

Analyzing the Social Ties and Structure of Contributors in Open Source Software Community

Mohammad Y. Allaho and Wang-Chien Lee

Department of Computer Science and Engineering, The Pennsylvania State University
University Park, Pennsylvania 16802
Email: mya111,wlee@cse.psu.edu

Abstract—We conduct a statistical analysis on the social networks of contributors in Open Source Software (OSS) communities using datasets collected from two most fast-growing OSS social interaction sites, Github.com and Ohloh.net. Our goal is to analyze the connectivity structure of the social networks of contributors and to investigate the effect of the different social tie structures on developers' overall productivity to OSS projects. We first analyze the general structure of the social networks, e.g., graph distances and the degree distribution of the social networks. Our analysis confirms that the social networks of OSS communities follow power-law degree distributions and exhibit small-world characteristics. However, the degree mixing pattern shows that high degree nodes tend to connect more with low degree nodes, suggesting collaborations between experts and newbie developers. Second, we study the correlation between graph degrees and the productivity of the contributors in terms of the amount of contribution and commitment to OSS projects. The analysis demonstrates evident influence of the social ties on the developers' overall productivity.

I. INTRODUCTION

Large collaborative online communities have appeared and grown rapidly in the presence of Web 2.0 technology, witnessed by the massive success of Open Source Software (OSS) projects such as the Apache projects and GNU/Linux. As members of OSS projects are usually volunteers [1], their motivations usually involve personal goals/interests, e.g., sharpening their programming skills and gaining experience, following fellow peers, networking with the OSS community members, or simply supporting free open software projects [2]. In contrast to commercial software with dedicated responsibility to software engineers, the amount of participation and commitment by the volunteering developers are crucial factors for the success of OSS projects [3].

To investigate and better understand the success factors of OSS, several research studies have been reported in the literature projects [4]–[9]. However, these studies are mainly focusing on success factors related to the user-developer interaction (e.g., user ratings and bug reporting) and project related factors (e.g., programming language used, type of operating system, type of open source license, etc). Recently, the importance of social factors in OSS community has drawn a growing amount of attention. Studies in [10] and [11] suggest that the rapport created from collaboration among individuals in open collaborative communities indeed plays a crucial role leading to successes of OSS projects. However, the potential correlation between the productivity of OSS developers and the social ties among those developers in the community has not been quantitatively studied before. In this paper we

aim to answer these research questions: What are the general structure and features of the OSS community social networks? What are the social ties patterns in OSS community social networks that affect the volunteering developers' contribution and commitment?

In order to answer these critical research questions, we first analyze the general structure of the OSS Social Network (SN) and then investigate the correlation between various social tie structures and the contribution and commitment of volunteering developers to OSS projects. We define the OSS SN as a directed graph where nodes represent developers and links represent a following/follower relationship (similar to the Twitter network). We collect data from two fast-growing OSS collaboration sites, *Github.com* and *Ohloh.net*. These two sites are unique since they provide many social interaction tools for developers, including the ability to form a network of followings/followers relationship which gives explicit information of acquaintances. We emphasize here that most previous studies on the OSS SNs typically construct the SN based on communication links which results in an implicit SN (i.e., the links may not represent true acquaintances for various reasons). Aiming to better understand the characteristics of OSS SNs, we perform a series of analysis on the social network structure by measuring SNs distances (i.e., diameter and average path length), analyzing degree distributions, testing power-law model, and examining degree mixing patterns between nodes (developers) in the OSS SN.

Our social network analysis confirms a small-world characteristics for the OSS SNs. Also, the analysis shows that the degree distribution of the SNs follows a power law distribution that fits the scale-free network model. Moreover, the degree mixing pattern analysis for the SNs shows that high degree nodes tend to connect more with low degree nodes, suggesting collaboration between experts and newbie developers.

In addition to the analysis of network structures, we also investigate the factors of social ties under the context of global network graph to explore the social factors that affect the developers' contribution and commitment. A global network analysis studies the connectivity among individuals in the whole OSS community. The network analysis demonstrates evident influence of the social ties on the developers' overall productivity.

