定义为：

$$P(k')\approx(k')^{-\gamma} \tag{2}$$

其中，k' 为重整化之后的网络的节点度，定义为每个盒子中最大的节点度；γ 为与原网络度分布相同的幂指数。

自从复杂网络中的分形理论被提出后，研究者开始广泛关注分形的起源问题，以及分形与小世界现象之间的关联。学者们普遍认为，网络中大量的长程连接（即捷径边）导致了网络从分形到小世界的相变现象[10]。最近，Gallos 等人从生物网络、大脑网络和演员合作网络的分析中发现网络中的强连接构成了分形网络结构，而由弱连接组成的捷径边导致了小世界现象[5,11]。因此，本文根据强、弱连接的概念，对开源社区合作网络结构的变化进行了分析。

3 数据采集和网络构建

本文采用 GHTorrent 项目所收集的社交合作数据。GHTorrent 建立了基于 GitHub 事件流和持久性数据的可扩展的离线镜像[12]，该项目提供了由 MySQL 存储的结构化数据。本文使用用户 id 和项目 id 作为唯一标识符来获取数据，避免数据的重复获取。本文所获取的数据包括从 2007 年到 2012 年之间注册的 1028472 个项目和 483438 个开发者。

以 $G=(V,E)$ 定义网络，其中，$V=\{v_1,v_2,\cdots,v_n\}$ 为节点集合，$E=\{e_1,e_2,\cdots,e_n\}$ 为边集合。节点代表开发者，边代表两个开发者在某一个项目中合作，边的权重表示合作的次数。同一个项目中所有的开发者之间两两相连。如图 1 所示，其中有项目 $P_1=\{v_1,v_2,v_3,v_4\}$，$P_2=\{v_4,v_5,v_6\}$ 和 $P_3=\{v_2,v_4,v_7,v_8\}$ 用虚圆圈表示，开发者 v_2、v_4 在项目 P_1、P_2 中都有合作，该边权重为 2，其余的边权重都为 1。

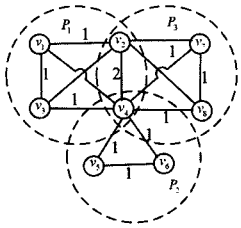

图 1 由项目和开发者构建的网络

4 网络分析

4.1 无标度和小世界特性

根据以上创建网络的方式，本文所创建的 GitHub 网络的每条边都具有一个权值，代表两个开发者合作的次数，并称权值较小的为弱连接的边，权值较大的为强连接的边。这里，引入权重的阈值 w_T 来分析网络的特性和结构。网络 Gw_T 的定义为移除 G 中所有权值小于 w_T 的弱连接边所形成的网络。为了保证网络的连通性，提取网络的最大连通子图进行分析，所得到的网络的直径 ℓ_{max} 和节点数量 N 如表 1 所列。

表 1 网络的直径与节点数量

	$w_T=1$	$w_T=2$	$w_T=4$	$w_T=8$
N	82914	20283	7231	1819
ℓ_{max}	25	25	18	13
$\ln(N)/\ell_{max}$	0.453	0.3967	0.4936	0.577

首先，分析了网络的无标度性。无标度特性是指网络的度分布服从幂律分布，如式(3)所示。

$$P(k)\approx(k)^{-\gamma} \tag{3}$$

其中，不同阈值下的网络的度分布如图 2 所示，坐标轴采用双对数，表现出不同幂指数的重尾的幂律分布。在 w_T 为 1 和 2 时，幂律分布比较显著，其中，$w_T=1$ 时的幂指数为 1.205；而在 w_T 为 4 和 8 时，度分布较分散，幂律分布并不明显。

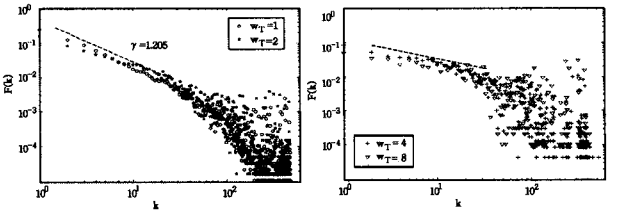

图 2 不同阈值 w_T 下的度分布

众所周知，小世界网络的定义为：

$$\ln(N)\propto\ell_{max} \tag{4}$$

从式(4)可知，网络的直径接近于网络中节点的数量的自然对数。为分析网络的小世界特性，计算了各网络的直径，从表 1 可以看出，不同阈值的网络都符合小世界网络的定义。

