Open Source Software Developer and Project Networks

Matthew Van Antwerp[1] and Greg Madey[2]

[1] University of Notre Dame mvanantw@cse.nd.edu
[2] University of Notre Dame gmadey@cse.nd.edu

Abstract. This paper outlines complex network concepts and how social networks are built from Open Source Software (OSS) data. We present an initial study of the social networks of three different OSS forges, BerliOS Developer, GNU Savannah, and SourceForge. Much research has been done on snapshot or conflated views of these networks, especially SourceForge, due to the size of the SourceForge community. The degree distribution, connectedness, centrality, and scale-free nature of SourceForge has been presented for the network at particular points in time. However, very little research has been done on how the network grows, how connections were made, especially during its infancy, and how these metrics evolve over time.

1 Introduction to Complex Networks

The OSS network is defined as follows. Developers and projects are considered nodes in the graph. If a developer works on a project, there is an edge between the developer and that project. Since developers can only work on projects, the resulting graph will be bipartite, with developers and projects being the two groups having no edges within those groups. This bipartite graph can be easily transformed into a developer network or a project network. From the CVS database [8], users, groups, and timestamps were extracted. The timestamps are the dates (in unix time) of the oldest and most recent commits to that particular project. With this information, even if two users worked on the same project, a tie was only created between them if they worked on the project at the same time, i.e. their time frame windows overlapped.

1.1 Previous Work

Xu in her dissertation [13] analyzed many aspects of the developer and project networks in the SourceForge community. Xu examined the SourceForge developer network over time and determined it to be scale-free [12]. Xu also examined the community structure of the SourceForge developer network in [11] using metrics such as modularity [7, 6], identifying the largest communities and their populations. Gao examined the diameter, clustering coefficient, centrality, and other metrics of the SourceForge developer network over a timespan of a year and a half [3].

In [4], the authors apply social network analysis to CVS data, graphing network measurements such as degree distribution, clustering coefficient in modules, weighted clustering coefficient, and connection degree of modules for various projects at different time periods in the histories of Apache, Gnome, and KDE. They concluded that both the module network and the developer network exhibit small world behavior.

2 SourceForge, GNU Savannah, and BerliOS Developer

SourceForge was launched in November 1999. It is the world's largest OSS hosting site, with over 2.3 million registered users and over 180,000 projects at time of writing. It hosts numerous prominent and popular OSS projects. It is also the most studied hosting platform for the purposes of OSS research. SourceForge data is available at the SourceForge Research Data Archive (http://srda.cse.nd.edu) [9].

Many popular GNU/Linux utilities are or were at some time hosted at GNU Savannah, including gcc, emacs, libc, autoconf, automake, and make. The site has been up since around 1996, although many projects had their CVS logs imported and many of them date back to the early 1980s. They are strict about only hosting free software (SourceForge, for example, allows you to host a project that does not have a free software license). Many prominent OSS figures contribute to projects hosted here, including Richard Stallman, Ulrich Drepper, and Roland McGrath. Despite having far fewer developers than SourceForge, it is a very active community. Despite having not even 4000 developers, they have made nearly 5 million code commits. SourceForge has about 65 million code commits with orders of magnitude more members.

BerliOS Developer is a German website hosting 5,425 projects and 43,708 registered users at time of writing [2]. While difficult to tell exactly how old the site is, the BerliOS project itself was registered in June 2000 and the earliest CVS timestamp dates to 1996, although only a handful of projects have CVS commits dating prior to June 2000 and all of those projects were registered on BerliOS itself after June 2000. These projects likely had previously existing CVS archives imported into the BerliOS hosting platform. It is similar in functionality and services offered to SourceForge, but does not have the worldwide popularity of it. It is about as old as SourceForge as well, so the two share some similarities with BerliOS having a much smaller user base.

3 SourceForge Developer Network

73,829 users have made at least one CVS commit. Of those, 47,946 users are connected to at least one other developer (in other words, they are not the sole developer on all of the projects they work on), which is 64.94%. Of these connected users, the largest connected component contains 19,269 users, which is 40.19%, or 26.10% of all users who have made at least one commit. A visualization of a random sample of the developer network is found in figure 1. Just under

2/3 of the user-project ties were sampled to create this network. This resulted in 37,811 vertices in the network with the largest connected component containing 4687 vertices. This largest connected component is what is displayed in figure 1. There are many clusters of developers, but no central core in this sample. There are many "rings" of developers and towards the outside of the graph, there are linchpin developers where the graph would become disconnected without their presence. Visualizations were developed with Pajek [1].

