Analysis and Modeling of Open Source Software Community

Yongqin Gao
Department of Computer Science and Engineering
University of Notre Dame
ygao1@nd.edu

Vince Freeh
Department of Computer Science
North Carolina State University
vin@cse.ncsu.edu

Greg Madey
Department of Computer Science and Engineering
University of Notre Dame
gmadey@nd.edu

Abstract

Open Source Software [8] (OSS) development is a classic example and prototype of collaborative social networks [4, 6, 19]. Based on the empirical data we collected from SourceForge over the last two years, we are able to investigate the structure and the dynamical mechanisms that determine the topology and govern the evolution of such systems. SourceForge is the largest online collaboratory for open source software development projects, hosting over 50 thousand projects and over 80 thousand developers. We used data on projects and developers that were collected monthly start in January 2001 [2, 3]. In this paper, we analyze the empirical data we collected from SourceForge to obtain statistics and topological information of the OSS developer collaboration network. We extract the parameters of the evolution by inspecting the network over time. We generate a model that depicts the evolution of this collaboration network.

Degree distribution, diameter and clustering coefficient are frequent attributes used to describe a network [13, 14, 15] and have been used ever since the foundation of random network theory. We also used these attributes to characterize the empirical data we collected from SourceForge. Existing research tends to look at the network a single snapshot in its evolution, which means they all based their observations on network without respect to time. We are able to inspect the network with consideration of time using the empirical data collected over more than two years.

- For diameter, which is the maximal shortest path in distance between any pairs of nodes, the SourceForge developer collaboration network has a value between 8 and 6. This is quite small considering to the network size of over 70,000 nodes. If we define D as the diameter of network and p is the connection probability, the scale of the diameter with respect to the fraction p is not fixed. For small p, D scales linearly with the system size, while for large p the scaling is logarithmic.
- For clustering coefficient, which is the fraction of the number of present links over the number of potential links among the nodes in its neighborhood, the value of SourceForge developer collaboration network is over 0.7 and is much bigger than that of a random network (the clustering coefficient of a random network of similar size is around 0.2 [9, 18]).
- For degree distribution $P(k)$, which is the distribution of the degree k throughout the networks, we find that it follows power law instead of normal distribution as shown in Figure 1. The left figure is the project distribution in normal coor-

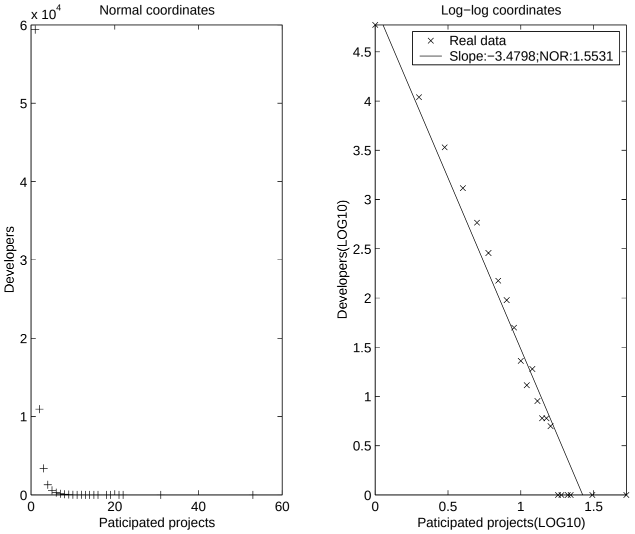

Figure 1: Degree distribution of SourceForge developer collaboration network.

dinates, and the right figure is the project distribution in log-log coordinates, where we can observe that the distribution fits a straight line well.

The previous attributes are all related to the topology of a collaboration network, which is studied by many researcheres [5, 12, 16, 20]. We also have some evolution related discoveries from these statistics we obtained from SourceForge.

- Cluster distribution. For a large complex network, it is normally composed of clusters, which are connected components of a network. Research about these clusters can help us to understand the mechanisms of complex network evolution. In our study, we find the cluster distribution of the SourceForge collaboration network also follows power law, and the size of major cluster of the network, which is the largest cluster in the network, increases toward fixed percentage (35%) of the overall network size and this percentage may be used to characterize the evolution of the network.

- Average degree. The average degree of the SourceForge network is also increasing. We cannot yet tell from the data is the average degree is approaching a limit, but it is increasing.

- Diameter. The diameter of the SourceForge developer collaboration network is used to measure the capability of the network to spread information. The smaller the diameter of a network, the faster information can spread through the network. We find the diameter of SourceForge is decreasing from 8 to 6 over two years, but the rate of decrease is diminishing.