4.2 重整化群分析

为了分析网络的分形性，对不同边权重阈值 w_T 下的 GitHub 网络进行重整化群分析，并采用基于贪婪着色的盒子覆盖法[13]。盒子覆盖的结果如图 3 所示，$w_T=1$ 的网络的盒子数量随盒子的大小表现出指数衰减，不具有分形性。而 w_T 为 2、4 和 8 时，3 个网络的盒子数量随着盒子的增大呈明显的幂律衰减，表现出明显的分形性。其中，通过拟合得到各个网络的分形维度非常接近，分别为：$w_T=2$ 时，$d_B=2.86$；$w_T=4$ 时，$d_B=2.87$；$w_T=8$ 时，$d_B=3.19$。

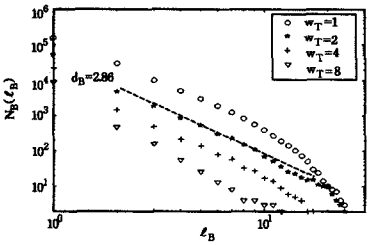

图 3　不同阈值 w_T 下的最小盒子数与盒子大小的关系

此外，分析了 $w_T=2$ 和 $w_T=4$ 时分形网络的自相似性，即在重整化之后的网络的度分布具有尺度不变性。如图 4 所示，在不同的尺度下，两个网络具有幂指数相似的度分布。其中，$w_T=2$ 的网络重整化之后的几个网络的度分布幂指数大约为 $\gamma=1.838$，而 $w_T=4$ 的网络的幂指数约为 $\gamma=1.568$。实验结果表明这些网络具有明显的自相似性。

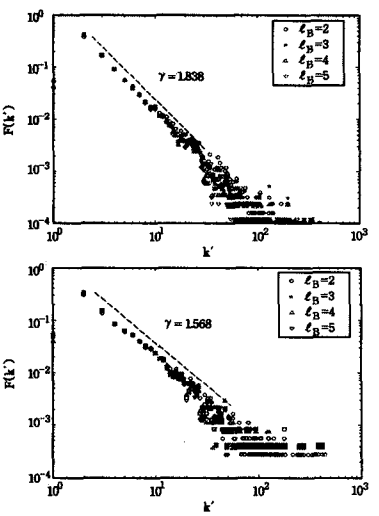

图 4　阈值 $w_T=2$ 和 $w_T=4$ 时度分布的尺度不变性

4.3　权重分布分析

由以上分析可知，移除网络中的弱连接边后，网络结构由小世界非分形网络变成了小世界分形网络。边的权重在网络的结构中起了关键的作用，因此这里分析了边的权重在整个网络中的分布。如图 5 所示，在 $w_T=1$，即未删除弱连接边时，网络中的权重分布近似于重尾的幂律分布。这种重尾分布的原因，可能是网络中存在富人俱乐部现象，即有一部分经验丰富的开发者形成小团体，团体内部有广泛的合作，而与外部的合作较少。

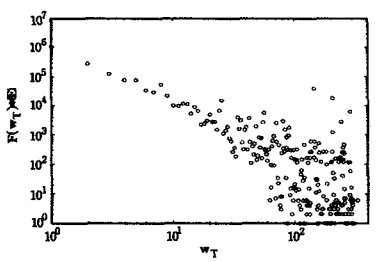

图 5　$w_T=1$ 的网络中边权重的分布

为了进一步了解边的权重对网络结构的影响，还分析了在不同尺度下的盒子覆盖后的网络中盒子内与盒子外的边（即捷径边）的平均权重。由于盒子覆盖法的结果具有随机性，本文打乱盒子覆盖的节点顺序，重复实验 30 次之后取平均值。如图 6 所示，盒子外的捷径边的平均权重〈w〉虽然有先下降后上升的趋势，但是总体都比盒子内的边的平均权重小，表明捷径边中的弱连接数量要多于盒子内的非捷径边。这与弱连接导致了网络从分形结构转变成小世界网络的结论吻合。

