

# 开源软件社区开发者偏好合作行为研究

何鹏<sup>1,2</sup> 李兵<sup>3,4</sup> 杨习辉<sup>1,2</sup> 熊伟<sup>1,2</sup>

(武汉大学软件工程国家重点实验室 武汉 430072)<sup>1</sup> (武汉大学计算机学院 武汉 430072)<sup>2</sup>

(武汉大学国际软件学院 武汉 430079)<sup>3</sup> (武汉大学复杂网络中心 武汉 430072)<sup>4</sup>

**摘要** 着重从开发者角度出发,先对 SourceForge.net 开源社区项目与开发者数量增长情况进行统计分析,以见证社区的快速发展;随后以两个月为时间段分析新增开发者、合作的数目,并将开发者之间新建的合作细分为4种情况,探析4种合作方式情况下的差异,进而判断社区开发者优先选择合作方式的顺序;最后针对新开发者与社区已有开发者之间的合作,分析了新开发者的合作偏好与已有开发者的度数中心性、介数中心性和接近中心性,以及他们的开发项目数与之前项目中角色的关系,发现新成员优先选择与介数中心性或度数中心性大的已有开发者合作,且这些开发者整体上都具有多次开发经验并在开发过程中担任过特定角色。研究结果有利于优化群体软件开发过程,为提高群体软件开发效率与质量水平奠定了基础。

**关键词** 群体开发, 社会网络分析, 偏好合作, 行为分析

**中图分类号** TP301 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.2.035

## Research on Developer Preferential Collaboration in Open-source Software Community

HE Peng<sup>1,2</sup> LI Bing<sup>3,4</sup> YANG Xi-hui<sup>1,2</sup> XIONG Wei<sup>1,2</sup>

(State Key Laboratory of Software Engineering, Wuhan University, Wuhan 430072, China)<sup>1</sup>

(School of Computer, Wuhan University, Wuhan 430072, China)<sup>2</sup>

(International School of Software, Wuhan University, Wuhan 430079, China)<sup>3</sup>

(Complex Network Research Center, Wuhan University, Wuhan 430072, China)<sup>4</sup>

**Abstract** This paper mainly focused on the analysis of developer's behavior in open-source community. At first, we analyzed the growth of the number of projects and developers in SourceForge.net community to witness its rapid development. Then, we investigated the quantities of new developers and collaborations in a two months interval, and divided the new collaborations into four categories to explore their differences and then judge the cooperation order among developers. Finally, with respect to the collaboration between new and old members, we further analyzed the relationship between preferential behavior and centrality measures such as degree centrality, betweenness centrality and closeness centrality, the number of projects developed and their roles. The result shows that a new developer will prefer collaborating with those who have great betweenness centrality or degree centrality, because they develop more projects and play important roles. Our work will optimize the development process of collaborative development, and lay a solid foundation to improve the productivity and quality of software.

**Keywords** Collaborative development, Social networks analysis, Preferential cooperation, Behavior analysis

随着云计算的日益普及和广泛应用,以互联网为运行平台的开源群体软件开发新模式被越来越多的人所认可和采用。互联网已成为一个面向公众的资源丰富(包括计算资源、数据资源和软件资源等)的公共基础设施,为群体开发提供了一个崭新舞台。在网络环境所带来的丰富资源面前,资源共享、人人参与的步伐得到急剧加速,依托于群体开发的各类虚拟社区开源软件生产组织模式也得到了快速发展,人与人之间社会关系分析也已成为研究热点。

开源软件社区(如 SourceForge.net,简称 SF.net)作为一类典型的虚拟群体开发社区,来自不同国家、有着不同背景的开发者通过加入社区与其他开发者互惠合作。一个开源软件起初通常由单个人/群体/组织向开源平台提供整个系统、项目组件的源代码和交互设施,社区其他感兴趣的开发者可以下载源码,参与到项目的开发工作。与传统软件开发不同,这些开发者大多都是自愿者,他们的动机是丰富经验、增长知识或娱乐<sup>[1]</sup>,但他们之间合作紧密。

到稿日期:2014-01-14 返修日期:2014-07-13 本文受国家重点基础研究发展计划(2014CB340401),国家自然科学基金(61273216, 61272111, 61202032, 61202048),湖北省重大科技创新计划(2013AAA020),江苏省电子商务重点实验室开发基金(JSEB2012-02)资助。

何鹏(1988-),男,博士生,CCF会员,主要研究领域为软件度量、缺陷预测、社会网络分析, E-mail: penghe@whu.edu.cn; 李兵(1969-),男,教授,博士生导师,CCF高级会员,主要研究领域为面向服务软件工程、复杂网络、云计算和人工智能, E-mail: bingli@whu.edu.cn(通信作者); 杨习辉(1987-),男,硕士生,主要研究领域为面向服务软件工程; 熊伟(1973-),男,讲师,主要研究领域为服务发现和服务推荐。

