A Study of SourceForge Users and User Network

Liguo Yu[1], Srini Ramaswamy[2]

1. Computer Science Department, Indiana University South Bend, USA, ligyu@iusb.edu
2. ABB Corporate Research Center, Bangalore, India, srini@ieee.org

Abstract

SourceForge is a web-based code repository for open-source projects. It is one of the most successful web sites that promote code sharing, collaboration, and open-source software development. Users of SourceForge are globally distributed and come from every corner of the world. They interact with each other in the virtual space through the formation of user groups. Accordingly a user network is formed on SourceForge. The user network is not only a virtual network, but also serves as a platform for open-source and distributed software development. A careful study of SourceForge can hence provide useful insight into both distributed software development and the emergence of developer social networks. In this paper, the SourceForge users and user network are studied, including user growth rate, geographical distribution, and network clusters. The findings presented in this study are an attempt to enrich our knowledge about online software developer networks.

Introduction

SourceForge is a web-based code repository for open-source projects (Maguire 2007). It has over millions of users and hosts nearly a million projects. It is one of the most popular platforms for open-source developers to share code, track changes, and report bugs. It also facilitates distributed development and global collaboration. Since SourceForge contains a huge amount of valuable information about projects, developers, and users, considerable research has been done to mine its repository and retrieve useful information about software development and management. Here, we review some of the important findings about SourceForge that have been reported.

Gao and Madey (2007a) studied the community network of the SourceForge. They performed three different analyses on SourceForge network, including structure analysis, centrality analysis, and path analysis. In another research (Gao and Madey 2007b), they used an iterated method to generate models simulating the evolution of SourceForge network. Their group also studied the social network properties of SourceForge, such as degree distribution, diameter, cluster size, and clustering coefficient (Xu, Christley, and Madey 2006). They found that SourceForge networks have scale-free properties and the small world phenomenon.

Christley and Madey (2007) studied the activities of different users of SourceForge, including project administrators, message posters, software developers, and handypersons. Robles and Gonzalez-Barahona (2006) studied the geographical locations of SourceForge developers. Krishna and Srinivasa (2012) studied the six top-rated projects in SourceForge. They found that most of the work in these projects is done by a small number of developers. Huang and Liu (2005) studied the learning process of programmers in SourceForge network. Kerr et al. (2006) used a diagram to visualize the evolution of an individual programmer's contributions to code and comments of SourceForge projects. Grechanik et al. (2010) studied java projects in SourceForge. They explained how object-oriented design principles are followed in open-source development.

Social network analysis of software developer network is another well addressed topic (Cataldo and Herbsleb 2008; Meneely et al. 2011; Zhang et al. 2011; Bird et al. 2008). For examples, studies are reported to investigate the effect of developer network on software development, and software quality (Singh 2010; Ehrlich and Cataldo 2012); some studies are performed to analyze software developer roles (Pohl and Diehl 2008; Yu and Ramaswamy 2007); and some work are reported to visualize developer network (Jermakovics et al. 2011. Sarma et al. 2009).

Data Source, Data Description, and Motivation of the Study

The data used in this study is retrieved from the SourceForge research data archive of the University of Notre Dame (Gao et al. 2007; Antwerp and Madey 2008;

Madey 2013). The archive contains complete monthly data of SourceForge from February 2005 until June 2012. Data in the archive are saved as database tables. Queries are used to retrieve relevant data. The retrieved data are saved as text file format for further processing.

We note here that the data archive has been mined and used by many researchers. Numerous research papers have been published based on the dataset. The motivation for this study is twofold: (i) to explore recent changes in SourceForge users and user network as it has changed tremendously over the past five years, when some of the earlier research were conducted. Hence it is worth studying the new data to find new information about the SourceForge user network. (ii) Cluster size is an important factor in virtual networks and hence it is worth studying the distributions of SourceForge clusters of different sizes.

Analysis and Results

General Results

Figure 1 shows the growth of the number of users of SourceForge. It can be seen that the number of registered users in SourceForge has increased about four times from February 2005 to June 2012. The number of users in each month is fitted with two models, linear model and exponential model, which are listed in Table 1. Although significance of the two models is both at the 0.001 level, the exponential model has a larger R-square value, which means that the number of registered users in SourceForge grows exponentially with time.

