The Meso-level Structure of F/OSS Collaboration Network: Local Communities and Their Innovativeness

Guido Conaldi[1] and Francesco Rullani[2]

[1] Centre for Organisational Research, Univerisity of Lugano, Via Giuseppe Buffi 13,
CH-6900 Lugano, Switzerland. guido.conaldi@usi.ch

[2] Department of Innovation and Organizational Economics, Copenhagen Business
School, Kilevej 14A, 2000 Frederiksberg, Denmark. fr.ino@cbs.dk

Abstract. Social networks in Free/Open Source Software (F/OSS) have been usually analyzed at the level of the single project e.g., [6], or at the level of a whole ecology of projects, e.g., [33]. In this paper, we also investigate the social network generated by developers who collaborate to one or multiple F/OSS projects, but we focus on the less-studied meso-level structure emerging when applying to this network a community-detection technique. The network of 'communities' emerging from this analysis links sub-groups of densely connected developers, sub-groups that are smaller than the components of the network but larger than the teams working on single projects. Our results reveal the complexity of this meso-level structure, where several dense sub-groups of developers are connected by sparse collaboration among different sub-groups. We discuss the theoretical implications of our findings with reference to the wider literature on collaboration networks and potential for innovation. We argue that the observed empirical meso-structure in F/OSS collaboration network resembles that associated to the highest levels of innovativeness.

1 Introduction

The production of F/OSS is an organizational phenomenon characterized by a strong bottom-up tendency, which hinges upon the creation of social networks of developers freely interacting and collaborating [11, 13, 26]. Therefore, given the central role as productive infrastructures that social networks play in F/OSS projects, it is not surprising that they have been object of several studies. Indeed, various studies have investigated the social networks generated by developers who take part to F/OSS projects focusing both on the social structure internal to individual projects, e.g.,[6, 17], and on the larger network of collaborations linking the wide population of F/OSS projects through common developers, e.g., [33]. Particularly the entire ecology of F/OSS projects hosted on SourceForge has been object of study because of the representativeness of the repository for the entire population of F/OSS projects and thanks to the availability of rich public data [29].

In this paper we also investigate the social network formed by F/OSS developers who collaborate to one or multiple F/OSS projects. However we concentrate on a different level of analysis. Instead of focusing on ‘macro’ or ‘micro’ networks, we investigate the overall collaboration network by looking at the *meso-level* structure of collaboration. We apply a technique able to detect sub-groups of densely connected developers whose connectivity and size is in between that of the whole network and that of single projects. These sub-groups are commonly known as ‘communities’ in the methodological literature on graph theory and network analysis, and constitute the meso-level structure we will investigate in the following.

We connect our empirical findings to a wider literature on collaboration networks and potential for innovation. More specifically, the theorization on the role of both strongly cohesive teams and brokerage among separated groups [4, 9] can be translated into the configurations characterizing the network of communities revealed by our analyses. Therefore we discuss theoretically which configurations of collaboration networks have the potential to foster innovation and argue that the observed empirical F/OSS collaboration network resembles that associated to the highest levels of innovation.

The paper is structured as follows: in the second section we discuss more in detail the existing evidence on the social networks of F/OSS projects. In the third section we firstly describe the data used for reconstructing the F/OSS collaboration network. We then relate our preliminary descriptive findings on the overall collaboration network with the existing empirical evidence. Subsequently we introduce the method adopted to find communities in the collaboration network and present the results. Finally, in the fourth section we discuss our structural findings with reference to the potential for innovation of the overall F/OSS collaboration network which we investigate.

2 Background and Related Work

A rapidly growing body of research adopts a network approach for the understanding of the structural characteristics of the F/OSS phenomenon. Several contributions focus on the internal network structure of F/OSS projects, e.g., [6, 3, 18, 17], whereas other contributions reconstruct the networks of collaboration among different projects, e.g., [33].