近年来,随着开源运动在全球逐渐流行,开源软件开发备受关注<sup>[2]</sup>,包括 Linux 操作系统、Mysql 数据库、Tomcat 服务器、Firefox 浏览器等。同时,各大开源社区(如 Sourceforge, Google code, Rubyforge 等)的开源软件与开发用户数大幅度增长,见证了开源的快速发展和广泛采纳,证明了源于网络用户贡献的集体群体智慧的力量。这引起了学术界的广泛关注,其开始着力探究开源社区群体开发成功的缘由、如何来衡量项目的成功以及决定成功的因素、开发者作为志愿者的动机和他们之间协调合作的机理等<sup>[3]</sup>。

群体软件工程源于开源软件开发,不仅是一种依托开放、开源的软件合作社群<sup>[4]</sup>,而且正向基于分享、交互与群体智慧的同侪生产(peer production)方式发展。在此方式下,使用者与设计者、开发者、维护者之间不再壁垒森严。交互既是群体开发的核心问题,也是研究的重要突破口。

本文分析了 SF.net 开源社区的演化特性和开源环境下开发者的群体交互合作偏好行为。本文第 1 节介绍了开源社区开发的相关研究工作;第 2 节主要对 SF.net 社区的演化特性进行分析;第 3 节引入中心性、开发项目数和角色身份 3 类指标来衡量节点的重要性及与偏好合作的关系;第 4 节主要陈述本文的研究方法;第 5 节是实验分析与验证过程;最后是相关讨论与全文总结。

## 1 相关研究

Hossain 和 Zhou<sup>[5]</sup>研究了度数中心性、介数和开发者网络密度对软件缺陷修复数和不同级别缺陷的影响。Wolf<sup>[6]</sup>等人结合多个网络指标检测了分布式软件团队开发产品的质量。Hind<sup>[7,8]</sup>等人先后研究了封闭性(Closure)、密度(Density)、中介性(Bridging)和中心性(Centrality)等社会网络结构对开源社区项目的成功的影响。毛清华<sup>[9]</sup>等人针对嵌入于社会网络的虚拟团队进行中心性分析,识别中心度高的成员以及经纪人角色,挖掘、利用隐含于网络中的知识,并依靠中心位置行动者的声望和权力的影响、支配作用,达到快速知识共享的目的。Tora<sup>[10]</sup>等人证实了中介者角色对社区取得成功的重要性,解决了社区团队间交互问题,类似工作可参考文献<sup>[11-13]</sup>。

Pinzger 等<sup>[14]</sup>构建开发者贡献网络,使用中心性指标衡量贡献网络的重要程度,指出中心模块比非中心模块存在更多的缺陷。Huang 等<sup>[15]</sup>研究了开发者与模块之间的组织结构,并对开发者的角色重要性进行计算,证实开源开发过程的边缘参与学习过程<sup>[16]</sup>。文献<sup>[17,18]</sup>根据中心性阐述了开源软件开发团队中的核心-边缘结构。

以上工作采用的实验对象大多是基于单个成功的或少量的几个开源软件,主要是探析项目开发网络结构,以及网络指标对项目质量的影响和缺陷预测。据统计,以整个开源社区为研究对象,分析新开发者的偏好合作行为的研究非常少。Hahn<sup>[19]</sup>等人分析了开源软件开发团队的形成和开发者的加入行为,并验证了原有的合作关系对软件团队的影响。Cavrak<sup>[20,21]</sup>等人对分布式软件开发者合作行为模式、动机和角色进行了分析,以便更好地理解分布式项目的开发动态。本文的主要贡献在于:

(1)已有工作侧重于开源软件自身的演化规律,而我们收集整个 SF.net 开源社区为期两年的开发者和项目信息,以每两个月为一个时间段,分析社区中两者的整体变化和在不同

项目属性下的演化特性。

(2)对比 4 种可能的合作关系,首次分析开发者间合作关系的优先顺序,并分析新加入成员的偏好合作行为与开发者中心性、开发项目数和角色之间的关系。

## 2 开源软件社区

文章整体结构框架描述如图 1 所示,首先是实验数据的获取,随后对开源社区的项目与开发者的增长分布情况进行演化分析,再以开发者为主体,构建开发者网络,研究网络中每个时间段新成员加入与新合作构建的特性,针对新开发者与已有开发者间的合作,应用 3 类指标探索社区新开发者的偏好合作行为。

