A TOPOLOGICAL ANALYSIS OF THE OPEN SOURCE SOFTWARE DEVELOPMENT COMMUNITY

Jin Xu, Yongqin Gao, Scott Christley & Gregory Madey
Department of Computer Science and Engineering
University of Notre Dame
Notre Dame, IN 46556

HICSS 38, Hawaii

January 2005

Partially Supported by NSF, CISE/IIS-Digital Science and Technology
INTRODUCTION

- **FLOSS: Complex/Self-organizing System**
  - Characteristic Structural/Topological Properties
  - Size distributions
  - Connectivity

- **Social Networks**
  - Small-world Models
  - Kevin Bacon Number
  - Erdos Numbers

- **Collaboration Networks**

- **Quantitative Structural/Topological Study of Project Communities**
  - Most developers participate on one project
  - Some developers participate on more than one project

- **Extension on prior study, using newer expanded data**
PREVIOUS RESEARCH

● Social Networks
PREVIOUS RESEARCH (cont.)

- Web Data Mining
PREVIOUS RESEARCH (cont.)

- Roles Classification
OSS COMMUNITY

- **User Group**
  - Passive Users: no direct attributable contribution in the data (downloads, user base, word-of-mouth publicity, etc.)
  - Active Users: bug reports, patch submissions, feature requests, help requests, etc.

- **Developer Group**
  - Peripheral Developer: *irregularly* contribute
  - Central Developer: *regularly* contribute
  - Core Developer: *extensively* contribute, manage CVS releases and coordinate peripheral developers and central developers.
  - Project Leader: guide the vision and direction of the project.
OSS DEVELOPMENT COMMUNITY

- Project Leaders
- Core Developers
- Co-developers
- Active Users

Jin Xu, University of Notre Dame
OSS SOCIAL NETWORK

- Two Entities: developer (community member) & project
- Project-Developer Network (bipartite graph)
  - Nodes: project, developer
  - Edges: developers in a project are connected to that project
PROJECT-DEVELOPER NETWORK

Jin Xu, University of Notre Dame
OSS SOCIAL NETWORK (cont.)

- **Project Network (unipartite graph)**
  - Node: project
  - Edge: two projects are connected if there is a developer participating on both.

- **Developer Network (unipartite graph)**
  - Node: developer
  - Edge: developers participating in a common project are connected to each other
OSS DEVELOPER NETWORK
DATA COLLECTION & EXTRACTION

- Data Source: SourceForge 2003 database dump, plus earlier “web-mined” data
- SourceForge.net approaching:
  - 100,000 registered projects
  - 1,000,000 registered users
- Project Leaders & Core Developers
  - Identification explicitly stored in data dump
- Co-developers & Active Users
  - Identification indirectly available
  - Forums: ask and/or answered questions
  - Artifacts: bug reports, patch submissions, feature requests, help requests, etc.
SUBSET OF THE DATABASE SCHEMA
Analysis: SourceForge.net Level

- Project Leaders (10%)
- Core Developers (5%)
- Co-developers (12%)
- Active Users (8%)
- Passive Users (65%)

- Project Leaders (28%)
- Core Developers (22%)
- Co-developers (34%)
- Active Users (16%)

Jin Xu, University of Notre Dame
### MEMBER DISTRIBUTION

<table>
<thead>
<tr>
<th>Project Size</th>
<th>Project Count</th>
<th>Project Leaders</th>
<th>Core Developers</th>
<th>Co-developers</th>
<th>Active Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤ 88</td>
<td>64847</td>
<td>47.8%</td>
<td>20.6%</td>
<td>19.8%</td>
<td>11.8%</td>
</tr>
<tr>
<td>&lt; 88</td>
<td>193</td>
<td>2.1%</td>
<td>5.7%</td>
<td>60.3%</td>
<td>31.7%</td>
</tr>
<tr>
<td>≥ 279</td>
<td>70</td>
<td>0.9%</td>
<td>2.7%</td>
<td>55.8%</td>
<td>40.6%</td>
</tr>
<tr>
<td>&lt; 279</td>
<td>193</td>
<td>2.1%</td>
<td>5.7%</td>
<td>60.3%</td>
<td>31.7%</td>
</tr>
</tbody>
</table>