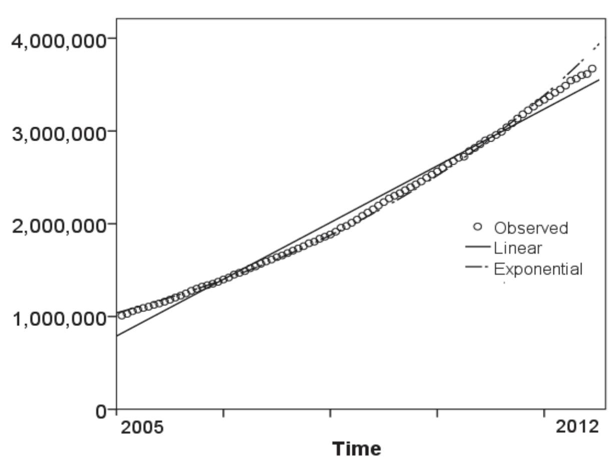


Figure 1. The Growth of the number of users in SourceForge.

Table 1. The models of user growth in SourceForge

Model	R Square	Significance
Linear	0.986	<0.001
Exponential	0.997	<0.001

Figure 2 shows the geographical distribution of SourceForge users as in June 2012. It can be seen that most users reside in America (including both North America and South America). Asia has more users than Europe. There are even users from Arctic and Antarctica (less than 300). Figure 3 shows the number of English users and non-English users. Figure 4 shows the distribution of non-English languages used by SourceForge users.

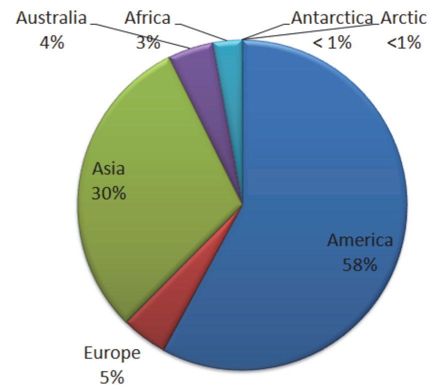


Figure 2. The geographical distribution of SourceForge users as of June 2012.

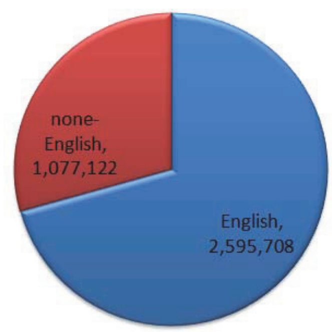


Figure 3. Number of English users and non-English users in SourceForge as of June 2012.

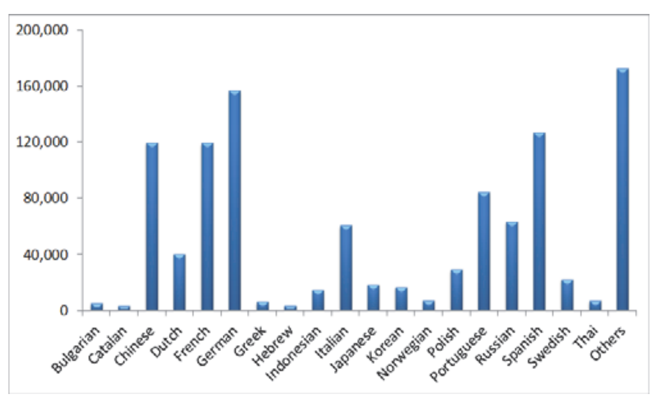
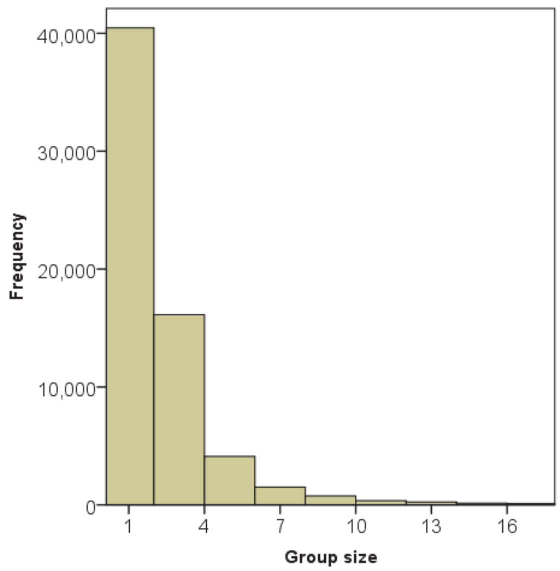