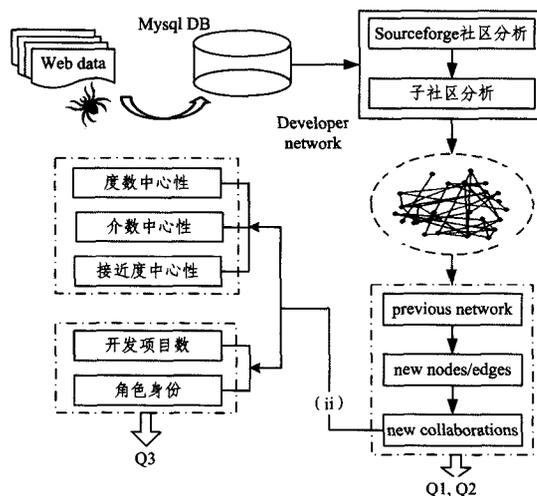


图 1 结构框架

本文以 SF.net 社区为实验对象,所使用的部分数据由 Flossmole<sup>[22]</sup>提供,时间跨度从 2007 年 4 月到 2009 年 6 月,以两个月为时间段。为了更好地了解社区,本节对社区开发者与项目的时间演化特性进行分析。图 2 显示了 2007、2008 两年社区发展呈线性增长,虽然 2008 年年底有一个小的下降趋势,但很快得到回升,见证了开源社区的发展。

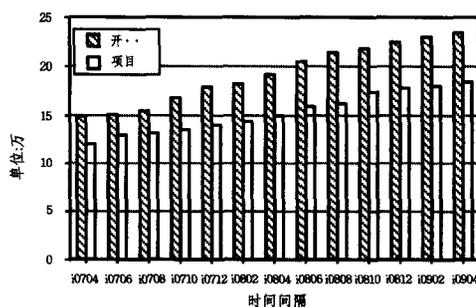


图 2 开发者与项目的增长趋势

本文的重点是探究开发者之间的合作交互关系,通过对数据的观察,每个项目包括如主题、目标受众、开发语言、数据库环境等属性,考虑到主题表示项目的方向以及目标受众表示项目的开发对象,我们选取目标受众与项目主题两个属性作为社区划分标准,图 3、图 4 分别为两种属性划分下前 8 个子社区的开发者数分布情况,其中项目主题包括:软件开发、游戏/娱乐、因特网、交互、办公、科学/工程、系统、多媒体;目标受众包括:开发者、终端用户、系统管理员、信息技术、科研、教育、其他、高级用户。从整体上看,开发者数呈一定的线性增长趋势,且 2007 年比 2008 年增长更快。

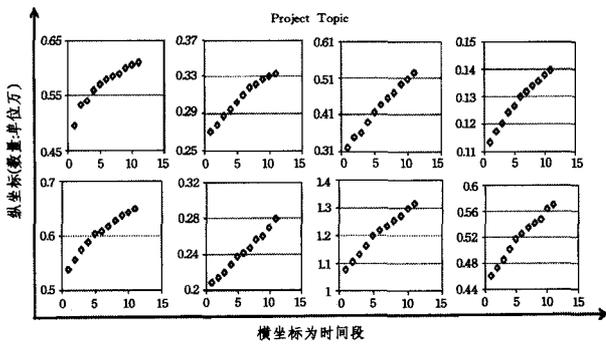


图3 Top-8项目主题开发者分布

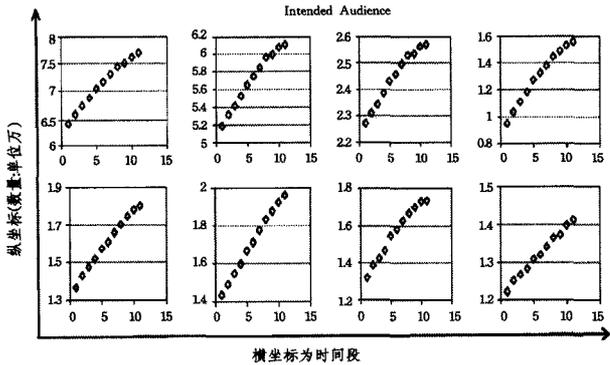


图4 Top-8目标受众开发者分布

在文献[23]中,我们对该社区的项目管理者与项目的分布做了分析,发现大部分管理者只负责比较少的项目。通过开发者与项目之间的度分布分析发现,77.1%的开发者只参与一个项目的开发,96%以上的开发者参与不足5个项目。由此可见,开源社区拥有庞大的人力资源,但社区资源的利用还不够充分。部分开发者没能及时找到自己感兴趣的项目和合适的合作对象,处于空闲状态。如果能够掌握他们的合作偏好,推荐可能的合作人,将有利于社区的可持续发展。

针对新成员在选择合作对象上是否存在偏好性以及新合作的类型差异与合作经验对再次合作的影响,本文提出3个研究问题:

Q1 新开发者:新加入的开发者是优先与其他新成员合作,还是与社区内已有的开发者合作?