Several characteristics manifested by the internal communication and collaboration networks of F/OSS projects are already known. Studies investigating the entire spectrum of F/OSS project demonstrate that a significant portion of them are formed by very few developers, or even only one [5], therefore introducing size as a dimension influencing the different structures F/OSS networks can assume. Furthermore, the configuration of F/OSS social networks has been demonstrated to change throughout the life of the projects. F/OSS projects indeed evolve over time. On the one hand they experience a high turnover rate among developers that is negatively correlated with the degree of involvement into the project [25, 12, 27]. On the other hand their overall structure reflects the different maturity and complexity a F/OSS project can assume over time. To this respect a progression pattern from single hub configurations to core-periphery structures is found in a longitudinal study of several F/OSS projects [17]. The variation over time of F/OSS network structures is also confirmed by a study that tracks the network centralization values of two F/OSS projects and shows how they varied over time [32].

Several studies present evidence of internal hierarchy in F/OSS internal communication and collaboration networks. It has been shown that well-established and large F/OSS communities manifest hierarchical structures [6, 17]. Sometimes the project founders assume a great authority on the entire development process [22, 28], whereas other equally relevant projects develop a complex meritocratic structure that relies on different status levels and voting procedures [21, 8].

Also the overall F/OSS phenomenon has been studied adopting a network perspective. The social network formed by all individuals connected through the F/OSS projects to which they co-collaborate has been shown to represent a prototypical complex evolving network [19]. Furthermore, this global F/OSS collaboration network has been analyzed by sub-dividing it in four subsets of different type of actors (project founders, core-developers, co-developers, and active users) and shown to be a self-organizing system that in all subsets obeys scale-free behaviors [33]. The same study also finds the same network to have a small average distance and a high clustering coefficient, therefore characterizing itself as a small world [31]. Finally, [33] discusses the role of individual actors in the overall F/OSS ecology and stresses the potential impact of co-developers and active users as direct connections among projects that could benefit from the fast sharing of information.

Here we adopt a global perspective on the overall F/OSS phenomenon similar to [33] and we reconstruct a similar F/OSS collaboration network. However, our focus is on the meso-level of the network. Consequently, we concentrate firstly on individuating dense communities of co-collaborating developers. The configuration of the network formed by these communities and their connections will then be at the core of our empirical analysis and theoretical discussion.

3 Methods and Analysis

3.1 Data

We use data describing the activities of 1,347,698 actors working on 170,706 F/OSS projects hosted on SourceForge [29] in September 2006. The period was chosen in accordance to the availability of information on individuals' emails, necessary for data cleaning. However, this should not be a problem, because we believe that the evolution of a self-organizing social network as that under analysis here follows general rules, such as growth and preferential attachment [2], that are very unlikely to change over a three-year period. SourceForge is likely not representative of the whole universe of F/OSS projects, as it is a company-owned platform and does not host some of the most famous project, such as Linux. However, it represents by far the largest repository of F/OSS projects worldwide, and it hosts extremely heterogenous projects, from very famous, active and large ones to very small or even 'dormant' ones. Thus, data relative to the activities it hosts have been already widely used in previous studies, e.g., [33, 16, 10]. In this study we will follow the same line of research.

Only projects labeled as 'active' have been retained in the dataset, as well as only 'active' actors registered with at least one project. This assured that we took into account all the projects and individuals relevant for our analysis, excluding only non-active projects or individuals who do not belong to the network. Different virtual identities belonging to the same individual have been aggregated through email address matching. This reduced the number of individuals to 161,983 and the number of projects to 115,112.

In order to reconstruct the F/OSS collaboration network we used individuals as nodes and affiliations to the same projects as ties, weighted by the number of projects in common. In other terms, we projected the weighted two-mode network formed by developers and projects that we originally collected into a one-mode network formed only by developers.

All analyses were performed using the igraph package [7] for the R environment.

3.2 The overall network structure of F/OSS collaboration

As a first step, we investigate the collaboration network similarly to what has been previously done by other studies on F/OSS, e.g., [33], and on other virtual networks, as for example Internet [1]. The main descriptive statistics for the generated network can be found in the first column of Table 1.