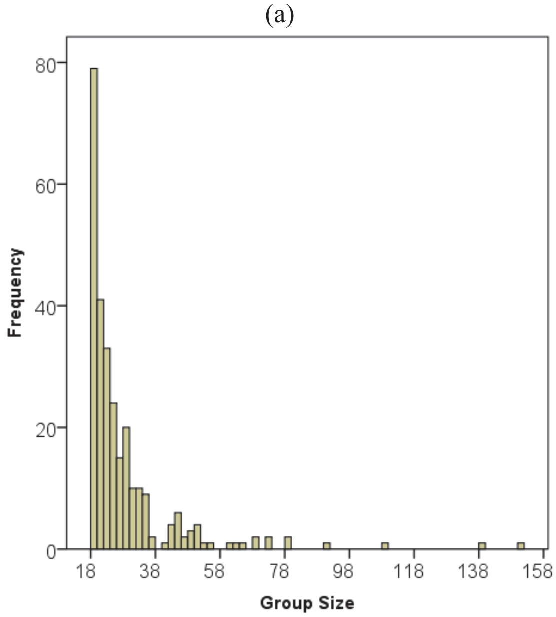
Figure 4. The distribution of non-English SourceForge users as of June 2012.

User Groups and Projects

In SourceForge, users form a group through working on a common project. Therefore, the number of users in a group is the number of users of a project. The user could take any role in this project, such as owner, manager, developer, or end user. Figure 5 shows the frequencies of different size of user groups. It can be seen that most projects (user groups) are small. Over 40 thousand projects just have less than two users; over 15 thousand projects have three or four users; and there are only 4 projects that have near or over one hundred users.



(a)



(b)

Figure 5. The distribution of user-groups with group size in SourceForge as of June 2012: (a) group size 1-17; (b) group size: 18 and greater.

A project's size does not necessarily represent its popularity and activity. While popularity could be measured with the number of downloads, number of page views; activity could be measured with number of messages posted, number of bugs reported, and the number of CVS commits recorded. To study the correlation between group size and a project's popularity and activity, Spearman's rank correlation tests are performed and the results are summarized in Table 2, where α represents the correlation coefficient and *p* represents the significance (2-tailed). We note here that the popularity measurement and activity measurement contain all the history data until June 2012. From Table 2, we can see that the Spearman's tests show that the group size has positive linear correlation with project popularity and activity. The significance is at the 0.001 level.

Table 2. Spearman's rank correlation between group size and its popularity and activity measurements.

		Number of downloads	**Number of page views**	**Number of posted messages**	**Number of bugs reported**	**Number of CVS commits**
Group size	α	0.286	0.336	0.315	0.362	0.423
	p	<0.001	<0.001	<0.001	<0.001	<0.001

User Network Clusters

In SourceForge, one user could join more groups (projects) and one project could have more users. Therefore, a user network could be formed through user groups (projects). In this network, users are nodes and user groups (projects) are edges connecting the nodes. In this user network, each isolated graph is called a cluster. Several special clusters are described below.

Single-user clusters. A single-user cluster has just one user. This user may participate in more projects. However, in all these projects, there is just one user, the single user. Examples of single-user clusters are illustrated in Figure 6(a-d), where rectangles represent users and circles represent projects.

Single-project clusters. A single-project cluster has just one project. This project might have more users. However, all these users only participate in one project, the single project. Examples of single-project clusters are illustrated in Figure 6(e-h), where rectangles represent users and circles represent projects. It is worth noting that Figure 6a and Figure 6e are the same cluster represented twice.

Table 3 shows the general results about user network clusters. Note here that not every user would join a group (project). Therefore, the number of users shown in Table 3 is only part of the registered users of SourceForge shown in Figure 1. Although there are over 374 thousand clusters, most of them are single-user and/or single-project clusters.