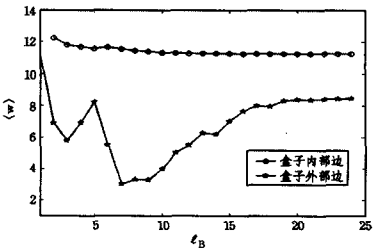

图 6　$w_T=1$ 的网络在重整化之后网络内外边的平均权重

4.4　同配性分析

通常的社交网络都具有同配性，即节点倾向于与自己度数相似的节点相连，高度数节点倾向于与高度数节点相连，低度数节点倾向于与低度数节点相连。这种同配性通常会导致网络中的富人俱乐部现象，而使得网络显示出重尾的幂律分布。因此本文分析了不同阈值 w_T 下的网络的同配性。

同配性通常有两个判断标准：Pearson 相关系数和邻居相关度[14]。Pearson 相关系数定义如下：

$$r=\frac{|E|^{-1}\sum_{i=1}^{|E|}j_ik_i-[|E|^{-1}\sum_{i=1}^{|E|}\frac{1}{2}(j_i+k_i)]^2}{|E|^{-1}\sum_{i=1}^{|E|}\frac{1}{2}(j_i^2+k_i^2)-[|E|^{-1}\sum_{i=1}^{|E|}\frac{1}{2}(j_i+k_i)^2]} \tag{5}$$

其中，j_i 和 k_i 为第 i 条边两边节点的度数，$|E|$ 为网络中边的总数。当 r 为正时表示网络具有同配性，数值越高同配性越强；当 r 为负时表示网络具有异配性。如表 2 所列，$w_T=1$ 的网络表现出非常弱的同配性，而当移除弱连接后，$w_T=2$，4 和 8 的网络表现出非常强的同配性。

表 2　网络的 Pearson 相关系数

	$w_T=1$	$w_T=2$	$w_T=4$	$w_T=8$
r	0.086	0.952	0.960	0.956

邻居相关度的定义为度数为 k 的所有节点的邻居节点的平均度数。公式为：

$$\langle k_{nn}\rangle=\sum_{k'}k'P(k'|k) \tag{6}$$

其中，$P(k'|k)$ 是从度数为 k 的节点连一条边到度数为 k 的节点的条件概率。邻居相关度可以更加直观地观察网络中不同度数节点连接的倾向性。如果曲线的斜率为正，则网络为同配，否则为异配网络。如图 7 所示，$w_T=1$ 的网络在 $k<440$ 时表现出同配性，而 $k>440$ 时网络表现出异配性。图 7 的内插图为坐标取双对数后的邻居相关度，当 $k<10$ 时，网络表现出异配性。图 8—图 10 分别表示 $w_T=2$，4 和 8 这 3 个网络的明显的同配性质。

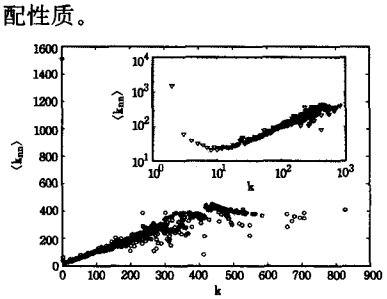

图 7　$w_T=1$ 网络的邻居相关度

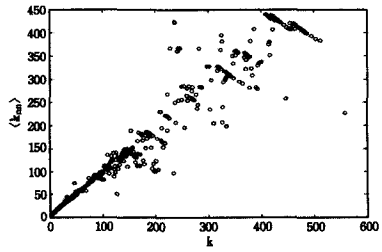

图 8 $w_T=2$ 网络的邻居相关度

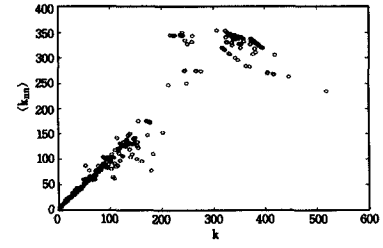

图 9 $w_T=4$ 网络的邻居相关度

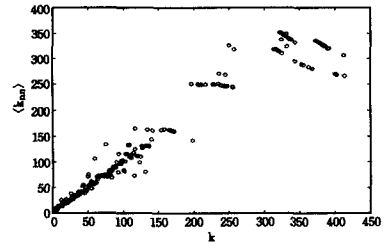

图 10 $w_T=8$ 网络的邻居相关度

以上分析表明 $w_T=2$、4、8 的网络具有分形性并有很强的同配性，而 $w_T=1$ 的网络具有非常弱的同配性，但是并不是分形网络。因此，本文实验的结果表明，异配性并非是分形产生的原因。在前期工作中，我们提出了一种同配并且 hub 吸引的分形网络模型，并指出演员合作网络也具有分形和同配性[8]。本文对开源社区合作 GitHub 网络的分析，更进一步验证了该结论。这对分形网络的理论研究具有一定的意义。