Table 1. Main characteristics of F/OSS collaboration network

Indicator	Whole network	Giant Component
Number of Nodes	161,983	58,481
Number of Ties	753,421	632,046
Global Clustering Coefficient	0.910	0.907
Average Path Length (APL)	7.105	7.106
APL for a comparable random network (Equal size and average number of ties)	7.804	4.612

We then isolate and analyze a giant component composed by 58,481 individuals, the second component spanning 201 nodes. The statistics relative to the degree distribution (mean: 21.62; standard deviation: 41.23; skewness: 5.12) signal the heterogeneity of the ego-networks of F/OSS project members (see Figure 1).

As the values reported in the second column of Table 1 show, the Global Clustering Coefficient (or, equivalently, Transitivity [30]) is extremely high and the Average Path Length is low (50% larger than that of a random graph, a proportion in the range of those reported by [1] for comparable cases). This shows that in the F/OSS world individuals not only gather locally in dense groups of neighboring collaborators, but also establish collaboration ties with members of local groups located elsewhere in the network, thus acting as 'brokers'. Thus, the network clearly resembles a 'small world' [31], a property detected also by [33]) in a similar context.

The mean (0.38), standard deviation (0.28) and skewness (0.79) of the distribution of Burt's measure of constraint [4] confirm this interpretation at the individual level. Indeed, Burt's constraint only focuses on the direct neighborhood of each F/OSS project in the network and captures the proportion of realized ties among its neighbors out of all the possible ones. The low average value of constraint found among the projects in the giant component, 0.38 with the constraint index varying in the range [0,2], confirms that projects tend to connect otherwise disconnected projects, therefore spanning so-called 'structural holes' in their neighborhood and thus acting, in Burt's terms,as brokers.

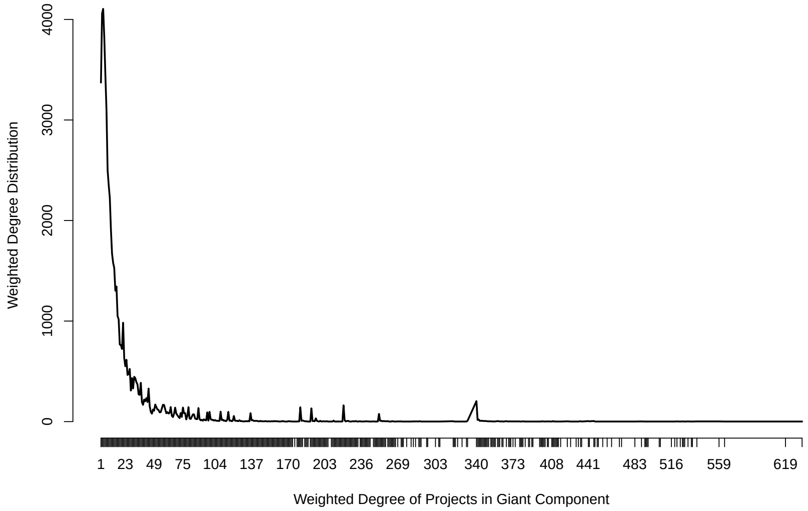

Fig. 1. Degree distribution of projects in the giant component of the collaboration network

3.3 Finding communities in the F/OSS collaboration network

The evidence at the global level of analysis just presented is consistent with what has been found in the field (e.g., [33]). We now focus on the meso-level of analysis and we test whether sub-groups of densely connected developers (i.e., communities) can be identified and whether they are connected through sparser collaborations.

In order to find communities in a network several algorithms are available, e.g., [23]. We apply the Walktrap algorithm [24]. This algorithm is based on the intuition that short random walks performed on a sparse network will tend to remain trapped in denser local areas of the network corresponding to communities. The Walktrap algorithm makes use of information on the weights of the ties in the network. This is a fundamental property for our purposes because of the wide variation existing in the F/OSS context concerning the number of collaborations in which different developers take part. A characteristic that is coherently reflected by the weighted degree distributions of our collaboration network.