Jin Xu, University of Notre Dame
TOPOLOGICAL PROPERTIES

- **Degree Distribution**
  - The total number of links connected to a node
  - Relative frequency of each index value
  - Power law distribution

- **Diameter**
  - The maximum longest shortest-path
  - The average longest shortest path

- **Cluster**
  - A social network consists of connected nodes

- **Clustering Coefficient**
  - The ratio of the number of links to the total possible number of links among its neighbors

Jin Xu, University of Notre Dame
FOUR SETS

- Subset A = \{ project leaders \}
- Subset B = \{ project leaders \} \cup \{ core developers \}
- Subset C = \{ project leaders \} \cup \{ core developers \} \cup \{ co-developers \}
- Subset D = \{ project leaders \} \cup \{ core developers \} \cup \{ co-developers \} \cup \{ active users \}

Prior work looked at Subset B only
DEVELOPER DEGREE DISTRIBUTION

Subset A

Subset B

Jin Xu, University of Notre Dame
DEVELOPER DEGREE DISTRIBUTION

Subset C

Subset D

Jin Xu, University of Notre Dame
# REGRESSION PARAMETERS

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Project-Network</strong></td>
<td>R-squared</td>
<td>0.9396</td>
<td>0.9704</td>
<td>0.6905</td>
</tr>
<tr>
<td></td>
<td>Slope</td>
<td>-3.5841</td>
<td>-2.6968</td>
<td>-1.3020</td>
</tr>
<tr>
<td><strong>Developer-Network</strong></td>
<td>R-squared</td>
<td>0.9870</td>
<td>0.9846</td>
<td>0.9469</td>
</tr>
<tr>
<td></td>
<td>Slope</td>
<td>-3.3747</td>
<td>-3.4676</td>
<td>-3.7793</td>
</tr>
</tbody>
</table>
Prior research suggests that OSS community network is a scale free network growing by two rules:
- Sequential addition of new developers
- Preferential attachment
- Related research on mechanisms that could plausibly generate observed topologies

All degree distributions are skewed

Most community members participate on 1 project

Linchpin members join multiple projects
- The largest number of communities a member joins increases from 29 (A) to 95 (D)
DIAMETER

- Diameter length
  - Subset A — N/A
  - Subset B — 10.2 out of 83118 members
  - Subset C — 2.7 out of 139570 members
  - Subset D — 2.7 out of 161691 members

- Project leader network is highly disconnected

- The degree of separation is significantly decreased with the participation of co-developers and active users.
CLUSTERS & CLUSTERING COEFFICIENT

<table>
<thead>
<tr>
<th>Property</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Largest cluster</td>
<td>737</td>
<td>15091</td>
<td>30794</td>
<td>40175</td>
</tr>
<tr>
<td>2nd largest cluster</td>
<td>197</td>
<td>34</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td># of clusters</td>
<td>43826</td>
<td>34280</td>
<td>27983</td>
<td>21659</td>
</tr>
<tr>
<td>Clustering coefficient</td>
<td>0.8406</td>
<td>0.8078</td>
<td>0.8867</td>
<td>0.8297</td>
</tr>
</tbody>
</table>

- All linked projects form a project cluster
- The largest cluster is much bigger than the 2nd largest cluster
- The largest cluster grows from A to D
- High clustering coefficient on all 4 subsets because members are fully collected in each project
A PROJECT CLUSTER
DISCUSSION & FUTURE WORK

- Small World Phenomenon
  - Small diameter & high clustering coefficient
- Scale Free
  - Power law distribution
- Effect of Co-developers & Active Users
- Agent-based Simulation
- More OSS Web Sites:
  - Apache, Bio.org, Savannah, etc.
THANK YOU