为更直观地观察网络的结构，本文用 Gephi 软件绘制出了 $w_T=4$ 和 $w_T=8$ 的最大连通子图。如图 11 和图 12 所示，在移除了大量弱连接的边之后，网络中仍然表现出明显的社团特性。大量的节点聚集到一起进行密切的合作，而与社团外部的合作较少，因此网络具有较强的同配性。

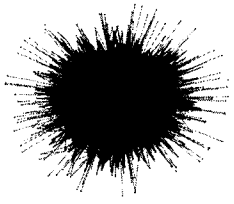

图 11 $w_T=4$ 的网络

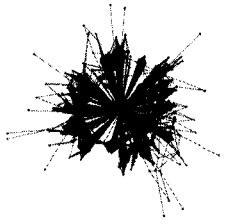

图 12 $w_T=8$ 的网络

结束语 本文以开源平台 GitHub 的项目和开发者数据构建合作网络，通过边的权重（即该边所连接的两个开发者的合作次数）将边划分为强连接和弱连接，发现去掉弱连接的边之后网络从小世界转为分形网络。为了了解强、弱连接在网络结构中的作用，比较了在不同尺度的重整化方法下，盒子内的边和跨盒子的边的平均权重，发现盒子间的捷径边权重小于盒子内的边。印证了强连接构成了分形结构，由弱连接产生的捷径导致了小世界的现象。此外，本文发现由强连接组成的分形网络具有强同配性，印证了异配性不是分形产生的原因的结论。针对网络的结构特性，研究开源合作网站中知识的积累及其如何影响项目的成败，将是后期工作的重点。

参 考 文 献

[1] Dabbish L, Stuart C, Tsay J, et al. Social coding in GitHub: transparency and collaboration in an open software repository [C]// Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work. ACM, 2012: 1277-1286

[2] Thung F, Bissyandé T F, Lo D, et al. Network structure of social coding in GitHub [C] // 2013 17th European Conference on Software Maintenance and Reengineering (CSMR). IEEE, 2013: 323-326

[3] Gallos L K, Potiguar F Q, Andrade Jr J S, et al. Imdb network revisited: unveiling fractal and modular properties from a typical small-world network[J]. PloS one, 2013, 8(6): e66443

[4] Song C, Havlin S, Makse H A. Self-similarity of complex networks[J]. Nature, 2005, 433(7024): 392-395

[5] Gallos L K, Song C, Havlin S, et al. Scaling theory of transport in complex biological networks[J]. Proceedings of the National Academy of Sciences, 2007, 104(19): 7746-7751

[6] Song C, Havlin S, Makse H A. Origins of fractality in the growth of complex networks[J]. Nature Physics, 2006, 2(4): 275-281

[7] Yook S H, Radicchi F, Meyer-Ortmanns H. Self-similar scale-free networks and disassortativity[J]. Physical Review, 2005, 72(4): 045105

[8] Kuang L, Zheng B, Li D, et al. A fractal and scale-free model of complex networks with hub attraction behaviors [J]. Science China Information Sciences, 2015, 58(1): 1-10

[9] Mandelbrot B B. The fractal geometry of nature[M]. Macmillan, 1983

[10] Rozenfeld H D, Song C, Makse H A. Small-world to fractal transition in complex networks: a renormalization group approach [J]. Physical Review Letters, 2010, 104(2): 025701

[11] Gallos L K, Makse H A, Sigman M. A small world of weak ties provides optimal global integration of self-similar modules in functional brain networks[J]. Proceedings of the National Academy of Sciences, 2012, 109(8): 2825-2830

[12] Gousios G. The GHTorent dataset and tool suite [C]// Proceedings of the 10th Working Conference on Mining Software Repositories. IEEE, 2013: 233-236

[13] Song C, Gallos L K, Havlin S, et al. How to calculate the fractal dimension of a complex network: the box covering algorithm [J]. Journal of Statistical Mechanics: Theory and Experiment, 2007, 2007(3): P03006

[14] Newman M E J. Assortative mixing in networks[J]. Physical review letters, 2002, 89(20): 208701