Our work is the first to focus on the correlations between social factors and developers' productivity. We quantify the developers' productivity from an abstract meaning into measurable metrics in order to comprehend the amount of contribution

TABLE I. COLLECTED DATA STATISTICS.

	<i>Ohloh</i>	<i>Github</i>
Contributions	671,835	9,268,644
Contributors	297,046	1,034,996
Contributors with account	14,958	652,040
SN Edges	36,205	1,788,803
Projects	456,811	2,332,749

and commitment each developer exerts in a project. Our findings may pave the road for future recommendation systems related to online open collaborative projects.

The rest of the paper is organized as follows. Section II introduces the dataset collected and used in this study. Section III shows the analysis of network structure. The analysis includes network structure measures, degree distribution and degree mixing patterns analysis. Section IV investigates the correlation between the social ties and developers' overall productivity. Section V introduces several related works and, finally, Section VI concludes the paper.

II. DATASET

We collect data from two fast-growing OSS social interaction sites, *Github.com* and *Ohloh.net*. These two sites provide a variety of information regarding the social relationship and contributions of their volunteering developers. Unlike other OSS collaboration sites (e.g., *sourceforge.net*, *freshmeat.net*), Github and Ohloh provide facilities for contributors to explicitly recognize each other through a *follow* link in Github and *Give kudo* in Ohloh which creates a directed social network of contributors/developers. We collect data over a million of contributors and over two millions of OSS projects. The data were collected via APIs provided by Github and Ohloh.

Table I summarizes some general statistics of the collected data from Github and Ohloh. In Table I, "Contributions" denotes the total number of distinct contributions in all OSS projects, whereas "Contributors", in the second row, is the distinct number of contributors in each OSS community. For example, if a community has two contributors *A* and *B*, and contributor *A* participates in 2 projects while contributor *B* participates in 3 projects, then this community has 2 contributors and total of 5 contributions. Furthermore, the "Contributors with account" are the contributors registered in the site with a profile information. In our analysis, we only consider contributors with account because it is possible to link between their contribution information and social attributes which is not possible with anonymous contributors. This may cause a threat to validity since anonymous contributors could add weight to any connectivity pattern. Another threat is that we only capture explicit acquaintances among the developers, where there might be acquaintances that are not declared. However, since we conduct the analysis on a large number of OSS developers and projects without selection criteria, the threats to validity is minimized. Moreover, the results in the next sections show consistency on both Ohloh and Github data.

III. NETWORK STRUCTURE ANALYSIS

In this section we study the general graph distance measures, the node degree distribution and the mixing pattern of

TABLE II. OSS SOCIAL NETWORK STRUCTURE MEASUREMENTS.

	Diameter	Avg. Path Length	Avg. Degree	Avg. CC
Github	22	6.25	4.31	0.112
Ohloh	23	7.11	3.21	0.25

TABLE III. POWER-LAW COEFFICIENT ESTIMATES (α) AND CORRESPONDING KOLMOGOROV-SMIRNOV GOODNESS-OF-FIT METRIC (D).

	Indegree		Outdegree	
	α	D	α	D
Github	2.31	0.0258	2.82	0.0433
Ohloh	2.77	0.0281	3.04	0.0231

degrees in both OSS communities' SNs.

A. Basic Measures of the Network Structure

We first measure the SNs graph distances to comprehend the general structure of the networks. Table II shows the diameter and average path length of Github and Ohloh SNs. Recall that the graph diameter is the longest distance of the shortest distances between every pair of nodes in the network, while the average path length is the average length of those shortest paths. The average path length found in both Github and Ohloh confirms the result by Stanley Milgram [12] which states that any two people are on average separated by six intermediate connections, known as the "six degrees of separation" in the "small world" phenomenon. Also, the diameter of the OSS SN is small compared to non-social networks (e.g., the Web network has a diameter of 905 [13]), which is a feature of most social graphs. Additionally, Table II includes the average degree and clustering coefficient (CC) of the OSS SNs. The structure measures of these two SNs show great similarity. In the next section, we test the OSS SN against the scale-free network model.