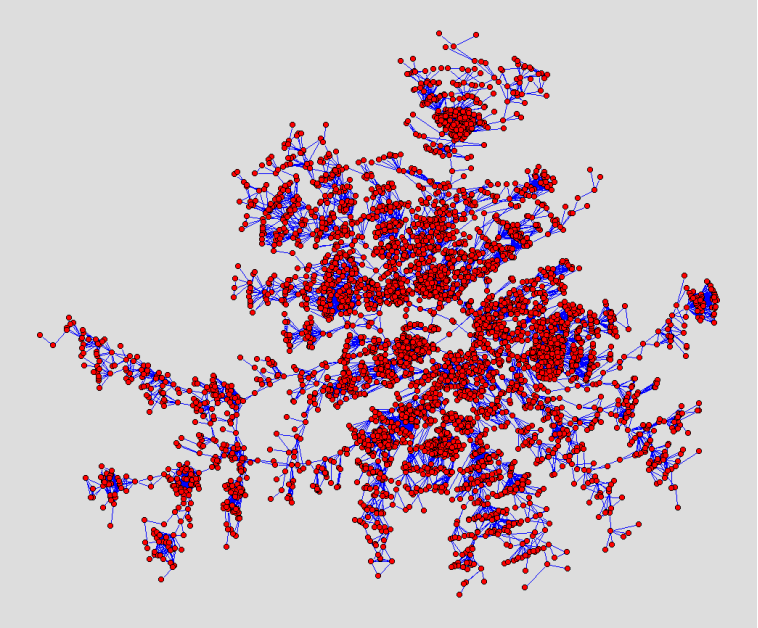

Fig. 1. Sample of the SourceForge developer network.

4 Savannah Developer and Project Networks

3889 users have made at least one CVS commit. Of those, 3042 users are connected to at least one other developer (they are not the sole developer on all of the projects they work on), which is 78.22%. Of these connected users, the largest connected component contains 1747 users, which is 57.42%, or 44.92% of all users who have made at least one commit. A visualization is provided in figure 2. In that figure, developers who are the sole developer on all projects they work on are excluded. They would be singletons in the network were they included. The Savannah project network can also be seen in figure 2. The network is well-connected with most projects in one large cluster. This is due to the long life of most Savannah projects, the rarity of new projects hosted at

Savannah, and the fact that many developers here work on multiple Savannah projects during their lifetime.

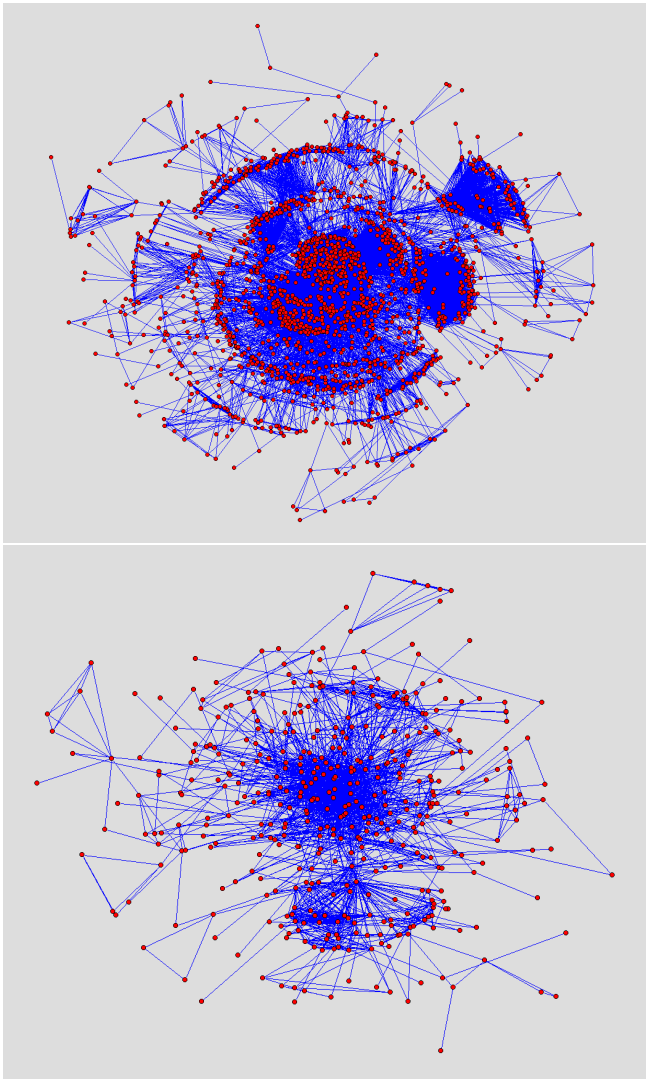

Fig. 2. (top) The largest connected component in the Savannah developer network visualized with the Kamada-Kawai algorithm for drawing graphs [5], a force-based algorithm. Distance between two nodes in the figure roughly corresponds with the length of the shortest path between them in the graph. (bottom) The Savannah project network.