Q2 新合作:社区开发者之间的4种新合作情况,(i)新开发者之间;(ii)新开发者与已有开发者之间;(iii)事先没有合作经验的已有开发者之间;(iv)已有开发者之间再次合作。关于开发者间的合作,图5给出了一个简单例子,然而,哪一种合作在社区间更为普遍?

Q3 角色指标:新开发者是否倾向于与更重要的已有开发者建立合作关系?

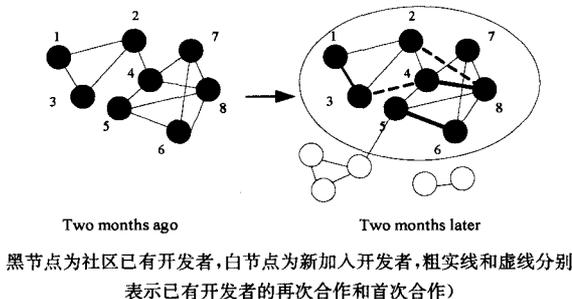


图5 开发者的合作关系

### 3 研究方法

#### 3.1 实验指标

本文从社会网络的研究视角出发,采用度数中心性、介数中心性和接近中心性3个典型指标分析其对开源社区开发者合作行为的影响。另外,我们还把开发者的开发项目数与他们在之前参与项目中的职位情况作为衡量一个开发者在社区开发经验和能力的度量指标。

(1)度数中心性(Degree Centrality),指与该节点直接相连的节点数,在网络中跟其他节点具有广泛连接的节点被视为网络的重要节点。本文用其表示一个开发者与其他开发者的直接交互能力,节点*i*的度数中心性表示为:

$$DC(i) = k(i) \quad (1)$$

其中, $k(i)$ 表示网络中与节点*i*直接相连的节点数。

(2)介数中心性(Betweenness Centrality),节点介数定义为网络中所有最短路径中经过该节点的路径的数目占最短路径总数的比例,是一种能够体现全局中心度的方法<sup>[24]</sup>,表示开发者通过中介路径发掘信息,完成任务的能力。具有较高介数中心性的节点,可视为信息传递的“信息桥”。在开源软件开发团队中,若存在较高介数中心性的成员,意味着团队的信息传递较为通畅。用 $g_i^{(s,t)}$ 表示节点*i*对*s*和*t*最短路径中经过*i*节点的数, $n^{(s,t)}$ 表示*s*和*t*之间存在的最短路径总数,则节点*i*的介数中心性可表示为:

$$BC(i) = \frac{2 * \sum_{s < t} g_i^{(s,t)} / n^{(s,t)}}{n(n-1)} \quad (2)$$

(3)接近中心性(Closeness Centrality),关注一个开发者与网络其他开发者的接近程度,表示开发者网络中节点快速获取所需信息的能力。在开源软件开发过程中,一个开发者接近中心性相对较高,表明他与其他人都比较邻近,获取信息更快,沟通成本越低。用 $d(i,j)$ 表示节点的距离,则节点*i*接近中心性可表示为:

$$CC(i) = \sum_{j=1}^N \frac{1}{d(i,j)} \quad (3)$$

(4)开发项目数(Project Numbers),一个参与了多个项目开发的开发者在社区相对比较活跃,这一类开发者在不同子社区间有着重要的作用。在开发者-项目构成的二分网络中,开发项目数为开发者的度,可表示为:

$$PN(i) = D_{BN}(i) \quad (4)$$

(5)开发者角色身份(Developer Position),在一个开源项目中,有项目管理者、开发者、分析设计师、发布者等身份。根据开发者在之前参与项目中所担当的职位来衡量其在社区的角色地位。在SF.net社区中,开发者的身份共21种,本文涉及20种,如图6所示。

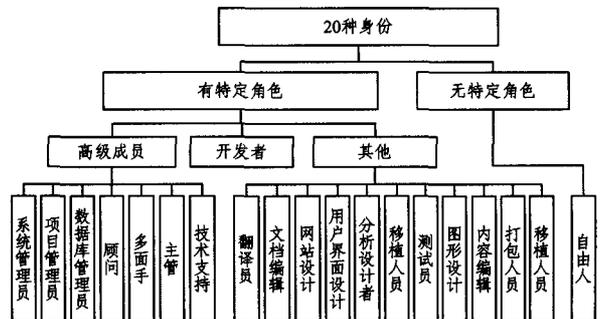


图6 开发者角色身份信息

### 3.2 开发者网络

假设两个开发者从事同一个项目的开发,则被视为存在一条合作连边。项目与开发者之间的隶属关系可用二部图<sup>[25]</sup>表示为 $G=(V,U,E)$ ,其中 $V$ 代表开发者的节点集, $U$ 代表软件项目节点集,且 $V,U$ 满足同类节点间没有连边, $E$ 是边集,如图7所示。