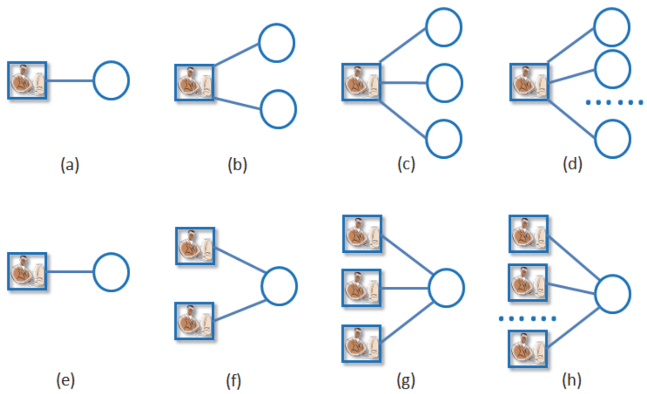
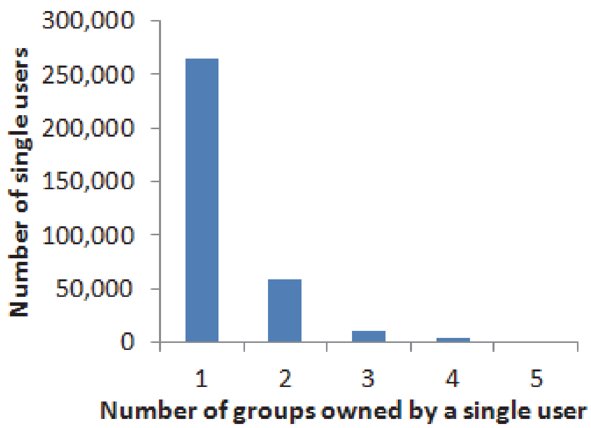


Figure 6. Examples of single-user clusters and single-project clusters: (a) single-user single-project; (b) single-user double-project; (c) single-user triple-project; (d) single-user more-project; (e) single-project single-user; (f) single-project double-user; (g) single-project triple-user; and (h) single-project more-user.

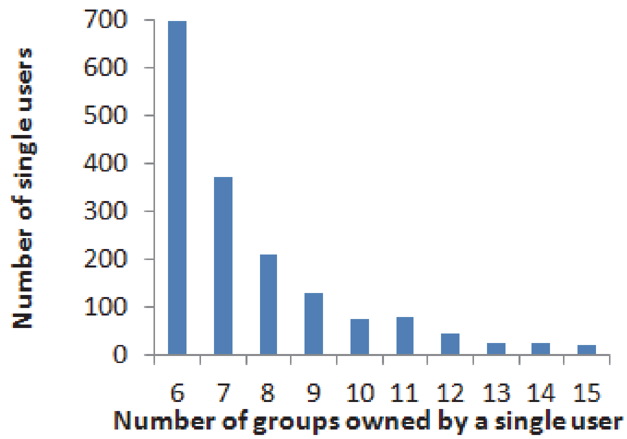
Table 3. General results about clusters.

Total number of users	562,198
Total number of projects	818,819
Total number of clusters	374,459
Single-user single-project clusters	265,037
Single-user multiple-project clusters	75,591
Single-project multiple-user clusters	12,709
Multi-user, multi-project clusters	21,122

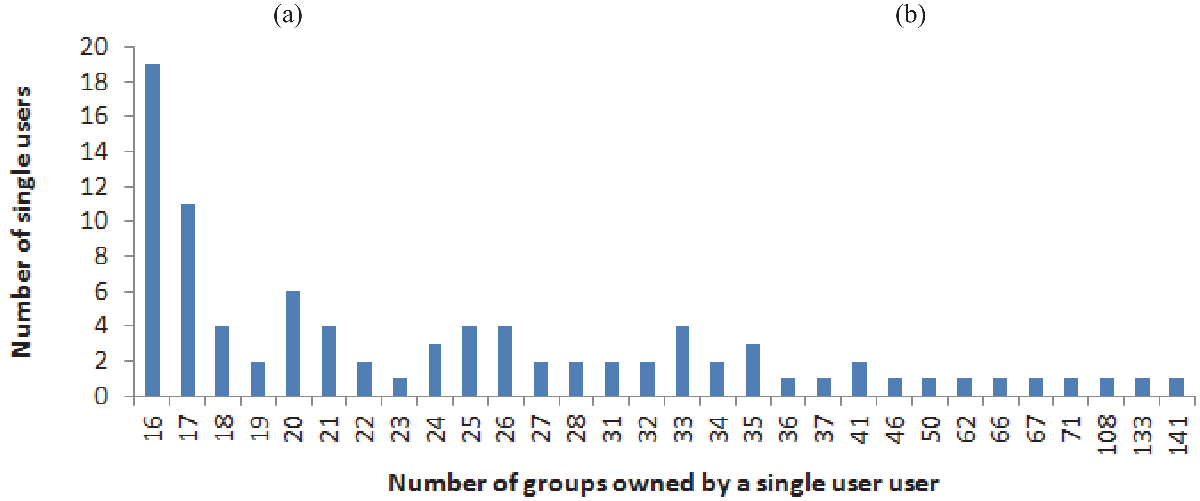
Figure 7 shows the distribution of single-user clusters. Figure 8 shows the distribution of single-project clusters. It should be noted that the single-user, single-project clusters are presented in Figure 7(a). So they are not shown in Figure 8(a). It is worth noting here that the reason to break single-user clusters into three figures (Figure 7) and to break single-project clusters into two figures (Figure 8) is that their data ranges are too big to be clearly illustrated in one figure.



