Understanding OSS as a Self-Organizing Process

Greg Madey
University of Notre Dame
Computer Science & Engineering
University of Notre Dame
574-631-8752
gmadey@nd.edu

Vincent Freeh
University of Notre Dame
Computer Science & Engineering
University of Notre Dame
574-631-9131
vin@nd.edu

Renee Tynan
University of Notre Dame
Department of Management
University of Notre Dame
574-631-6764
rtynan@nd.edu

ABSTRACT
We hypothesize that open source software development can be modeled as self-organizing, collaboration, social networks. We analyze structural data on over 39,000 open source projects hosted at SourceForge.net. We define two software developers to be connected — part of a collaboration social network — if they are members of the same project, or are connected by a chain of connected developers. Project sizes, developer project participation, and clusters of connected developers are analyzed. We find evidence to support our hypothesis, primarily in the presence of power-law relationships on project sizes (number of developers per project), project membership (number of projects joined by a developer), and cluster sizes.

Categories and Subject Descriptors
D.2.2 [**Software Engineering**]: General.

General Terms
Management, Measurement

Keywords
Self-organization, emergence, open source, power laws

1. INTRODUCTION
Several research streams converge to provide us with a number of tools and models for analyzing the open source software movement: social network theory, small world phenomenon, power-laws, self-organization, and graph theory. Social network theory models persons as nodes of a graph and their relationships as edges of the graph [1-4]. Thus two persons are directly connected if they have a relationship (e.g., friendship) with each other; they then are one link away from one another. More distant relationships are modeled as paths through the graph; a "friend of a friend" is two links away. Several studies reveal an interesting phenomenon present in many of these social networks; most persons are very few links from any other person – the Small World Phenomenon [3, 4]. This idea was popularized in the play (and movie) *Six Degrees of Separation* [5] which claims that all persons in the world are at most six friendship links away.

2. SOCIAL NETWORK THEORY
Collaborative networks are variations of social networks, where the relationships are collaborations, e.g., actors in movies [3, 6], or co-authors on research papers [7, 8]. Often entire populations are connected into one large cluster with characteristic cluster coefficients, maximum degrees of separation (diameter) [3]. Highly prolific actors or authors are linchpins in collaborative networks. Linchpin actors or researchers play key roles in bridging disparate groups into one large cluster.

Social networks, collaborative networks, and other self-organizing systems (e.g., the Internet, WWW pages, U.S. firm sizes, cities, economic systems, word usage in languages, ecosystems) often have another interesting property; they have highly skewed distributions, which under a log-log transformation results in a linear relationship. This is called a power-law relationship. Power-law relationships have been reported for the Internet [9-13], sizes of U.S. firms [14], city size distributions [15], ecosystems [16], word rank in languages and writing [17] and many others.

Why such systems have power-law relationships is an open research question. Some speculate that self-organizing processes, when modeled as growing networks, display non-random attachment of nodes (sometimes called preferential attachment) [8, 12, 18].

We analyze the open source movement by modeling it as a collaborative social network. The developers are nodes of a graph and joint membership on an open source project is a collaborative link between the developers. The open source software development movement is highly decentralized and is a volunteer effort where developers freely join projects that they find appealing – all attributes of typical self-organizing systems. We hypothesize that the open source movement displays power-law relationships in its structure. Our empirical analysis of structural data collected from SourceForge suggests that this is the case. If this is supported by more detailed investigations, and as additional general theories are developed about social and collaborative networks (e.g., distributions in networks with non-random growth), that theory may then be applied to the open source software development process.

3. DATA COLLECTION
We gathered data monthly over the 14 month period from January 2001 through March 2002 at SourceForge, a web-based project support site sponsored by VA Software. SourceForge provides project management tools, bug tracking, mail list services, discussion forums, version control software for over 33,000 open source developers, participating on over 39,000 projects, as of February 2002 [19, 20]. We note that not all open source projects are registered with SourceForge; many high profile projects maintain their own developer sites, e.g., Apache, Perl, sendmail, Linux. But some large projects have moved to SourceForge (e.g.,

Samba) and we speculate that there are many smaller projects that have not joined SourceForge. Our assumption is that the projects at SourceForge are representative of the overall open source movement, in part because of its popularity and the large number of projects and developers registered there.

The primary data required for this research is a table consisting of records with two fields: project number and developer ID. Because projects can have many developers and developers can be on many projects, neither field is unique primary key. Thus the composite key composed of both attributes serves as a primary key. Each project in SourceForge has a unique project number. Additionally, each developer is assigned a unique ID when registering with SourceForge.

A web crawler traversed the SourceForge web server to collect the necessary data. All project home pages in SourceForge have a similar top-level design.

4. RESULTS

We model the OSS developers and projects as a network in two complementary ways. First, each developer is a node in the network; an edge exists between nodes if both developers are on the same project. This representation is analogous to movie actors as nodes and movies as links, or research paper authors as nodes and joint authorship as a link in the collaboration networks discussed above. The second way uses projects as nodes. Our initial analysis of the structural data shows that the developer collaboration network at SourceForge fits a power-law model, as determined by ordinary least squares (OLS) regression in log-log coordinates. As shown in Figure 1, both the project-size (number of developers on the project) and the number of projects per developer (total number of projects-joined by a developer) have power-law distributions. The solid line is the OLS regression line though the data, with an adjusted $R^2 = .93$ for the project-size data, and an adjusted $R^2 = .97$ for the projects-joined data. This power-law distribution is often a property of such self-organizing systems.