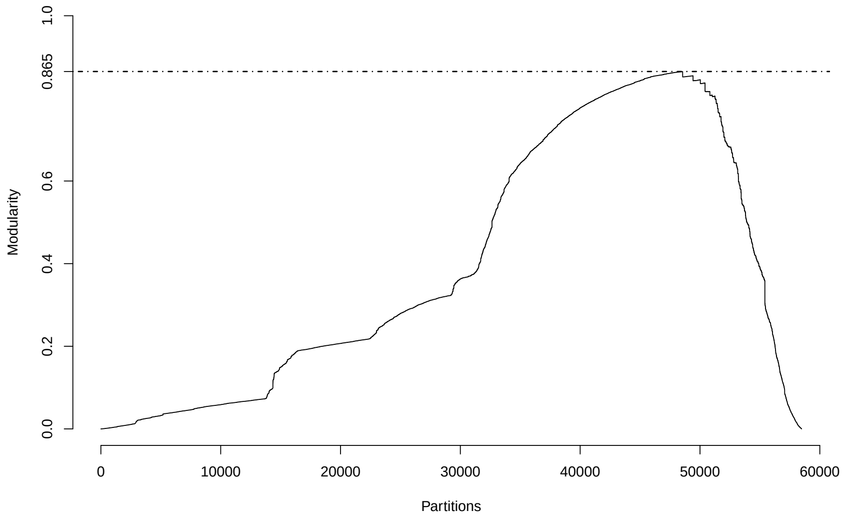

Fig. 2. Modularity values for all partitions in the giant component of the F/OSS collaboration network

The algorithm induces a sequence of partitions of the original network into communities. It starts with each node representing a community and ends with all nodes in one community. In order to find which partition best represents the community structure of the original network we adopt the most widely used criterion: the modularity Q index [23]. Q relies on the fraction of ties inside a community and the fraction of ties bound to that community: the best partition maximizes Q (that lies in the range [-1,1]) and therefore defines communities which are internally densely connected with only sparse connections among them.

When applying the Walktrap algorithm to the giant component of our F/OSS collaboration network we find that the best partition (with a high Q of 0.865, see Figure 2) individuate 9,931 communities, many of them extremely small, reaffirming the tendency to create dense, i.e., very close, local sub-groups of co-collaborators loosely interconnected. Nonetheless, the community-detection analysis shows the prominent role also of outward connections: 71% of individuals in communities with at least two members create ties beyond their community borders, and the average ratio between the number of their outward weighted ties and their total weighted degree is 0.28 (standard deviation: 0.19). This means that these many dense communities, whereas clearly distinguishable, are not almost isolated, but connected by a large number of brokers and through important ties. The network of the largest communities (see Figure 3)clearly shows the coexistence of brokerage and closure.

In Figure 3 each node is a sub-group of densely interconnected developers, i.e., a community, identified through the Walktrap algorithm. For simplicity, only communities having 25 or more members are depicted. The size of each node is proportional to the amount of collaborations in the same F/OSS projects among its members. The thickness and color shade of the ties linking the communities are proportional to the total number of F/OSS projects to which the members of the different communities co-collaborate. Figure 3 shows that high levels of collaborations can exist among communities of both comparable and different sizes.

Figure 4 shows the inner structure of the largest community, identified in brighter red with a yellow contour in Figure 3. The community is here magnified to show the connections among its 541 members (the round nodes). The thickness and color shade of the ties are here proportional to the total number of F/OSS projects to which two individuals co-collaborate, while the size of the nodes is proportional to the total number of collaborations each individual has with members of other communities. Figure 4 shows that inside a community both central and peripheral individuals manifest high levels of external collaboration.

