

# Understanding the Open Source Software Community

Presented by Scott Christley  
Dept. of Computer Science and  
Engineering  
University of Notre Dame

*Supported in part by National Science Foundation, CISE/IIS-Digital Society & Technology, under Grant No. 0222829*

# Contributors

- Vincent Freeh, Computer Science, North Carolina State University (Principal Investigator)
- Greg Madey, Computer Science & Engineering, University of Notre Dame (Principal Investigator)
- Renee Tynan, Department of Management, College of Business, University of Notre Dame (Principal Investigator)
- Jeff Bates, Acting Director of SourceForge.net, OSTG Inc. (Industrial Collaborator)
- Scott Christley, Computer Science and Engineering, University of Notre Dame (Doctoral Student)
- Yongqin Gao, Computer Science and Engineering, University of Notre Dame (Doctoral Student)
- Jin Xu, Computer Science and Engineering, University of Notre Dame (Doctoral Student)
- Jeff Goett, University of Notre Dame (REU Student)
- Chris Hoffman, University of Notre Dame (REU Student)
- Nadir Kiyancilar, University of Notre Dame (REU Student)
- Carlos Siu, University of Notre Dame (REU Student)

# GNU Open Source Software (OSS) Linux



- Free ...
  - to view source
  - to modify
  - to share
  - of cost



## Savannah

- Examples

- Apache
- Perl
- GNU
- Linux
- Sendmail
- Python
- KDE
- GNOME
- Mozilla
- Thousands more



# Unanswered Questions

- What is the motivation of the developers?
- Is this a new form of software development?
- Is this the future of work?
- Why do some projects “succeed” while others fail?
- ...and many more

# Data Set

- SourceForge.net
- Over 100,000 software projects and 1,000,000 registered users as of May 2005.
- Recent agreement between ND and SourceForge.net to get monthly data.
- <http://www.nd.edu/~oss>