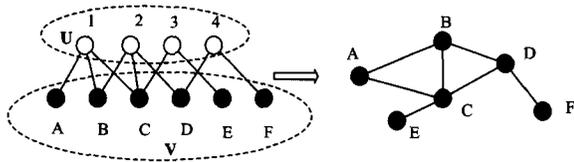


图7 一个简单的二部图(右边是一个对应的开发者网络)

### 3.3 偏好合作分析

Barabasi 和 Albert<sup>[26]</sup>提出复杂网络的无标度特性,并把这种特性总结为增长和偏好连接机制:随着时间的推移,网络新节点不断加入,新加入的节点优先与网络中度数较大的节点连接,表现出“富者更富”的现象。开源社区开发者合作网络作为一类复杂的社会网络,社区成员和合作随时间不断地变化,分析网络中新加入开发者的偏好合作行为有利于了解社区开发者网络结构的动态变化与合作趋势的走向。

表1 Topic="Desktop Environment"下节点和边的统计信息

Time interval	Total nodes	(iv)		(iii)		(ii)			(i)		sum			
		E	N	E	N	E	new_N	old_N	E	N	new_N	new_E	E/N	
i0704	5363													
i0706	5544	4	6	6	9	638(81%)	63(35%)	344	140(18%)	140(77%)	181	788	4.35	
i0708	5780	10	12	10	14	562(69%)	55(23%)	329	238(29%)	209(89%)	236	820	3.47	
i0710	5951	9	15	21	25	267(55%)	47(27%)	202	191(39%)	150(85%)	171	488	2.85	
i0712	6151	7	10	26	29	315(58%)	64(32%)	258	196(36%)	170(85%)	200	544	2.72	
i0802	6253	16	17	27	31	373(79%)	48(47%)	234	55(12%)	77(75%)	102	471	4.62	
i0804	6405	10	12	13	14	598(79%)	74(49%)	329	136(18%)	130(86%)	152	757	4.98	
i0806	6565	17	20	20	24	770(70%)	102(64%)	385	299(27%)	129(81%)	160	1106	6.91	
i0808	6700	9	13	27	30	694(84%)	47(35%)	344	100(12%)	114(84%)	135	830	6.15	
i0810	6855	16	24	80	58	500(59%)	60(32%)	321	253(30%)	161(87%)	185	849	4.59	
i0812	7065	17	18	84	59	550(56%)	78(43%)	290	334(34%)	160(89%)	180	985	5.47	
Avg(* )	6242	11.5	29.3	31.4	29.3	527(68%)	63.8(39%)	303.6	194(26%)	144(84%)	171	764	4.61	

注:i,ii,iii,iv 分别代表问题 Q2 中的 4 类合作,i\* 表示对应时间段,E 表示新建立的合作,N 为社区新加入的开发者

表1给出了每个时间段的节点总数、边数(合作数)和前两种合作关系中新节点与边数的比值。第二种合作情况下的节点分为新节点和老节点,最后三列为每个时间段新节点、新合作和每个新开发者的平均合作数。最后一行为整个表中对应列的均值,用于衡量子社区新开发者的整体演化行为。从表1可以发现,新成员数平均每两个月增加171,新建立的合作平均为764,平均84%的新开发者间存在合作关系,相比之下,只有39%的新开发者选择与已有开发者建立合作,前者是后者的2倍以上。另外,虽然与已有开发者合作的新成员占有新成员比例不多,但68%以上的新合作建立在新开发者与已有开发者之间,2008年8月尤为明显。新成员之间建立起来的合作总数只占有合作中的1/4。

同时,一些新开发者既与其他新开发者存在合作,又与社区一些已有开发者建立了合作。整体上,更多的开发者优先选择与其他新开发者建立合作,在建立的合作数量上,少量的新开发者却与已有开发者建立了更多的合作,换言之,新开发者间的合作不如新开发者与已有开发者间合作频繁,因此问题 Q1 得到解决。对于 Q2,表1显示4种情况下的合作从多

文章引入开发者的中心性、开发项目数和在已有项目中所拥有的角色身份度量指标来表示其在开发者网络中的重要性,根据新加入的开发者在合作对象上的选择,探索开源社区开发者合作网络的增长特性与偏好行为。

## 4 实验分析

### 4.1 预备分析

根据项目主题属性将社区划分为272个子社区,由于一些子社区包括的项目和开发者数相对较少,还有一些子社区增长很不明显,最终选取9个数据信息比较齐全、数据量相对较大的子社区作为实验分析对象,包括软件开发(Software Development)、游戏/娱乐(Games/Entertainment)、因特网(Internet)、交互(Communication)、办公/业务(Office/Business)、科学/工程(Science/Engineering)、系统(System)、多媒体(Multimedia)和桌面环境(Desktop Environment),统计结果如表1所列。尽管9个子社区中因特网和交互两个主题下的(iii)和(iv)两种合作关系的趋势与其他7个在某些时间段里稍有不同,但其他两种合作关系表现出来的现象完全一致,因此,表中仅给出了最后一个主题为桌面环境的社区节点与边的统计信息。