B. Power-law Degree Distribution

Node degree analysis is essential to reveal the connectivity trends between nodes (developers) in a graph. Thus, we first investigate whether the OSS SNs follow a scale-free network model or not. A scale-free network has a degree distribution following a power-law model where the exponent α is typically $2 < \alpha < 3$ in most cases. Figure 1 shows the outdegree and indegree Complementary Cumulative Distribution Function (CCDF) of the OSS SN for Github and Ohloh, respectively. Basically, the plots show the probability of any node having a degree greater than or equals the given degree x . Both outdegree and indegree distributions are consistent with the fitted power-law distribution model. To confirm this result, Table III shows the maximum likelihood estimate of the scaling exponent α and the Kolmogorov-Smirnov goodness-of-fit statistic D , where a value close to zero indicates that the data is closely estimated by the given model. Both networks follow a power-law distribution and confirm properties of scale-free networks where few nodes have high degrees and the majority have low degrees.

C. Node Degree Mixing Pattern

In this section, we investigate the mixing pattern of degrees between connected nodes. In other words, we would like to

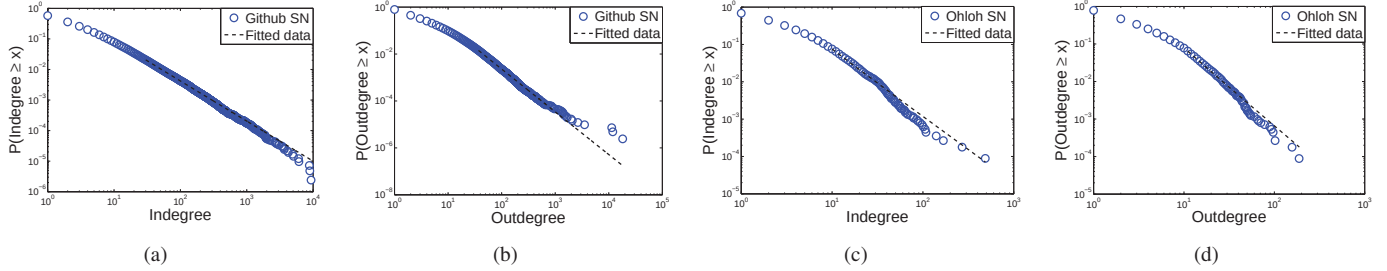


Fig. 1. Github and Ohloh Out/Indegree Complementary Cumulative Distribution Functions (CCDF) for the OSS SN ($x=Out/Indegree$). All social networks show properties consistent with power-law networks.

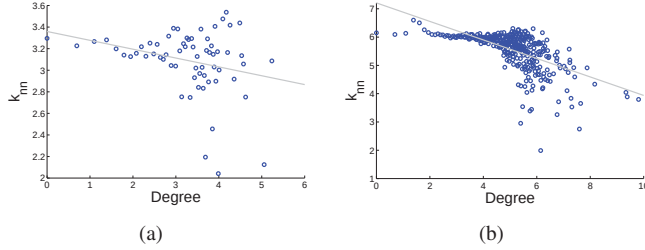


Fig. 2. Ohloh (left) and Github (right) k_{nn} functions show that, in OSS SN, high degrees tend to connect to lower degrees which indicate a mixture of degree connectivity.

know whether similar degree nodes tend to connect to each other or is it a mixture? This study is important since it reveals important insights to the structure of the network.

1) *Degree correlation*: First we study the *Joint Degree Distribution* (JDD), also referred to as the *mixing pattern* [14] and the *2K distribution* [15]. JDD gives the probability of a node of degree k_1 to be connected with a node of degree k_2 . JDD is approximated by the *degree correlation function* k_{nn} . For an undirected graph, k_{nn} is a mapping between every degree and the average degree of all neighboring nodes (connected) to nodes of that degree. On the other hand, k_{nn} for a directed graph is a mapping between every outdegree and the average indegree of all neighboring nodes (connected) to nodes of that outdegree. Figure 2 shows the OSS SN k_{nn} plots. Both Ohloh (Figure 2(a)) and Github (Figure 2(b)) show that high-degree nodes tend to connect to lower-degree nodes (note that the slope of the k_{nn} distribution is negative), which indicates that OSS SN have a mixture of degree connectivity. This mixing between different degree nodes shows that there are interactions between experts and newbie developers which is very positive in an open collaborative environment.

