Chapter 7.11
Modeling the Free/Open Source Software Community: A Quantitative Investigation

Gregory Madey
University of Notre Dame, USA

Vincent Freeh
North Carolina University, USA

Renee Tynan
University of Notre Dame, USA

ABSTRACT

In this chapter we summarize the latest results from an ongoing study examining Free/Open Source Software (F/OSS) Development communities as self-organizing systems. Using publicly available data about projects, developers, and their relationships at F/OSS hosting sites such as SourceForge, we have found the existence of several power-law relationships, which is consistent with the contention that F/OSS communities are self-organizing systems. The F/OSS community is modeled as a collection of ad hoc, social networks consisting of heterogeneous agents, self-organizing into projects and clusters of projects. A computer simulation of the F/OSS community model is developed using SWARM, an agent-based simulation toolkit. Empirical data is used to parameterize the simulation, which in turn is used to investigate a social psychological model of communication and team effectiveness in F/OSS projects.

INTRODUCTION

Our investigation aims to increase the understanding of the Free/Open Source Software (F/OSS) movement by providing a quantitative investigation of the network properties of the community. In some ways, the F/OSS movement is a prototypical example of a self-organizing system (Axelrod &

Cohen, 1999; Barabasi, 2002; Barabasi, Albert, & Jeong, 2000; Faloutsos, Faloutsos, & Faloutsos, 1999; Holland, 1998; Huberman & Adamic, 1999; Kuwabara, 2000; Madey, Freeh, & Tynan, 2002a, 2002b, 2002c), but it also possesses some unique properties that may affect the development of the network.

The lack of central planning or control in F/OSS projects challenges conventional economic assumptions, turns conventional software engineering and project management principles inside out, threatens traditional proprietary software business strategies, and presents new legal and governmental policy questions regarding software licensing and intellectual property. Understanding F/OSS is far from an academic enterprise—F/OSS is a major component of the IT infrastructure enabling global e-commerce. Free/open source software includes BIND, sendmail, Apache, Linux, INN, GNU utilities, MySQL, PostgreSQL, and Perl, all critical elements of the Internet.

In this chapter we describe a social network investigation of almost 60,000 F/OSS projects at SourceForge (2003), a web-based project support site sponsored by VA Software. With permission, we collected data on developers and projects over time at SourceForge. We analyzed the data using cluster analysis to learn more about the structure of the developer-project network, and then used the data to create a model of the network for agent-based simulations. We ran simulations of the network using the model to validate the model and to discover emergent properties of the network that can only be observed by studying the network growth over time.

We find that both project size and the number of projects on which developers are working can be modeled with the power-law relationship, providing empirical evidence for the claim that the F/OSS community is a self-organizing system. We also find that the cluster size of connected developers fits the power law, if the largest and most connected cluster, comprising almost 35% of the developers, is removed, and we discuss the possible causes behind this dual structural nature of the network. Finally, extending Barabasi's construct of a network fitness component (Barabasi, 2002), we find that a dynamic lifecycle fitness parameter for projects is necessary to best model the project data at SourceForge.

We begin with a discussion of social network theory and the utility of using simulation modeling to understand self-organizing systems. We then describe our data collection, cluster analysis results, model development and simulation results, followed by a discussion of the theoretical and practical ramifications of our results and directions for future research.

MODELING SOCIAL NETWORKS

Why should we invest the effort to do a quantitative simulation of the F/OSS network? The rationale behind such an investigation is that the F/OSS community is a social network that possesses several prototypical features of complex systems, systems that prior investigations have shown possess temporal and emergent properties that can be discovered only through modeling the system as a whole over time. For example, Axelrod (1984) found that certain types of "guarded cooperation" emerged as the most effective strategies for maximizing long-term joint outcome of dyads in a community, a result that could not have been obtained without simulation modeling.

Social network theory seeks to understand the network properties of people in relation to one another. Social network theory models persons as vertices or nodes of a graph and their relationships as edges or links of the graph (Barabasi, 2002; Jin, Girvan, & Newman, 2001; Wasserman & Faust, 1994; Watts, 1999; Watts & Strogatz, 1998). Thus two persons are directly connected if they have a relationship (e.g., friendship) with each other; they then are one edge away from one another, a distance of one. More distant relationships are modeled as paths through the