到少依次为 ii、i、iii、iv,其中值得注意的是,已有开发者当中,更多之前没有合作经验的开发者不断地选择彼此之间合作,并且这种趋势从开始的6快速增长到84。

对于 Q3,计算每个时间段子社区开发者网络中已有开发者的3类中心性指标。图8给出了被新开发者选中合作的已有开发者的3个中心性指标均值与开发者网络的整体指标均值的比值。结果表明,新开发者更倾向于选择与中心性值比较大的已有开发者合作,且度数中心性与度中心性更明显。

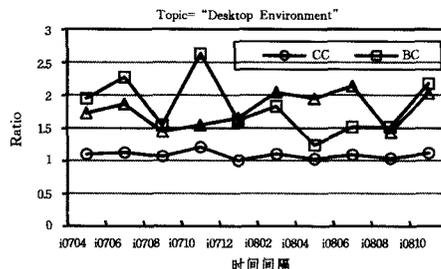


图8 被选中的已有开发者的中心性指标均值比例

对于开发者的开发项目数与在社区的身份分析,表2给

出了被选中合作的已有开发者中特定角色(SR)、高级成员(AM)、开发者(DEV)及其他成员(OTHERS)所占的比例,4种身份的比值中,至少86%以上被选中中的开发者在开发过程中都有过特定的身份,而平均值达到93.6%。其中具有高级成员身份的开发者占23.4%的比重,纯粹为开发者身份的占一半以上,无特定身份的占6.4%,其他为16.5%。不难发现,新成员会优先选择与社区内具有特定身份的已有开发者合作,很少选择与自由人身份的开发者合作。表中最后还给出了其他8类主题下的比率均值,整体表现一致:具有特定身份的比值范围为[83.4%,96%],高级成员为[14%,32.6%],开发者为[41.4%,65.1%],没有特定身份和其他的比值都比较少。在每个时间段的均值AVG(PN(\*))在(1.8,4)范围内,相对整个社区所有开发者参与项目数均值1.4而言,可以理解为被选中中的开发者中,大部分都是参与项目数相对比较多的开发者。

表2 Topic="Desktop Environment"开发者身份与平均项目数

	SR	AM	DEV	NSR	OTHERS	AVG(PN(*))
i0704	0.953	0.213	0.62	0.047	0.12	2.193
i0706	0.963	0.183	0.587	0.037	0.193	2.083
i0708	0.941	0.228	0.522	0.059	0.191	2.059
i0710	0.868	0.186	0.667	0.132	0.016	2.163
i0712	0.962	0.181	0.543	0.038	0.238	1.819
i0802	0.879	0.333	0.326	0.121	0.22	3.277
i0804	0.908	0.339	0.431	0.092	0.138	3.917
i0806	0.958	0.172	0.622	0.042	0.164	1.962
i0808	0.933	0.353	0.48	0.067	0.1	2.947
i0810	0.989	0.151	0.57	0.011	0.268	1.972
AVG(*)	0.936	0.234	0.537	0.064	0.165	2.439
Internet	0.896	0.19	0.565	0.104	0.141	1.96
Sci/Eng	0.834	0.182	0.556	0.166	0.097	1.857
Office/B	0.854	0.188	0.542	0.146	0.123	1.920
System	0.960	0.326	0.414	0.040	0.220	2.368
Game/En	0.891	0.201	0.558	0.109	0.131	2.332
Soft Dev	0.891	0.209	0.626	0.109	0.056	2.287
Commu	0.892	0.207	0.556	0.108	0.129	1.843
Mult Med	0.876	0.14	0.651	0.124	0.086	1.806

本节首先应用社会网络分析中3个典型的中心性指标来衡量开发者网络中被选中合作的已有开发者的不同重要性,再利用他们在社区的开发项目数与在之前参与项目中身份作为评价他们开发经验和能力的指标,通过开发者中心性、开发项目数和身份角色的分析,验证已有开发者的重要性对新开发者偏好合作行为的影响。

#### 4.2 验证分析

上一节结果回答了3个研究问题,为进一步验证结果的一般性,选用社区的目标受众属性进行子社区划分,对结果进行验证。目标受众共有18种,相对项目主题少了很多,从实验分析角度考虑,我们选取开发者(Developer)、终端用户(End Users/Desktop)、系统管理员(System Administrators)、高级终端用户(Advanced End Users)、信息技术(Information Technology)、科研(Science/Research)、教育(Education)和其他(Other)8种进行分析。采用与4.1节相同的步骤分析每个时间段的开发者合作行为。表3给出了8种目标受众在每个时间段下4种合作类型的平均值。在整个过程中,每种目标受众的子社区均表现出新开发者与已有开发者建立的合作比与其他新开发者建立的多,且已有开发者更多地是选择与其他没有合作过的已有开发者建立合作,这与开发者由于之

