Research on the Influence Mechanism of Open Source Projects using Python and Stata

Xincheng Wu[1], Yaqi Wang[2] *

[1, 2]School of Business, Sichuan University, Chengdu, 610065, China

[1]2021225020104@stu.scu.edu.cn
[2] *wangyaqi0221@stu.scu.edu.cn

Abstract. As a successful practice of open innovation (OI) in software field, open source community provides a new and sufficient impetus for the development of soft-ware industry with increasing technology intensity. Community influence of open source projects is not only an important indicator to measure project performance, but also the spread effectiveness of the project in open source projects. This paper holds that open source projects are excellent examples of projects under the topic of OI. Through the analysis of community influence factors of open source communities discussed, the relationship between OI mechanism at the company level and project level is expounded. The Python crawler technology was used to collect the open source project data of 25 well-known Chinese enterprises from GitHub, and the data were processed for the mixed cross section data of the studio, and the double fixed effect regression analysis was performed on the data using Stata16.

Keywords: Open innovation, GitHub, Open source community

1 Introduction

As one of the important indicators of sustainable development in 2030 proposed by the United Nations [7], open innovation (OI) has become a mainstream phenomenon in various commercial fields and occupies an increasingly important position. In 2003, Professor Henry Chesbrough put forward the concept of OI for the first time [4], many scholars put forward ideas on the mechanism exploration of OI at the enterprise level. It is proved that project complexity. resource investment, cooperation mode and R&D intensity have an impact on enterprise performance, and the exploration of OI mechanism based on project level needs to be further improved [2].

The project performance output mechanism of open source community is one of the OI mechanisms at the project level, but the project performance generation mechanism of open source community still cannot be well explained by the existing theoretical logic. In a review of the relevant literature, scholars from the perspective of resource input [9] studied the output mechanism of com-munity performance, from social network theory [11] research project community communication mechanism and from intellectual property theory [1] studied the role of open source license agreement on

V. Escudero et al. (Eds.): BADS 2022, AHCS 11, pp. 1355–1362, 2023.
https://doi.org/10.2991/978-94-6463-102-9_140

project property rights and open source culture, and studied the participation of the community itself and individuals and enterprises.

To sum up, in order to make up for the lack of mechanism exploration of open innovation at the project level, the important mechanism of the influence of open source community projects is revealed. This paper analyzes the influence generation mechanism of open source projects based on the GitHub Open source community project that 25 well-known Chinese Internet enterprises participated in from 2010 to 2020. The conclusion of the appeal can provide the basis and reference for the further development of the open source community in China.

2 Research hypothesis

2.1 Resource Investment

The main body of community influence is community members. According to the open source project discussed in this paper, community members can be divided into project companies, project participants, and project users, which have different demands. In general, participation in open source community is more influenced by economic factors, while non-corporate project participants are more influenced by non-utilitarian factors such as social interaction, while open source project users are more concerned about the specific use value of the project. According to the input-output model, the final output of the open source community is a line of code that can be executed and understood as a result of the mental effort of the members of the open source project. In this paper, the influence of open source projects is regarded as one of the invisible outputs invested by the mental work of the open source community, so the input of resources has an impact on the influence, even the most important impact. According to the input-output model, re-source input plays a non-negligible role in community influence output, which is an important mechanism of influence source. From the perspective of enterprises participating in open source projects, such open resource input is included in the R&D expenditure of enterprises, and such research expenditure is generally considered to be related to the innovation performance of enterprises [6].

Enterprise investment project members (usually employees), especially the open source project director is not only in the project specific development produces influence on improving the quality of project, and the community is a mesh structure, the relationship between knowledge according to the network transmission, by project members of the community individual network spread to other members. Stam proposed and verified that there is an inverted U-shaped relationship between community participation and enterprise innovation performance [10].

Therefore, hypothesis 1 is proposed: H1: Project resource investment is positively correlated with the influence of open source projects.

2.2 Property Rights Protection

Communities are zbased on a culture in which recognition of property rights and incentives is seen as antithetical to the over-all ethos and motivation of contributors [3]. Based on the goal of improving corporate performance, the company will have a higher awareness of property rights protection, but the concern about property rights is a deviation from the project community culture, which will reduce the cultural identity of community members. The open source license agreement is the most direct means for the project company to adjust the community culture, which directly reflects the company's concern about the property rights of the project. Companies need to strike a balance between project ownership and open source culture. After a long time of development, open source license agreement has formed a set of relatively mature system, but the complex system increases the difficulty of the selection of open source license agreement. Kapitsaki and Charalambous adopted the method of user similarity and project similarity to implement open source license recommendation [5]. From the point of view of the community. Intellectual property is one of the research hotspots in the field of open innovation. The choice of OI mode of enterprise projects determines the ownership rights of enterprises. On the one hand, it represents the openness degree of enterprises, and on the other hand, it affects the performance of enterprises. Under the conditions of technological opportunities and technological slack at different research levels, how OI patterns of firms are deter-mined by their R&D intensity is illustrated in this literature[8]. As non-originator development participants driven by non-utilitarian factors, they are often confused about the subtle differences of various open source protocols. Because they are not the core needs of their participating communities, community users often ignore the content of open source license agreements, causing the risk of infringement in reality and exposing users to legal risks.