Therefore, we can affirm that both the network of communities and the network of developers belonging to a same community assume a similar configuration that combines densely connected sub-groups with the presence structural holes [4], i.e. lack of ties, isolating the sub-groups and some relevant brokers spanning these borders thanks to inter-community co-collaborations. This combination, in line with the literature on similar phenomena, e.g., the Internet [1], places F/OSS in a sort of 'middle range' between full closure and extensive brokerage.

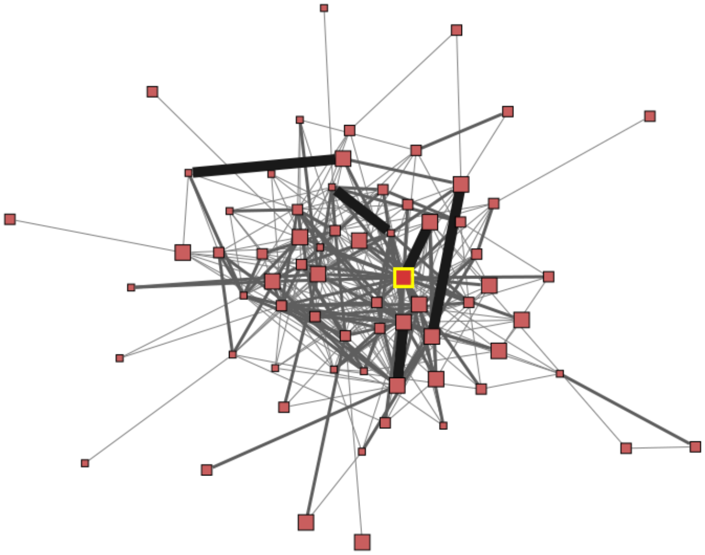

Fig. 3. Network of identified communities (with size > 25) in the giant component of the F/OSS collaboration network

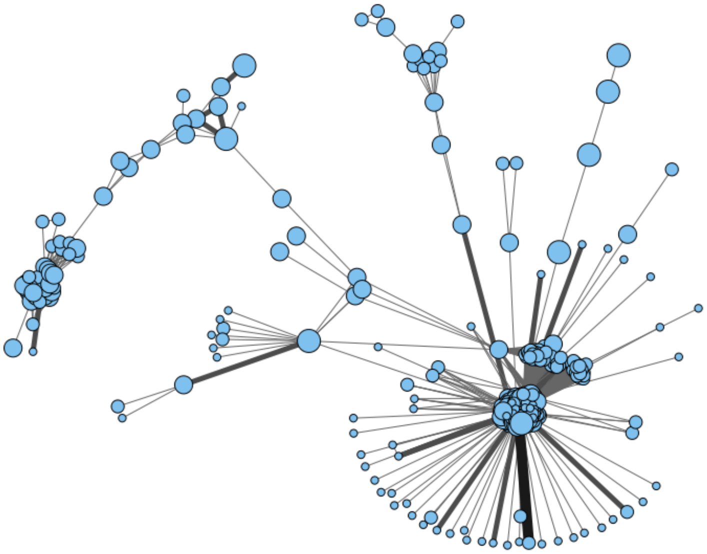

Fig. 4. Closer view of the largest identified community in the giant component of the F/OSS collaboration network

4 Discussion and Conclusions: Community Structure and Innovativeness

Among many others, one question that our analysis raises is whether the community structure that we just described hinders or fosters innovation. In a recent article Mayer-Schönberger [20] warns against the possibility of overestimating the innovation capabilities of open systems such as F/OSS or the Internet itself. If diversity is crucial for producing novelty, a *conditio sine qua non* for ideas recombination, extremely densely connected collaboration networks, such as he assumes the F/OSS collaboration network to be, have difficulties to reach high level of innovation because of their intrinsic tendency to establish many redundant connections, creating thus homogenization and group-thinking. In other words, Mayer-Schönberger puts forward a positive and monotonic relationship between innovative performance and the importance of structural holes [4] in the structure.