5 BerliOS Developer and Project Networks

1582 users have made at least one CVS commit. Of those, 1113 users are connected to at least one other developer (they are not the sole developer on all of the projects they work on), which is 70.35%. Of these connected users, the largest connected component contains only 100 users, which is 8.98%, or 6.32% of all users who have made at least one commit. The BerliOS project network is mostly disconnected. There are however a handful of interesting cliques present in this network.

6 Repeat Network Connections

The SourceForge developer network, of 396,590 developer-developer ties, only 10,491 are duplicates or the original links that were later duplicated. This comprises only 2.65% of all developer pairs. However, for Savannah, of 46,937 developer pairs, there are 4620 pairs that are duplicates or links that were later duplicated, nearly 10%. For BerliOS, there are 3349 developer ties and 84 of them are repeats or the links that would later be duplicated. This is 2.51% of all pairs, comparable to SourceForge. The phenomena of repeat network connections in developer networks has not been extensively studied. The abundance of presumably fruitful developer ties in Savannah indicates that the projects here were likely successful. This also likely indicates that the typical project on Savannah is more successful than the typical project at SourceForge or BerliOS. This phenomena is examined further in [10].

7 Evaluation of the Communities

BerliOS is not a very globally connected developer community. While many developers are connected to someone else, there does not seem to be any sort of small-world effect in this network. SourceForge is a very large community and is better connected than BerliOS. About one quarter of all developers (CVS committers) in SourceForge are in the largest connected component. However, Savannah has nearly half of all developers in the largest connected component, an impressive aspect. A summary of the aforementioned statistics is available in table 1.

Table 1. Size of largest connected component in the developer networks

Hosting Site	**Total Size**	**Number Connected**	**% Connected**	**Largest CC**	**% of total**
SourceForge	73,829	47,946	64.94%	19,269	26.10%
Savannah	3889	3042	78.22%	1747	44.92%
BerliOS	1582	1113	70.35%	100	6.32%

8 Conclusions

We presented initial statistical analysis of the project and developer networks of three different OSS forges. The evolutionary trends displayed by these networks may offer crucial insight into OSS phenomena. Software versioning logs provide a great resource for building and studying these networks.

9 Acknowledgments

Research reported in the paper was supported in part by the National Science Foundation's CISE IIS-Digital Society & Technology program under Grant ISS-0222829 and by the National Science Foundation's CISE Computing Research Infrastructure program under Grant CNS-0751120

References

1. Vladimir Batagelj and Andrej Mrvar. Pajek - program for large network analysis. *Connections*, 21:47–57, 1998.
2. BerliOS Developer. http://developer.berlios.de.
3. Yongqin Gao. *Computational Discovery in Evolving Complex Networks*. PhD thesis, University of Notre Dame, 2007.
4. Luis Lopez-Fernandez Gregorio. Applying social network analysis to the information in cvs repositories. In *Proceedings of the First International Workshop on Mining Software Repositories (MSR 2004)*, Edinburgh, UK, 2004.
5. T. Kamada and S. Kawai. An algorithm for drawing general undirected graphs. *Inf. Process. Lett.*, 31(1):7–15, April 1989.
6. M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69:066133, 2004.
7. M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69:026113, 2004.
8. Matthew Van Antwerp. Studying open source versioning metadata. Master's thesis, University of Notre Dame, Notre Dame, IN, April 2009.
9. Matthew Van Antwerp and Greg Madey. Advances in the sourceforge research data archive. In *Workshop on Public Data about Software Development (WoPDaSD) at The 4th International Conference on Open Source Systems*, Milan, Italy, 2008.
10. Matthew Van Antwerp and Greg Madey. The importance of social network structure in the open source software developer community. In *The 43rd Hawaii International Conference on System Sciences (HICSS-43)*, Hawaii, January 2010.
11. J. Xu, S. Christley, and G. Madey. The open source software community structure. In *NAACSOS2005*, Notre Dame, IN, June 2005.
12. J. Xu and G. Madey. Exploration of the open source software community. In *NAACSOS 2004*, Pittsburgh, PA, June 2004.
13. Jin Xu. *Mining and Modeling the Open Source Software Community*. PhD thesis, University of Notre Dame, 2007.