Therefore, hypothesis 2 is proposed: H2: The awareness of project property rights protection is negatively correlated with the community influence of open source projects.

3 Data and method

3.1 Data

In this paper, a total of 1876 open source project data of 25 well-known Chinese companies deeply involved in GitHub from 2010 to 2020 were collected, forming a mixed cross-sectional data sample. In the first step, the number of activities of all warehouses of well-known Chinese companies on GitHub was counted, and all the open source projects of the top 25 companies with the number of activities in 2020 were used as the basis for the experimental samples. In the second step, the research samples are collected by crawling through the GitHub Search API. By December 2020, Python crawler was used to crawl all the open source project code of 25 target companies, and a total of 1,936 pieces of data were crawled. The third step is to clean the crawling projects, excluding the data of tutorial sharing and empty projects without code amount. After cleaning, 1876 pieces of valid data are obtained.

3.2 Method

An open source project data contained 13 fields as shown in Table 1. In this paper, Stars, Forks and Subscripts were processed as OSCI, a representation variable of project community influence. Among the explanatory variables, the number of people involved in the project is taken as the labor force of the open source project into DLI; Use the project's open source license agreement as an indicator of the company's concern about property rights. Among the control variables, three types are used: whether the project company is listed (LC), whether the project has a display Page (Page) and project Size (Size).

Some scholars have reduced the dimension of seven indicators, namely watch, star, fork, commit, contributors, total submitted discussion topics and total PR quantity, to obtain a comprehensive quantitative indicator of project success[11], which takes more into account the driving factors of individuals as project initiators. The research object of this paper is community influence, and more attention is paid to the indicators of community influence that will be affected by the above indicators. The indicators of Forks, Stars and Subscripts are based on the data generated by the users of the project, which are the direct reflection of community recognition of the project. Companies use the open source license agreement, the assignment of the rights and interests from the company more in the project to the entire community of users, the rights and interests of the company can get from the project is the less.

Table 1. Variable names, symbols, and meanings (Owner-drawing)

Variable types	Variable name	Variable symbol	Variable meaning
Explained variable	Open Source Community influence	OSCI	The influence of open source project in GitHub
Explanatory variables	Degree of labour input	DLI	The contributor in the project.
	Concern about property rights	CAPR	Whether the project uses an OSL
Control variables	Listed Company	LC	Whether the company is listed or not
	The Pages	Page	Whether the project has a display page in GitHub
	The size	Size	Project size (MB)
	The Year	Year	The year corresponding to the project
	The Language	Language	The programming language used by the project

4 Empirical Results and Analysis

4.1 Descriptive Statistics

Table 2 shows the descriptive statistical results for the main variables. As can be seen from the table, the minimum value of digital transformation degree of all sample enterprises is 0, the maximum value is 463, and the standard deviation is 18.915, indicating that the degree of digital transformation degree of the whole industry is very different. And the median degree of digital transformation is 1, indicating that most enterprises have only carried out a minimal degree of digital transformation. At the same time, the difference between the minimum and maximum values of supplier concentration Herfindahl index and customer concentration Herfindahl index is obvious. Descriptive statistics indicate that the differences between samples are large enough to facilitate empirical research.

Table 2. Descriptive statistics of variables (Owner-drawing)

The Variable	Number of Samples	The Average	The Standard Deviation	The Minimum	The Maximum	The Median
OSCI	1876	176.4	595.791	0.00	9190	6.40
DLI	1876	22.640	60.536	1.00	460.000	4.00
CAPR	1876	0.562	0.496	0.00	1.000	1.00
LC	1876	0.687	0.464	0.00	1.000	1.00
Page	1876	0.093	0.291	0.00	1.000	0.00
Size	1876	0.039	0.46	0.00	2.487	0.002

4.2 Mixed Cross-section Regression Analysis

Benchmark model in order to verify the hypothesis 2 set in a set of models, to test whether the model is moderate, clean and tidy for mixed samples of existing data to cross section data item by item, regression analysis, in order to im-prove the robustness of measuring results, in stata16 joined robust regression model after all parameters to remove effects in the model, the result of the following table:

Table 3. The result of mixed cross-section regression (Owner-drawing)