(a)

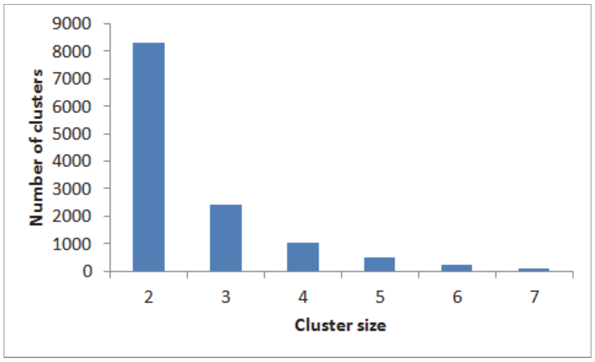


(b)

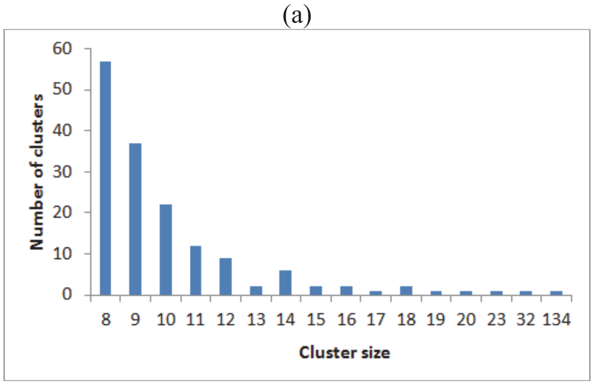


(c)

Figure 7. The distribution of single-user clusters according to the number of groups owned by a user: (a) 1-5; (b) 6-15; (7) 16 or more.



(a)



(b)

Figure 8. The distribution of single-project clusters according to the cluster size: (a) 2-7; and (b) 8 or more.

Most of the single-user clusters or single-project clusters are small scale user networks. Table 3 shows there are over 21 thousand multi-user, multi-project clusters. They are relatively large scale user networks and hence worthy of greater attention. The largest cluster in SourceForge contains 97,357 users and 68,083 projects. Other clusters are relatively small, which are in this paper referred as scattered clusters. Most of the scattered clusters are in the regular range of (2, 2) to (50, 50) as shown in Figure 9, where notation (a, b) represents user count a and project count b of a cluster. Only 34 scattered clusters are not in this range as shown in Figure 10.

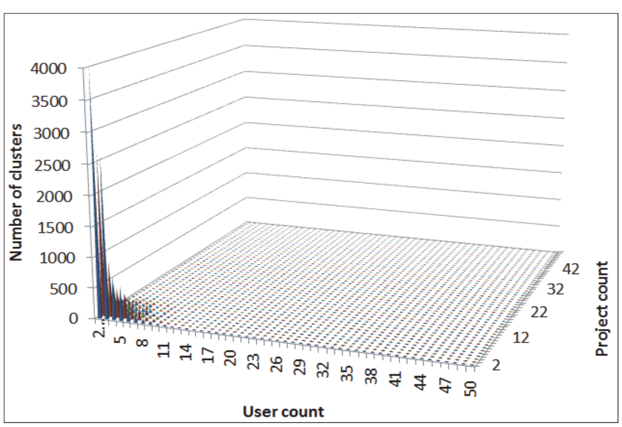


Figure 9. Number of (2, 2) to (50, 50) clusters.

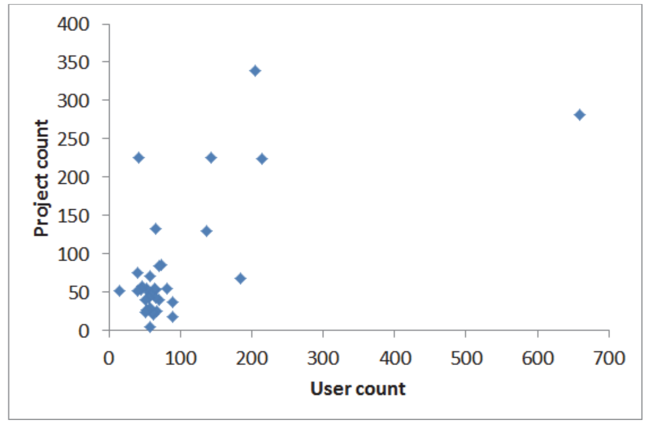


Figure 10. Thirty-four out-of-regular-range scattered clusters.