5. REFERENCES

1. Jin, E.M., Girvan, M., Newman, M. E. J., *The Structure of Growing Social Networks*, in *Santa Fe Institute Working Papers*. 2001: Santa Fe. p.1- 9.

2. Wasserman, S., K. Faust, *Social Network Analysis: Methods and Applications*. Structural Analysis in the Social Sciences, ed. M. Granovetter. 1999, Cambridge, UK: Cambridge University Press.

3. Watts, D., *Small Worlds*. 1999, Princeton: Princeton University Press.

4. Watts, D., Strogatz, S. H., *Collective Dynamics of Small-World Networks.* Nature, 1998. 393: p. 440-442.

5. Guare, J., *Six Degrees of Separation*. 1990, New York: Vintage Books.

6. Tjaden, B., *The Kevin Bacon Game*. 1996, http://www.cs.virginia.edu/oracle/.

7. Barabasi, A.L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., Viscek, T., *Evolution of the Social Network of Scientific Collaborations*. 2001.

8. Newman, M.E.J., *Clustering and Preferential Attachment in Growing Networks*, in *Santa Fe Institute Working Papers*. 2001: Santa Fe. p. 1-13.

9. Huberman, B.A., Adamic, L. A, *Growth Dynamics of the World Wide Web.* Nature, 1999. 401: p. 131.

10. Faloutsos, M., Faloutsos, P., Faloutsos, C. *On Power-Law Relationships of the Internet Topology*. in *SIGCOMM'99*. 1999. Cambridge, MA: ACM.

11. Albert, R., Jeong, H., Barabasi, A. L., *Diameter of the World Wide Web.* Nature, 1999. 401: p. 130-131.

12. Barabasi, A.L., Albert, R., *Emergence of Scaling in Random Networks.* Science, 1999. 286: p. 509-512.

13. Barabasi, A.L., Albert, R., Jeong, H, *Scale-free Characteristics of Random Networks: The Topology of the World Wide Web.* Physica A, 2000: p. 69-77.

14. Axtell, R.L., *Zipf Distribution of U.S. Firm Sizes.* Science, 2001. 293(5536): p. 1818-1820.

15. Pumain, D., Moriconi-Ebrard, F., *City Size Distributions and Metropolisation.* GeoJournal, 1997. 43(4): p. 307-314.

16. Jorgensen, S.E., Mejer, H., Nielsen, S. N., *Ecosystem as Self_organizing Critical Systems.* Ecological Modeling, 1998: p. 261-268.

17. Schroeder, M.R., *Fractals, Chaos, Power Laws*. 1991, New York: W. H. Freeman and Company.

18. Callaway, D.S., Hopcroft, J. E., Kleinberg, J. M., Newman, M. E. J., Strogatz, S. H., *Are Randomly Grown Graphs Really Random*, in *Santa Fe Institute Working Papers*. 2001: Santa Fe. p. 1-8.

19. SourceForge, *SourceForge Home*. 2002,

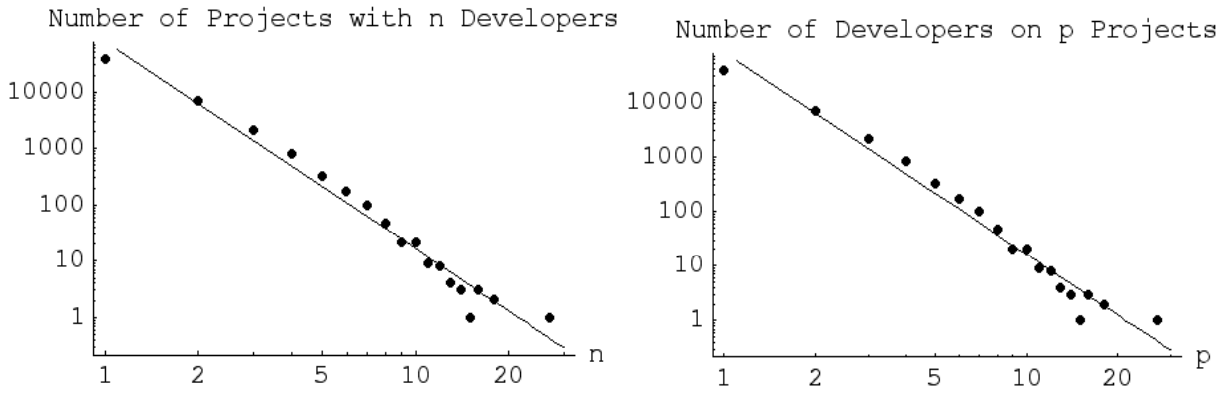

Figure 1: Power Law Relationships: OSS Project Size and Developer Project Membership

http://sourceforge.net/.

20. Wu, M.W., Lin, Y. D, *Open Source Development: An Overview*. IEEE Computer, 2001: p. 33-38.