Variable	Modle1	Modle2	Modle3
Constant	-355.520*** (-6.07)	-349.151*** (-5.79)	-213.017*** (-5.79)
DEL	130.580*** (7.44)	132.955*** (7.52)	81.146*** (7.36)
OSL	—	-45.316* (-1.73)	-25.589* (-1.65)
LC	√	√	√
Page	√	√	√
Size	√	√	√
∑Year	√	√	√

∑Language	√	√	√
Adj-R2	0.133	0.134	0.136
N	1876	1876	1,876

*** represent significant at 10%, 5% and 1% levels respectively

According to model 1, the elasticity coefficient β1 of project participating members to the number of project collections is (130.580) positive and significant at the level of 1%, which verifies that the increase of project resource investment is positively related to the increase of project collections, and supports hypothesis 1. According to model 2, after adding OSL variable, the sign of original β1 remains unchanged, and the coefficient of β2 (-45.316) is negative and significant at 10% level, which proves the negative correlation between project complexity and project influence and supports hypothesis 2. According to model 3, after adjusting the Y value, Subscripts and Star weighted average were taken as the Y value, and the robustness test was conducted on the original model 1 and Model 2. The signs of coefficients β1 and β2 did not change, and the robustness test passed.

5 Research Conclusions and Implications

5.1 Research Conclusions

This paper explores the influence mechanism of open source community from the perspective of "open innovation". By using the mixed cross section data of 25 well-known Chinese Internet companies in GitHub from 2010 to 2020, it discusses the influence mechanism of Chinese enterprises' participation in GitHub community influence. The results are as follows: There is a positive relationship between project influence and project resource input; It has a negative relationship with project complexity, while company strength and the use of open source license agreement have a positive relationship.

5.2 Implications

The above conclusions firstly show that: The company needs to further enhance the importance of the project feasibility report. The accurate prediction of project complexity in the feasibility report is an important predictor of project success. Because of the open source community properties, open source code as mental work, its main is human input and output performance, and the company pay the cost is low, so to get a higher community resources in the community influence the will get higher and the influence of community higher and community will be further resource input, this is the project company to obtain a higher efficiency. In general, it is better to reduce the number of open source projects and adopt loose open source license agreements in the competition with open source projects of the same type.

5.3 Limitations

In this paper, the collection of data samples is slightly lacking, and innovative research materials are only collected from the perspective of the open source project in China on GitHub. In the future, data samples will be expanded in two directions: The first is to collect samples from Chinese and foreign companies on the same open source platform GitHub, and further reflect the differences in various indicators between Chinese and foreign companies through sample comparison, so as to find new growth points for improving the performance of Chinese open source projects. Second, the same company collects open source projects in multiple open source communities, especially the same open source project data collection, compares and obtains the characteristics of different open source platforms, analyzes and obtains new community development strategies.

References

1. Andersen-Gott M, Ghinea G, Bygstad B. (2012). Why do commercial companies contribute to open source software?. J. International journal of information management, 32(2): 106-117.
2. Bagherzadeh M, Markovic S, Bogers M. (2019). Managing open innovation: A project-level perspective. J. IEEE Transactions on Engineering Management, 68(1): 301-316.
3. Benkler Y. (2004). Sharing nicely: On shareable goods and the emergence of sharing as a modality of economic production. J. Yale Lj, 114: 273.
4. Chesbrough, Henry William. 2003. Open Innovation: The New Imperative for Creating and Profiting from Technology. Harvard Business Press.
5. Kapitsaki G M, Charalambous G. (2016). Find your open source license now! //2016 23rd Asia-Pacific Software Engineering Conference (APSEC). C. IEEE, 2016: 1-8.
6. Mina A, Bascavusoglu-Moreau E, Hughes A. (2014). "Open Service Innovation and the Firm's Search for External Knowledge.". J. Research Policy 43(5): 853–66.
7. Phelps C C. (2010). A longitudinal study of the influence of alliance network structure and composi-tion on firm exploratory innovation. J. Academy of management journal, 53(4): 890-913.
8. Rauter R, Globocnik D, Perl-Vorbach E. (2019). Open innovation and its effects on economic and sustainability innovation performance. J. Journal of Innovation & Knowledge, 4(4): 226-233.
9. Kim S, Lim Y T, Soltesz E G. (2004). Near-infrared fluorescent type II quantum dots for senti-nel lymph node mapping. J. Nature biotechnology, 22(1): 93-97.
10. Lee C Y, Wu H L, Pao H W. (2014). How does R&D intensity influence firm explorativeness? Evi-dence of R&D active firms in four advanced countries. J. Technovation, 34(10): 582-593.
11. Raghuram S, Hill N S, Gibbs J L, et al. Virtual work: Bridging research clusters[J]. Academy of Management Annals, 2019, 13(1): 308-341.