As described before, the largest cluster contains 97,357 users and 68,083 projects. To see how closely these users or projects are connected, we constructed two network graphs: top-100 user graph and top-25 project graph. Figure 11 shows the top-100 most popular users according to the number of projects they participated and Figure 12 shows their connections. Figure 13 shows the top-25 most popular projects according to the number of users participated in them and Figure 14 shows their connections. It is worth noting that (1) in Figure 12, circles represent users, users are connected through projects, and the edge width indicates the number of common projects two users have; (2) in Figure 14, circles represent project, projects are connected through users, and the edge width indicates the number of common users the two projects have.

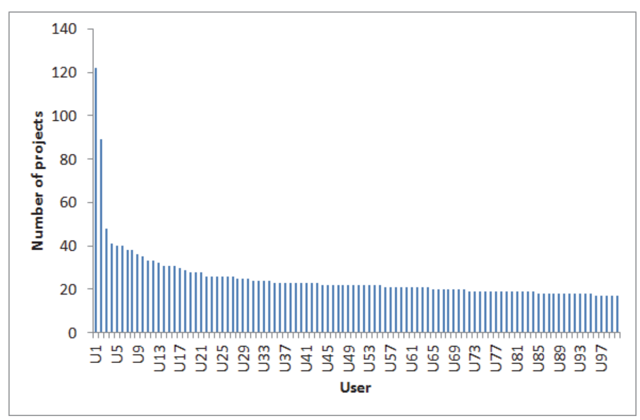


Figure 11. Number of projects in the top-100 users of the largest cluster

It can be seen that although the largest cluster is big, the inter-connections between users and projects are loose. It is true that as we add more users to the largest cluster, eventually all of the 97,357 users should be connected. To see the connectivity of the users in the largest cluster, we calculate the number of edges (connections) of each user. The result is illustrated in Figure 15. We can see most of the users have less than 5 edges. The average number of edges (connections) per user is 28. The largest number of connections a user has is 873. It is interesting to notice that this user is not the user that has the largest number of projects. Instead it is a user that participates in several largest projects. The user who is involved with the largest number of projects only has 49 connections. In other words, in order to get more connections, a user should join large projects with more users.