前的合作经验而再次合作并不违背,只是在SF.net社区中只有少部分的开发者合作比较频繁,大部分开发者的实际合作范围都比较狭小,而开源虚拟社区可供开发者合作的空间又很大,因而一些开发者有足够的机会和空间与其他没有合作过的开发者建立合作。

表3 不同目标受众的4类合作均值统计信息

avg (*)	(iv)		(iii)		(ii)		(i)		
	edges	nodes	edges	nodes	edges	new_nodes	old_nodes	edges	nodes
auth1	854	563	2705	2219	20183	1207	7576	3526	1580
auth2	241	227	1204	1055	11855	767	4641	2483	1198
auth3	242	96	204	180	4462	302	1792	1086	451
auth4	102	70	103	98	3394	360	1420	1721	709
auth5	92	53	283	215	7564	325	2273	1477	526
auth6	153	134	366	328	7180	413	2774	1162	555
auth7	142	71	275	244	7512	309	2108	1759	518
auth8	12	13	83	86	4026	136	1304	551	296

注:表中的authx分别对应于开发者、终端用户、系统管理员、高级终端用户、信息技术、科研、教育和其他目标受众属性

在实验结果中,需要注意的是新开发者的偏好行为与已有开发者的中心性关系,图9给出了8种目标受众在每个时间段的3种中心性指标的比例均值,可以发现,其与之前结果有所不同,图中度数中心性比介数中心性要更明显。在接近度中心性中,开发者、终端用户、高级终端用户的比值小于1。结果验证了接近度中心性效果最不明显,介数中心性与度数中心性效果更好。然而,正如弗里曼所说,哪个指标最合适需要依赖于研究问题的背景。

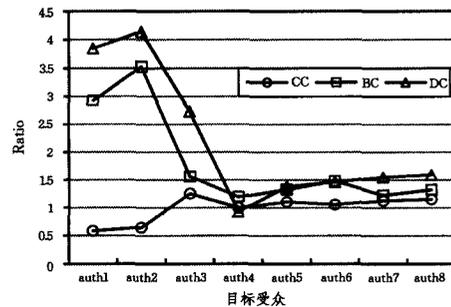


图9 被选中中的已有开发者的中心性指标均值比例

对于已有开发者的开发项目数和所拥有的身份角色与新开发者的偏好合作行为关系的验证,表4显示在8种目标受众属性下,具有特定角色、项目管理员、开发者和其他开发者身份的比例都相对稳定,并且与项目主题表现结果非常一致,开发者所占的比例依旧一半以上,项目管理员在17%—20%之间,已有开发者的开发项目数也在1.8以上,比整体的均值大。验证结果表明,新开发者确实具有优先选择开发项目数多且在社区拥有过特定角色身份的已有开发者合作。

表4 不同目标受众的各指标均值

	SR	PM	DEV	NSR	OTHERS	AVG (PN(*))
Auth1_AVG(*)	0.866	0.182	0.596	0.134	0.087	1.939
Auth2_AVG(*)	0.891	0.187	0.565	0.109	0.139	1.962
Auth3_AVG(*)	0.884	0.199	0.566	0.116	0.119	2.005
Auth4_AVG(*)	0.861	0.202	0.553	0.139	0.106	1.848
Auth5_AVG(*)	0.876	0.189	0.577	0.124	0.110	1.834
Auth6_AVG(*)	0.863	0.174	0.619	0.137	0.070	1.848
Auth7_AVG(*)	0.876	0.170	0.610	0.124	0.097	1.865
Auth8_AVG(*)	0.912	0.192	0.610	0.088	0.111	1.905

**结束语** 本文以 SF.net 开源社区为研究对象,首先对社区项目与开发者的增长情况进行了统计,发现整体上呈线性增长趋势。随后重点分析了开源社区开发者的群体合作行为,统计每个时间段新增开发者与合作数,并将新建合作细分为 4 种情况,探析 4 种情况下新建合作数情况,判断社区开发者优先选择合作的顺序。最后,针对新开发者与社区已有开发者之间的合作,进一步分析新开发者的合作偏好与已有开发者的中心性、开发项目数和在项目角色身份的关系。研究发现,新成员存在优先选择与介数中心性或度中心性大、开发项目数多且拥有过特定角色身份的已有开发者合作。结论与文献[27]的科学家合著网络中科学家的偏好合作一致,从而进一步验证了偏好连接模型在社会网络中的适用性。