2) *Assortativity*: Assortativity is a measure of the likelihood for nodes to connect to other nodes with similar degrees. The assortativity coefficient r ranges between -1 and 1. Positive assortativity coefficient means that nodes tend to connect to other nodes with similar degree, while negative assortativity coefficient means that nodes tend to connect to other nodes with different degrees. Table IV shows the assortativity coefficient for OSS SN for both Ohloh and Github. The negative assortativity coefficients indicate that high-degree nodes tend to connect to low-degree nodes, which confirms the observation of degree correlation discussed earlier. The small negative value of the assortativity coefficient for the OSS SN indicates that the degree mixture is not high between nodes.

TABLE IV. ASSORTATIVITY COEFFICIENTS.

	Ohloh	Github
Social Net.	-0.0177	-0.0465

IV. SOCIAL NETWORKING FACTORS AND DEVELOPER CONTRIBUTION ANALYSIS

To discover the social factors that affect the contributors' participation and contribution in OSS projects, we conduct an extensive analysis to investigate the correlations between the social connectivity ties and the amount of contribution exerted by the contributors. We conduct a global graph analysis, where we analyze the developers' connectivity in the whole community graph. In particular, we study the effect of node (developer) indegree and outdegree on the developer's contribution and commitment. Next, we elaborate in more details the social factors and contribution measures.

We analyze the correlation between node indegree/outdegree and the developer's contribution and commitment. The amount of contribution for each developer is estimated by two metrics: (i) the number of participations in distinct OSS projects, and (ii) the total number of commits to all OSS projects, where a commit is an update submitted to a project by a developer. Furthermore, a developer commitment is estimated by the number of months where a developer submitted at least one commit denoted by *months-work*. This metric is collected from Ohloh only since Github does not provide the number of active months for developers. We note that months-work is not simply the number of months between first and last commit for a developer. In particular, the number of idle months are not counted. Note that this information is hard to monitor and collect for a large number of developers, unless it is provided by the social platform provider.

Figure 3 shows the in/outdegree correlation with the number of participations for Ohloh and Github data. In particular, for each degree we take the average participation of nodes with that degree. Both indegree and outdegree show positive correlation with the number of participations indicating that developers with high indegree or outdegree have more tendency to participate in more OSS projects.

Figure 4 shows the in/outdegree correlation with the total number of commits in OSS projects for Ohloh and Github datasets. In particular, for each degree we take the average commits of nodes with that degree. Similar to participations, both indegree and outdegree show positive correlation with the

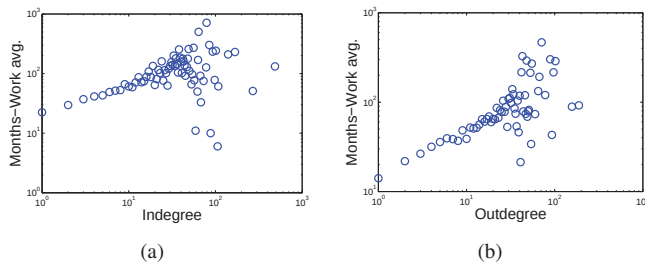


Fig. 5. Indegree (left) and outdegree (right) vs. months-work for Ohloh.

total commits indicating that developers with high indegree or outdegree have more tendency to commit more work in OSS projects.

In Figures 3 and 4, we notice a bulky head for the correlation points distribution, which suggests a wider range of participation and total commits values for nodes with high in/outdegree. This is more obvious in Github dataset than Ohloh dataset since the amount of data in Github is much more than the amount of data in Ohloh. Since we are aggregating nodes based on in/outdegree and the node degree follows a power-law distribution, where low degree nodes are much more than high degree nodes, then the data points aggregated for low-degree nodes are much more than high-degree nodes which makes the results for low degree nodes more consistent than high degree nodes.

Finally, Figure 5 shows the in/outdegree correlation with the total number of months-work in OSS projects for Ohloh data. For each degree we take the average months-work of nodes with that degree. Similar to participations and total commits, both indegree and outdegree show positive correlation with the total months-work indicating that developers with high indegree or outdegree have more tendency to stay longer time committing to OSS projects.