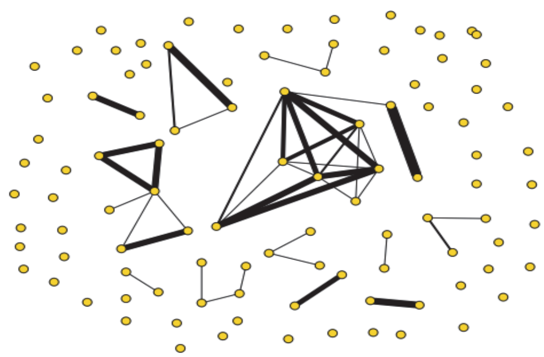


Figure 12. Interconnections among top-100 users in the largest cluster

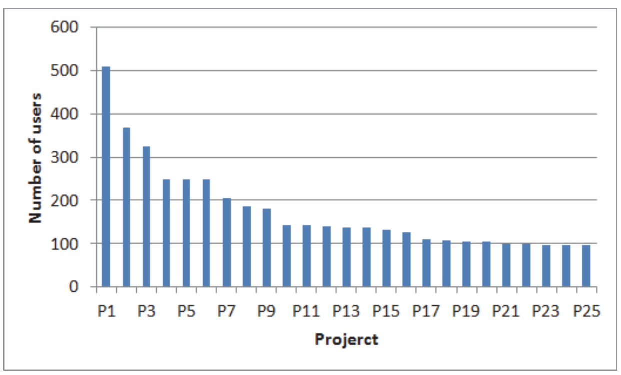


Figure 13. Number of users in the top-25 projects in the largest cluster

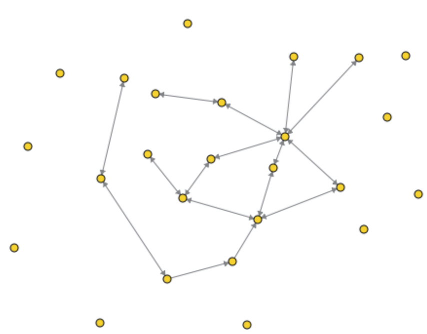
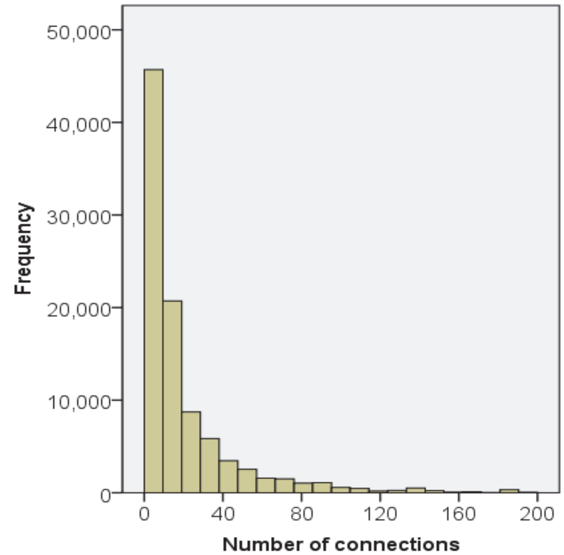
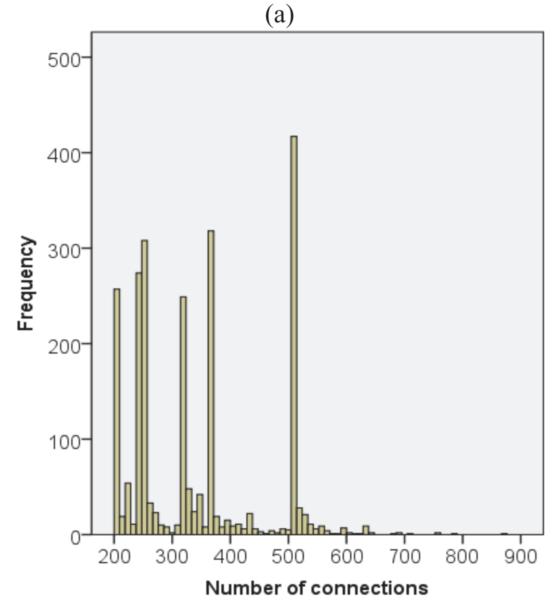


Figure 14. The connections of top 25 projects in the largest cluster.



(a)



(b)

Figure 15. The distribution of users with respect to the number of connections they have (the largest cluster): (a) users with number of edges (0-200); (b) users with number of edges 200 and more.

Conclusions

In this paper, we studied the user network of SourceForge. Our results can be summarized as the followings: (1) we found the user network of SourceForge has grown tremendously in the past several years and users are distributed all over the world; (2) we studied the clusters of user networks and found some interesting behaviors of the clusters. These include: (i) Most of the single-user clusters or single-project clusters are small scale user networks, (ii) Even in large clusters the inter-connections between users and projects are loose, and (iii) The number of connections a user has is proportionally related to the largeness of the project they are involved in, the complement – i.e. when a user participates in a large number of projects, it is not necessarily true that they have a large number of user connections.

Acknowledgements

The authors would like to thank Prof. Greg Madey of The University of Notre Dame for sharing SourceForge research dataset with us.

References

Antwerp, M. V.; and Madey, G. 2008. Advances in the SourceForge Research Data Archive (SRDA). In *Proceedings of the Fourth International Conference on Open Source Systems (WoPDaSD 2008)*, Milan, Italy.

Bird, C.; Pattison, D.; D'Souza, R.; Filkov, V.; and Devanbu, P. 2008. Latent Social Structure in Open Source Projects. In *Proceedings of the Sixteenth ACM SIGSOFT International Symposium on Foundations of Software Engineering*, 24–35, ACM, New York, NY, USA.

Cataldo, M.; and Herbsleb, J. D. 2008. Communication Networks in Geographically Distributed Software Development. In *Proceedings of the ACM conference on Computer supported cooperative work*, 579–588, ACM, New York, NY, USA.

Christley S.; and Madey, G. 2007. Analysis of Activity in the Open Source Software Development Community. In *Proceedings of the Fortieth Annual Hawaii International Conference on System Sciences*. Computer Society Press.

Ehrlich, K.; and Cataldo, M. 2012. All-For-One and One-For-All?: A Multi-Level Analysis of Communication Patterns and Individual Performance in Geographically Distributed Software Development. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, 945–954, ACM, New York, NY, USA.