This point of view certifies the importance of brokerage for innovation, however other studies balance this perspective warning against possible excessive disconnection. Gilsing et al. notice that 'access to heterogeneous sources of knowledge [that creates] potential for novel combinations . . . requires an emphasis on diversity and disintegrated network structures. . . . On the other hand [actors] need to make sure that such novel knowledge, once accessed, is evaluated, and . . . absorbed. This process favours more homogenous network structures' [9, p. 1718].

Studies undertaken in related fields confirm this view. Laursen and Salter [14] found that as the number of different external knowledge sources a firm can use (e.g., universities, or clients) increases, the positive impact of one more source on firms' innovative performance significantly decreases, because absorption and combination become more difficult. Similarly, Lazear argues that multicultural teams benefit from members' diversity, but only if this is coupled with a certain degree of commonality because 'Without communication, there can be no gains from diversity'[15, p. 20].

Thus, according to the full spectrum of studies just introduced, the relationship between the importance of structural holes and innovative performance should be rather described as an inverted U-shaped curve. This means that structures able to produce the highest innovation rate should be located somewhere in the middle in the continuum between full closure and extensive brokerage [4].

When exploring the social network generated by individuals collaborating in F/OSS projects on SourceForge we discovered a distinguishable meso-level network linking communities of closely co-collaborating individuals. Furthermore, we investigated also the collaboration network internal to the largest of these communities. We found that the structure of these networks both clearly resemble a mixture between densely connected local areas, where information flows pervasively and diversity is reduced, and a great number of structural holes that are spanned by several brokers. As the above-mentioned studies on the structure of innovativeness suggest, in such a structure the idiosyncratic knowledge created in one community can flow throughout the entire network thereby mixing with diverse knowledge, increasing the probability of generating novel recombinations.

This mix of closure and brokerage suggests that the F/OSS collaborative environment is not a fully connected network in which everybody is co-collaborating with everybody else, as [20] assumes. On the contrary, the F/OSS collaborative environment appears to possess the structural characteristics necessary to place itself in that middle-range area that the literature on innovation associates to the highest segments of the innovativeness curve.

Therefore, our results give a preliminary insight on a more complex relationship between the structural dimension of collaboration in the F/OSS world and innovativeness than the one prosed by [20]. Consequently they also call for a more in-depth research on the actual innovative performances achieved by different local areas of the overall F/OSS collaboration network in order to elaborate beyond a first description of the meso-level of community structure that represented the aim of this study.

References

1. Albert, R., Barabási, A.L.: Statistical mechanics of complex networks. Reviews of Modern Physics 74(1), 47–97 (2002)
2. Barabási, A.L.: Scale-free networks: a decade and beyond. Science (New York, N.Y.) 325(5939), 412–3 (2009)
3. Bird, C., Goey, A., Devanbu, P., Swaminathan, A., Hsu, G.: Open Borders? Immigration in Open Source Projects. Proceedings of the Fourth International Workshop on Mining Software Repositories (2007)
4. Burt, R.: Structural Holes: The Social Structure of Competition. Harvard University Press
5. Capiluppi, A., Lago, P., Morisio, M.: Characteristics of Open Source Projects p. 317 (2003)
6. Crowston, K., Howison, J.: The social structure of free and open source software development. First Monday 10(2)
7. Csardi, G., Nepusz, T.: The igraph software package for complex network research. InterJournal Complex Systems, 1695 (2006)
8. Germán, D.M.: The GNOME project: a case study of open source, global software development. Software Process: Improvement and Practice 8(4), 201–215 (2004)
9. Gilsing, V., Nooteboom, B., Vanhaverbeke, W., Duysters, G., Oord, A.V.D.: Network embeddedness and the exploration of novel technologies: Technological distance, betweenness centrality and density. Research Policy 37(10), 1717–1731 (2008)
10. González-Barahona, J.M., Robles, G., Andradas-Izquierdo, R., Ghosh, R.A.: Geographic origin of libre software developers. Information Economics and Policy 20(4), 356–363 (2008)
11. von Hippel, E., von Krogh, G.: Open Source Software and the "Private-Collective" Innovation Model: Issues for Organization Science. Organization Science 14(2), 209–223 (2003)
12. Howison, J., Inoue, K., Crowston, K.: Social dynamics of free and open source team communications. In: Damiani, E., Fitzgerald, B., Scacchi, W., Scotto, M. (eds.) Proceedings of the IFIP Second International Conference on Open Source Software