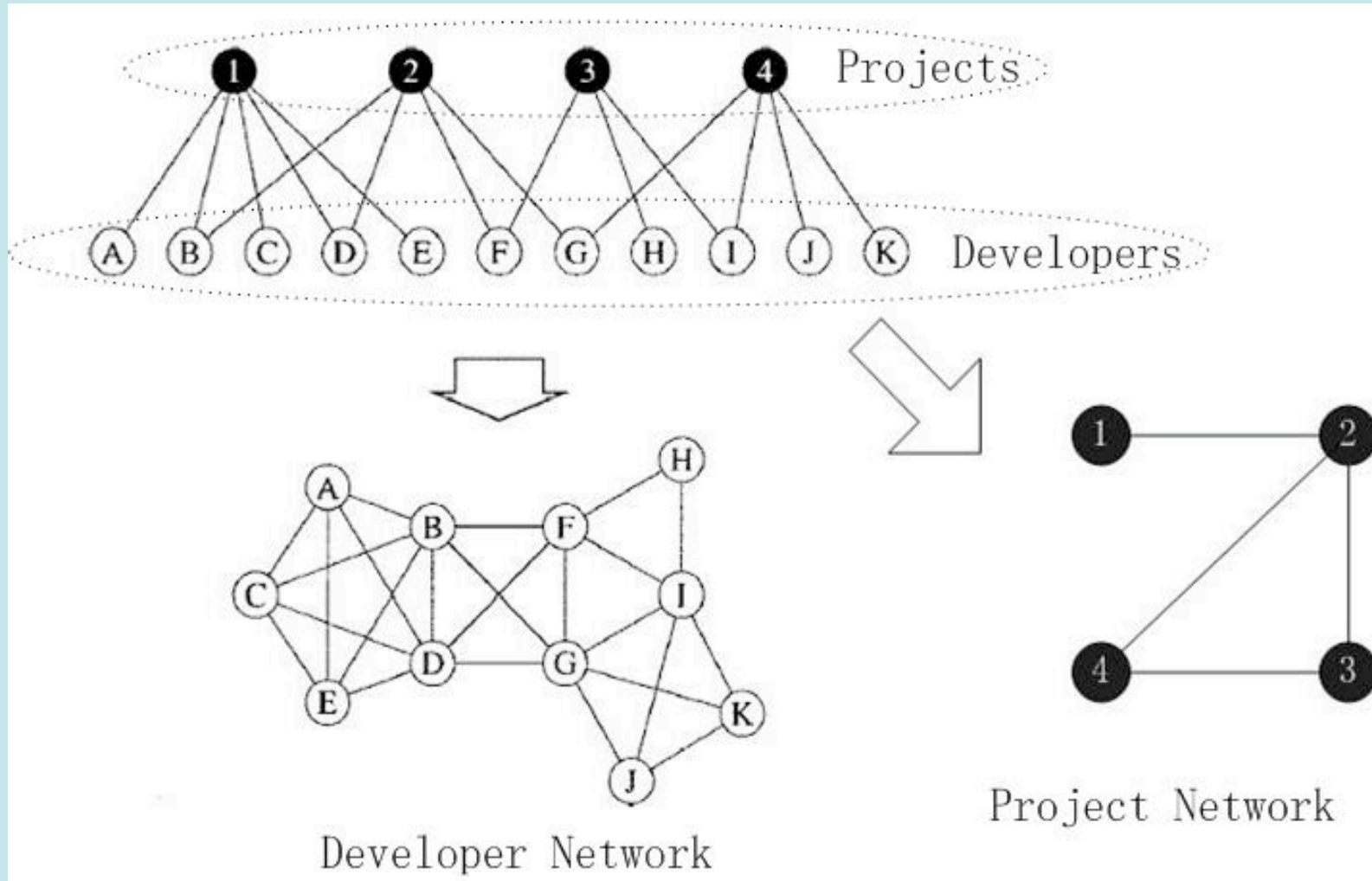
# Methodology

- Data Mining
- Network Topological Analysis
- Agent-based Simulation
- Social Network Analysis
- Public Goods Theory
- Future Directions

# Data Mining

- Gao, Huang, and Madey; NAACCSOS 2004
- Algorithms for prediction, categorization, clustering, and pattern finding.
- Computer scientists love this stuff!
- Just plug the data into the algorithms and out pops the answers, no social theory required!
- Conclusion-- Able to predict very well the failed projects but not the successful projects.
- Limitation-- Algorithms are not sophisticated enough for highly-(inter)dependent, temporal data produced by self-organizing phenomenon.

# Social Network



# Network Topology

- Xu, Gao, Christley and Madey; HICSS 2005
- Degree distribution, Connected components, Clustering coefficient, Diameter
- Small-world network
- Scale-free network (within a range)
- Conclusion-- Interesting global properties but gives little insight about the underlying mechanisms.

# Agent-based Simulation

- Gao and Madey, ADS 2005
- Simulate the growth and evolution of the social network.
- Projects and users are agents; rate of new projects and users join the community and/or projects calibrated with data set.
- Users join projects based upon preferential attachment.
- Conclusion-- Had to introduce the notion of project fitness to match the data set.
- Limitation-- Model of reality? ...Not exactly. Many models can produce the same global structures. Users don't have global knowledge about all projects.

# Social Network Analysis

- Xu, Christley and Madey; NAACSOS 2005
- Community structure, Betweenness, Assortativity (homophily) between projects within each community.
- Conclusion-- Over 1500 communities; mild assortativity between projects on attributes like OS, programming language.
- Limitations-- What do the communities mean?
- Future-- temporal analysis

# Public Goods Theory

- Christley and Madey, Agent 2004 Workshop
- Collective action out of mutual self-interest, jointness of supply, impossibility of exclusion, free rider phenomenon, connectivity, communality
- Characteristics of Individuals, Group, Environment; Action processes
- Conclusion-- Fits well as a descriptive model, All elements are dynamic and change over time, Critical mass is non-monotonic decision function, agent-based model as future work.
- Issues-- Calibration difficult, Time evolution and dynamics versus some “final result”. Complexity out of complexity.

# Future Directions

- Positional Analysis
  - Weighted, multi-relational (21) social network.
  - Approximate structural equivalence
  - Clustering and temporal analysis
- Hackman's model of team effectiveness

# Thank You!