## 参 考 文 献

- [1] Hinds D, Lee R M. Social Network Structure as a Critical Success Condition for Virtual Communities[C]//Proceedings of the 41st Annual, Hawaii International Conference on System Science. IEEE, 2008; 323
- [2] 吴江, 胡斌, 张金隆. 开源软件开发者和源代码协调性的网络分析[J]. 科研管理, 2011, 8(32): 133-141
- [3] Sen R, Singh S S, Borle S. Open source software success: Measures and analysis[J]. Decision Support Systems, 2012, 52(2): 364-372
- [4] Whitehead J, Mistrík I, Grundy J, et al. Collaborative Software Engineering: Concepts and Techniques[M]//Collaborative Software Engineering, 2010; 1-30
- [5] Hossain L, Zhou D. Measuring OSS quality through centrality [C]//Proceedings of the 2008 International Workshop on Cooperative and Human Aspects of Software. 2008; 65-68
- [6] Wolf T, Schroter A, Damian D, et al. Predicting build failures using social network analysis on developer communication[C]//Proceedings of the 31st International Conference on Software Engineering. ACM, 2009; 1-11
- [7] Hinds D. Social Network Structure as a Critical Success Condition for Open Source Software Project Communities[D]. Florida International University, 2008
- [8] Hinds D, Lee R M. Social network structure as a critical success condition for virtual communities[C]//Proceedings of the 41st Annual Hawaii International conference on System Sciences. Washington DC, USA, 2008; 323
- [9] 毛清华, 高杨. 基于社会网络中心性分析的虚拟团队知识共享促进策略[J]. 情报杂志, 2010, 29(10): 130-133
- [10] Toral S L, Martínez-Torres M R, Barrero F. Analysis of virtual communities supporting OSS projects using social network analysis[J]. Information and Software Technology, 2010, 52(3): 296-303
- [11] Hossain L, Zhu D. Social networks and coordination performance of distributed software development teams[J]. The Journal of High Technology Management Research, 2009, 20(1): 52-61
- [12] Datta S, Kaulgud V, Sharma V S. A Social Network Based Study of Software Team Dynamics[C]//ISEC. 2010; 33-41
- [13] Datta S, Sindhgatta R, Sengupta B. Evolution of developer collaboration on the jazz platform a study of a large scale agile project[C]//ISEC. 2011; 21-30
- [14] Pinzger M, Nagappan N, Murphy B. Can developer-module networks predict failures? [C]//Proceedings of the 16th ACM SIGSOFT International Symposium on Foundations of software engineering. ACM, 2008; 2-12
- [15] Huang S K, Liu K M. Mining version histories to verify the learning process of legitimate peripheral participants[C]//Proceedings of 2005 International Workshop on Mining Software Repositories. New York, USA, 2005; 1-5
- [16] Lave J, Wenger E. Situated Learning: Legitimate Peripheral Participation[M]. Cambridge: Cambridge University Press, 1991
- [17] Crowston K, Howison J. Assessing the health of open source communities[J]. Computer, 2006, 39(5): 89-91
- [18] Sureka A, Goyal A, Rastogi A. Using Social Network Analysis for Mining Collaboration Data in a Defect Tracking System for Risk and Vulnerability Analysis[C]//Proceeding of 4th India Software Engineering Conferene. ACM, 2011; 195-204
- [19] Hahn J, Moon J Y, Zhang C. Emergence of New Project Teams from Open Source Software Developer Networks: Impact of Prior Collaboration Ties[J]. Information Systems Research, 2008, 19(3): 369-391
- [20] Cavrak, Orlic M, Crnkovic I. Collaboration patterns in distributed software development projects[C]//ICSE 2012, 2012; 1235-1244
- [21] Bosnic, Cavrak I, Orlic M, et al. Student Motivation in Distributed Software Development Projects[C]//Proceedings of Collaborative Teaching of Globally Distributed Software Development; Community Building Workshop (CTGDSD 2011). 2011; 31-35
- [22] <http://flossmole.org/>
- [23] 何鹏, 李兵, 潘伟丰. 基于管理者合作网络的开源软件开发社区中心性分析[J]. 小型微型计算机系统, 2013, 34(1): 54-57
- [24] Bosnic', Cavrak I, Z'agar M, et al. "Customers' Role in Teaching Distributed Software Development[C]//IEEE Conference on Software Engineering Education and Training. 2010; 73-80
- [25] Conaldi G, Lomi A, Tonellato M. Dynamic Models of Affiliation and the Network Structure of Problem Solving in an Open Source Software Project[J]. Organizational Research Methods, Jan. 2012, 15(3): 385-412
- [26] Barabasi A L, Albert R. Emergence of scaling in random networks[J]. Science, 1999, 286: 509-511
- [27] Abbasi, Hossain L, Leydesdorff L. Betweenness centrality as a driver of preferential attachment in the evolution of research collaboration networks[J]. Journal of Informetrics, 2012, 6(3): 403-412