(Lake Como, Italy). IFIP International Federation for Information Processing, vol. 203, pp. 319–330. Springer, Boston (2006)

13. von Krogh, G., von Hippel, E.: The Promise of Research on Open Source Software. Management Science 52(7), 975–983 (2006)
14. Laursen, K., Salter, A.: Open for innovation: the role of openness in explaining innovation performance among U.K. manufacturing firms. Strategic Management Journal 27(2), 131–150 (2006)
15. Lazear, E.P.: Globalisation and the Market for Team-Mates. The Economic Journal 109(454), C15–C40 (1999)
16. Lerner, J., Tirole, J.: The Scope of Open Source Licensing. Journal of Law, Economics, and Organization 21(1), 20–56 (2005)
17. Long, Y., Siau, K.: Social Network Structures in Open Source Software Development Teams. Journal of Database Management 18(2), 25–40 (2007)
18. López-Fernández, L., Robles, G., González-Barahona, J.M., Herraiz, I.: Applying Social Network Analysis Techniques to Community-driven Libre Software Projects. International Journal of Information Technology and Web Engineering 1(3), 27–48 (2006)
19. Madey, G., Freeh, V., Tynan, R.: Modeling the free/open source software community: A quantitative investigation. In: Koch, S. (ed.) Free/Open Source Software Development, pp. 203–220. Idea Group Publishing (2005)
20. Mayer-Schönberger, V.: Can we reinvent the Internet? Science 325(5939), 396–7 (2009)
21. Mockus, A., Fielding, R.T., Herbsleb, J.D.: Two case studies of open source software development: Apache and Mozilla. ACM Transactions on Software Engineering and Methodology (TOSEM) 11(3) (2002)
22. Moon, J.Y., Sproull, L.S.: Essence of Distributed Work: The Case of the Linux Kernel. First Monday 5(11) (2000)
23. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Physical Review E 69(2), 026113 (2004)
24. Pons, P., Latapy, M.: Computing Communities in Large Networks Using Random Walks. Journal of Graph Algorithms and Applications 10(2), 191–218 (2006)
25. Robles, G., González-Barahona, J.M.: Developer identification methods for integrated data from various sources. Proceedings of the 2005 international workshop on Mining software repositories 30(4) (2005)
26. Scacchi, W., Feller, J., Fitzgerald, B., Hissam, S., Lakhani, K.: Understanding Free/Open Source Software Development Processes. Software Process: Improvement and Practice 11(2), 95–105 (2006)
27. Shah, S.K.: Motivation, Governance, and the Viability of Hybrid Forms in Open Source Software Development. Management Science 52(7), 1000–1014 (2006)
28. Shaikh, M., Cornford, T.: Version Control Tools: A Collaborative Vehicle for Learning in F/OS. In: Collaboration, Conflict and Control: The 4th Workshop on Open Source Software Engineering 2004. Edimburgh, Scotland (2004)
29. Van Antwerp, M., Madey, G.: Advances in the SourceForge Research Data Archive (SRDA), paper presented at the , Milan, Italy, September 2008, IFIP 2.13. In: Fourth International Conference on Open Source Systems. Milan (2008)
30. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge, UK (1994)
31. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature 393(6684), 440–2 (1998)

32. Wiggins, A., Howison, J., Crowston, K.: Social Dynamics of FLOSS Team Communication Across Channels. In: Proceedings of the Fourth International Conference on Open Source Software. Milan, Italy (2008)
33. Xu, J., Gao, Y., Christley, S., Madey, G.: A Topological Analysis of the Open Souce Software Development Community. In: Proceedings of the 38th Annual Hawaii International Conference on System Sciences (2005)