V. RELATED WORKS

There exist several works that analyze the OSS social network. Both [17] and [18] study the structure of a collaborator communication network extracted from the e-mail network and find that small teams have a centralized structure, whereas larger teams have a modular structure. However, they do not study the correlation between network structure and developers productivity. In contrast to the above works, our datasets include social networks based on explicit acquaintances, which facilitate us to investigate the success factors of OSS projects. In another work, Subramaniam et al. [4] study other factors that may affect the development activity, such as developer interest, user interest, project activity (i.e., bug reports and user comments), project status (i.e., new project, development phase, maintenance phase) and project attributes (i.e., OSS license, operating system, programming language). On the contrary, we focus on the social tie factors of the contributors in the OSS projects. Moreover, Casaló et al. [5] collect information via questionnaires and study the correlation between developers commitment to their OSS project and the reputation of those projects. They conclude that OSS reputation have an indirect effect on the commitment of collaborators. In contrast to questionnaires, the contribution and commitment

information in our study are collected from control version repositories that monitor the commits log for each developer.¹

VI. CONCLUSION

We conducted a statistical social network analysis on more than a million of online volunteering developers. Our study included a global network analysis (i.e., OSS community level) showing that the OSS SNs follow a power-law distribution and the SNs exhibit small world properties. Also, the OSS SN degree mixing pattern shows that nodes with different degrees are connected together, suggesting interactions between experts and newbie developers. Second, we show positive correlations between indegree/outdegree and the developers' contribution and commitment in OSS projects. These results lead into better understanding of the OSS social factors that influence team productivity and commitment essential to the success of OSS projects. Moreover, this study paves the road for future recommendation systems related to online open collaborative projects, which is our future work.

REFERENCES

- [1] M. Michlmayr and B. M. Hill, "Quality and the reliance on individuals in free software projects," in *Proceedings of the 3rd Workshop on Open Source Software Engineering*, Portland, OR, USA, 2003, pp. 105–109.
- [2] G. Hertel, S. Niedner, and S. Herrmann, "Motivation of software developers in open source projects: an internet-based survey of contributors to the linux kernel," *Research Policy*, vol. 32, no. 7, pp. 1159–1177, July 2003.
- [3] G. Robles, J. M. Gonzalez-Barahona, and M. Michlmayr, "Evolution of volunteer participation in libre software projects: Evidence from Debian," in *Proceedings of the First International Conference on Open Source Systems*, Genova, Italy, 2005, pp. 100–107.
- [4] C. Subramaniam, R. Sen, and M. L. Nelson, "Determinants of open source software project success: A longitudinal study," *Decis. Support Syst.*, vol. 46, no. 2, pp. 576–585, Jan. 2009.
- [5] L. V. Casaló, J. Cisneros, C. Flavián, and M. Guinaliú, "Determinants of success in open source software networks," *Industrial Management and Data Systems*, vol. 109, no. 4, pp. 532–549, 2009.
- [6] S.-Y. T. Lee, H.-W. Kim, and S. Gupta, "Measuring open source software success," *Omega*, vol. 37, no. 2, pp. 426–438, Apr. 2009.
- [7] V. Garousi, "Investigating the success factors of open-source software projects across their lifetime," *Journal of Software Engineering Studies*, vol. 4, p. 115, 2009.
- [8] I. Samoladas, G. Gousios, D. Spinellis, and I. Stamelos, "The sqo-oss quality model: Measurement based open source software evaluation," in *OSS*, 2008, pp. 237–248.
- [9] J. Wu and K.-Y. Goh, "Evaluating longitudinal success of open source software projects: A social network perspective," in *HICSS*, 2009, pp. 1–10.
- [10] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan, "Group formation in large social networks: membership, growth, and evolution," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD '06. New York, NY, USA: ACM, 2006, pp. 44–54.
- [11] J. Hahn, J. Y. Moon, and C. Zhang, "Emergence of new project teams from open source software developer networks: Impact of prior collaboration ties," *Information Systems Research*, vol. 19, no. 3, pp. 369–391, 2008.
- [12] S. Milgram, "The small world problem," *Psychology Today*, vol. 61, pp. 60 – 67, May 1967.
- [13] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, ser. IMC '07. New York, NY, USA: ACM, 2007, pp. 29–42.