Gao, Y.; and Madey, G. 2007a. Network Analysis of the SourceForge.net Community. In *Proceedings of the Third International Conference on Open Source Systems (OSS 2007),* Limerick, Ireland.

Gao, Y.; and Madey, G. 2007b. Towards Understanding: A Study of the SourceForge.net Community Using Modeling and Simulation. In *Proceedings of the Agent-Directed Simulation (ADS 2007)*, 145–150, Norfolk, VA, USA.

Gao, Y.; Antwerp, M.; Christley, S; and Madey, G. 2007. A Research Collaboratory for Open Source Software Research. In Proceedings of the Twenty Ninth International Conference on Software Engineering and Workshops, Minneapolis, MN, USA.

Grechanik, M.; McMillan, C.; DeFerrari, L.; Comi, M.; Crespi, S.; Poshyvanyk, D.; Fu, C.; Xie, Q.; and Ghezzi, C. 2010. An Empirical Investigation into a Large-Scale Java Open Source Code Repository. In *Proceedings of the ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*, Article 11, 10 pages, ACM, New York, NY, USA.

Huang, S.; and Liu, K. 2005. Mining Version Histories to Verify the Learning Process of Legitimate Peripheral Participants. *SIGSOFT Software Engineering Notes* 30(4): 1–5.

Jermakovics, A.; Sillitti, A.; and Succi, G. 2011. Mining and Visualizing Developer Networks from Version Control Systems. In *Proceedings of the Fourth International Workshop on Cooperative and Human Aspects of Software Engineering*, 24–31, ACM, New York, NY, USA.

Kerr, B.; Cheng, L.; and Sweeney, T. 2006. Growing Bloom: Design of a Visualization of Project Evolution. In *Proceedings of CHI Extended Abstracts on Human Factors in Computing Systems*, 93–98, ACM, New York, NY, USA.

Krishna, P. M.; and Srinivasa, K. G. 2012. Empirical Studies of Volunteer Collaboration in the Development of Free and Open Source Software: Analysis of Six Top Ranked Projects in SourceForge.net. *SIGSOFT Software Engineering Notes* 37(2): 1–11.

Madey, G. 2013. The SourceForge Research Data Archive (SRDA). University of Notre Dame, http://srda.cse.nd.edu/

Maguire, J. The SourceForge Story. *Datamation*, 2007. Available at http://www.datamation.com/cnews/article.php/3705731.

Meneely, A.; and Williams, L. 2011. Socio-Technical Developer Networks: Should We Trust Our Measurements? In *Proceedings of the Thirty Third International Conference on Software Engineering*, 281–290, ACM, New York, NY, USA.

Pohl, M.; and Diehl. S. 2008. What Dynamic Network Metrics can Tell Us about Developer Roles? In *Proceedings of the International Workshop on Cooperative and Human Aspects of Software Engineering*, 81–84, ACM, New York, NY, USA.

Robles G.; and Gonzalez-Barahona, J. M. 2006. Geographic Location of Developers at SourceForge. *In Proceedings of the 2006 international workshop on Mining software repositories*, 144–150, ACM, New York, NY, USA.

Sarma, A.; Maccherone, L.; Wagstrom, P.; and Herbsleb, J. 2009. Tesseract: Interactive Visual Exploration of Socio-Technical Relationships in Software Development. In *Proceedings of the Thirty First International Conference on Software Engineering*, 23–33, IEEE Computer Society, Washington, DC, USA.

Singh, P. V. 2010. The Small-World Effect: The Influence of Macro-Level Properties of Developer Collaboration Networks on Open-Source Project Success. *ACM Transactions on Software Engineering Methodology* 20(2): Article 6, 27 pages.

Xu, J.; Christley, S.; and Madey, G. 2006. Application of Social Network Analysis to the Study of Open Source Software. In *The Economics of Open Source Software Development*, Elsevier Press.

Yu L.; and Ramaswamy, S. 2007. Mining CVS Repositories to Understand Open-Source Project Developer Roles. In *Proceedings of the Fourth International Workshop on Mining Software Repositories*, Paper 8, IEEE Computer Society, Washington, DC, USA.

Zhang, W.; Yang, Y.; and Wang, Q. 2011. Network Analysis of OSS Evolution: An Empirical Study on ArgoUML Project. In *Proceedings of the Twelfth International Workshop on Principles of Software Evolution and the Seventh Annual ERCIM Workshop on Software Evolution*, 71–80, ACM, New York, NY, USA.