- Fitness. Fitness is a new idea proposed by A.L. Barab$\acute{a}$si [7] to characterize the phenomenon that a new project can sometimes outrun the old one. We can find this kind of “fitness” effect in our SourceForge network. And also we observed in the empirical data on SourceForge that the average monthly growth of individual project is diminishing with time. This phenomenon can not be explained by the fitness proposed by A.L. Barab$\acute{a}$si, which is a constant value throughout time for each project. In a scale-free network with constant fitness, the node with higher degree will always have higher preference, which is based on fitness and link counts, and thus make its degree constantly increasing, which should often be decreasing based on the SourceForge empirical data. Constant fitness can not explain this phenomenon. We proposed a dynamic fitness factor to explain this phenomenon.

- Life cycle pattern. We find a life cycle like pattern in the SourceForge empirical data (the development patterns of individual developer and project). This can not be explained by the simple fitness.

Mathematical methods are not always able to capture all the features of a complex social network like SourceForge, especially the behaviors of heterogeneous individuals in the network. This is why we use agent-based simulation to help us understand and investigate such a complex network. We propose a model to describe the evolution of the SourceForge developer collaboration network, so that we can build simulations based on this model [1]. Since the collaboration network can be clearly described by a bipartite graph, we build our model based on a bipartite network. In order to remove the ambiguity of the model, we made several assumptions about the collaboration network. This assumption is based on the statistics we obtained from the empirical data.

The model contains two parts:

- Procedure definition: This is the description of the steps that the simulation should follow to reproduce the evolution of the SourceForge developer collaboration network.
- Simulation parameters: These are the parameters we use to control the simulation to reproduce the evolution of SourceForge developer collaboration network.

After defining the model, we use simulation to verify the topological and evolution related properties of SourceForge developer collaboration network we obtained by statistical study of the empirical data.

References

[1] Y. Gao, V. Freeh and G. Madey, *Conceptual framework for agent-based modeling and simulation*, submitted to NASOS 2003.

[2] G. Madey, V. Freeh and R. Tynan, *Agent-based modeling of open source using swarm*, 8^{th} Americas Conf. of Information Systems, 2002.

[3] G. Madey, V. Freeh and R. Tynan, *The Open Source Software development phenomenon: an analysis based on social network theory*, 8^{th} Americas Conf. of Information Systems, 2002.

[4] L. A. Adamic and B.A. Huberman, *Scaling behavior of the world wide web*, Science, 287(2115), 2000.

[5] Z. Neda, E. Ravasz, A. Schubert and A.L. Barab$\acute{a}$si, *Evolution of the social network of scientific collaborations*, PhysicA, 311(590), 2002.

[6] R. Albert and A.L. Barab$\acute{a}$si, *Dynamics of complex systems: scaling laws for the period of boolean networks*, Physics Reviews.

[7] R. Albert and A.L. Barab$\acute{a}$si, *Emergence of scaling in random networks*, Science, 286:509-512, 1999.

[8] G. Drummond, *Open source software and documents: a literature and online resource review*, 1999.

[9] P. Erd$\ddot{o}$s and A. R$\acute{e}$nyi, *On random graphs*, Publications mathematicae, 6:290-297, 1959.

[10] D. Hiebeler, *The swarm simulation system and individual-based modeling*, Advanced technology for natural resource management, 1994.

[11] P. J. Kiviat, *Simulation, technology, and the decision process*, ACM Tran. on modeling and computer simulation, 1(2):89, 1991.

[12] D. J. Watts, M.E.J. Newman, *Random graph models of social networks*, Physics reviews, 64(026118), 2001.

[13] M.E.J. Newman, *Scientific collaboration networks: I. network construction and fundamental resules*, Physics reviews, 64(016131), 2001.

[14] M.E.J. Newman, *Scientific collaboration networks: II. shortest paths, weighted networks, and centrality*, Physics reviews, 64(016131), 2001.

[15] M.E.J. Newman, *Clustering and preferential attachment in growing networks*, Physics reviews, 64(025102), 2001.

[16] D. J. Watts and S. H. Strogatz, *Collective dynamics of small-world networks*, Nature 393(440), 1998.

[17] Liuyuan Lu, W. Aiello and F. Chung, *Random evolution in massive graphs*, IEEE symposium of foundations of computer science, 2001.

[18] B. Bollob$\acute{a}$s, *Random graphs*, London:academic, 1985.

[19] R. Kumar, F. Maghoul, P. Raghavan and *et. al.*, *Graph structure in the web: experiments and models*, Computer networks, 33(309), 2000.

[20] M. Barthelemy, L.A.N. Amaral and *et. al.*, *Classes of small-world networks*, Proceedings of the national academy of sciences, 97(21), 2000.