¹The commits log is collected by *Ohloh.net* and *Github.com*.

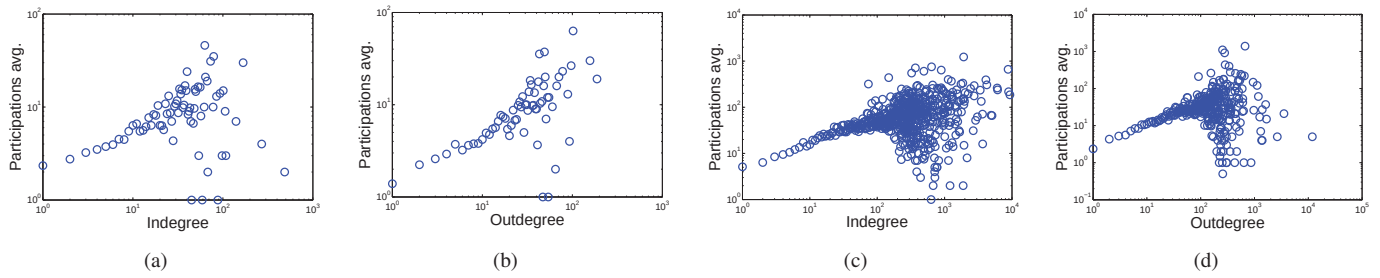


Fig. 3. Indegree and outdegree vs. number of participation for Ohloh ((a) and (b)) and Github ((c) and (d)).

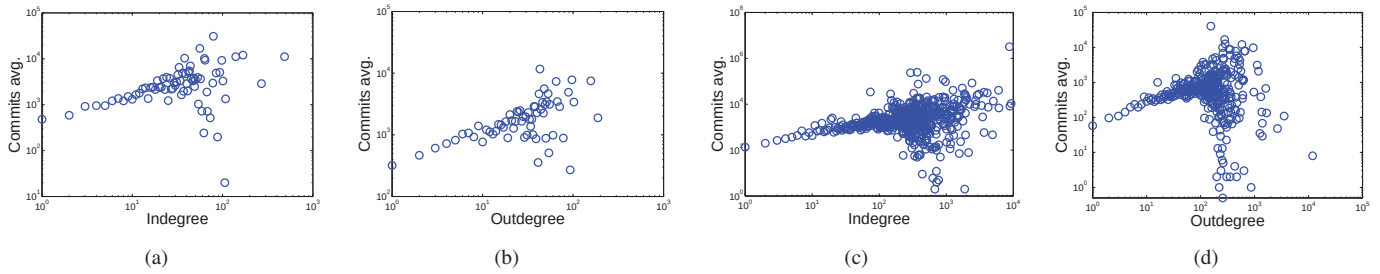


Fig. 4. Indegree and outdegree vs. total number of commits for Ohloh ((a) and (b)) and Github ((c) and (d)).

- [14] M. E. J. Newman, "Mixing patterns in networks," *PRE*, vol. 67, no. 2, p. 026126, February 2003.
- [15] P. Mahadevan, D. Krioukov, K. Fall, and A. Vahdat, "Systematic topology analysis and generation using degree correlations," in *Proceedings of the 2006 conference on Applications, technologies, architectures, and protocols for computer communications*, ser. SIGCOMM '06. New York, NY, USA: ACM, 2006, pp. 135–146.
- [16] G. Madey, V. Freeh, and R. Tynan, "The open source software development phenomenon: An analysis based on social network theory," in *Proceedings of the Americas Conference on Information Systems (AMCIS 2002)*, Dallas, Texas, 2002, pp. 1806–1813.
- [17] C. Bird, D. Pattison, R. D'Souza, V. Filkov, and P. Devanbu, "Latent social structure in open source projects," in *Proceedings of the 16th ACM SIGSOFT International Symposium on Foundations of software engineering*, ser. SIGSOFT '08/FSE-16. New York, NY, USA: ACM, 2008, pp. 24–35.
- [18] K. Crowston and J. Howison, "The social structure of open source software development teams," *First Monday*, vol. 10, no. 2, 2003.