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
  - Wasserman, S. Faust, K., *Social Network Analysis*, 1994.
  - Watts, D., *Small Worlds*, 1999.
  - Newman, M.E.J. "Clustering and Preferential Attachment in Growing Networks," 2001.
  - Barabasi, A. L., *Linked*, 2003
  - Gao Y. Q., Vincent F., Madey G., "Analysis and Modeling of the Open Source Software Community", *North American Association for Computational Social and Organizational Science (NAACSOS 2003),* Pittsburgh, PA, 2003.
  - Madey G., Freeh V., and Tynan R., "Modeling the F/OSS Community: A Quantitative Investigation," in *Free/Open Source Software Development*, ed., Stephan Koch, Idea Publishing, 2004.

PREVIOUS RESEARCH (cont.)

- Web Data Mining
  - Xu J., Huang Y., Madey G., "A Research Support System Framework for Web Data Mining", Workshop on Applications, Products and Services of Web-based Support Systems at the Joint International Conference on Web Intelligence (2003 IEEE/WIC) and Intelligent Agent Technology, Halifax, Canada, October 2003.
  - Howison J. and Crowston K.", "The perils and pitfalls of mining SourceForge", Proceedings of Mining Software Repositories Workshop, International Conference on Software Engineering (ICSE 2004), Edinburgh, Scotland, 2004.

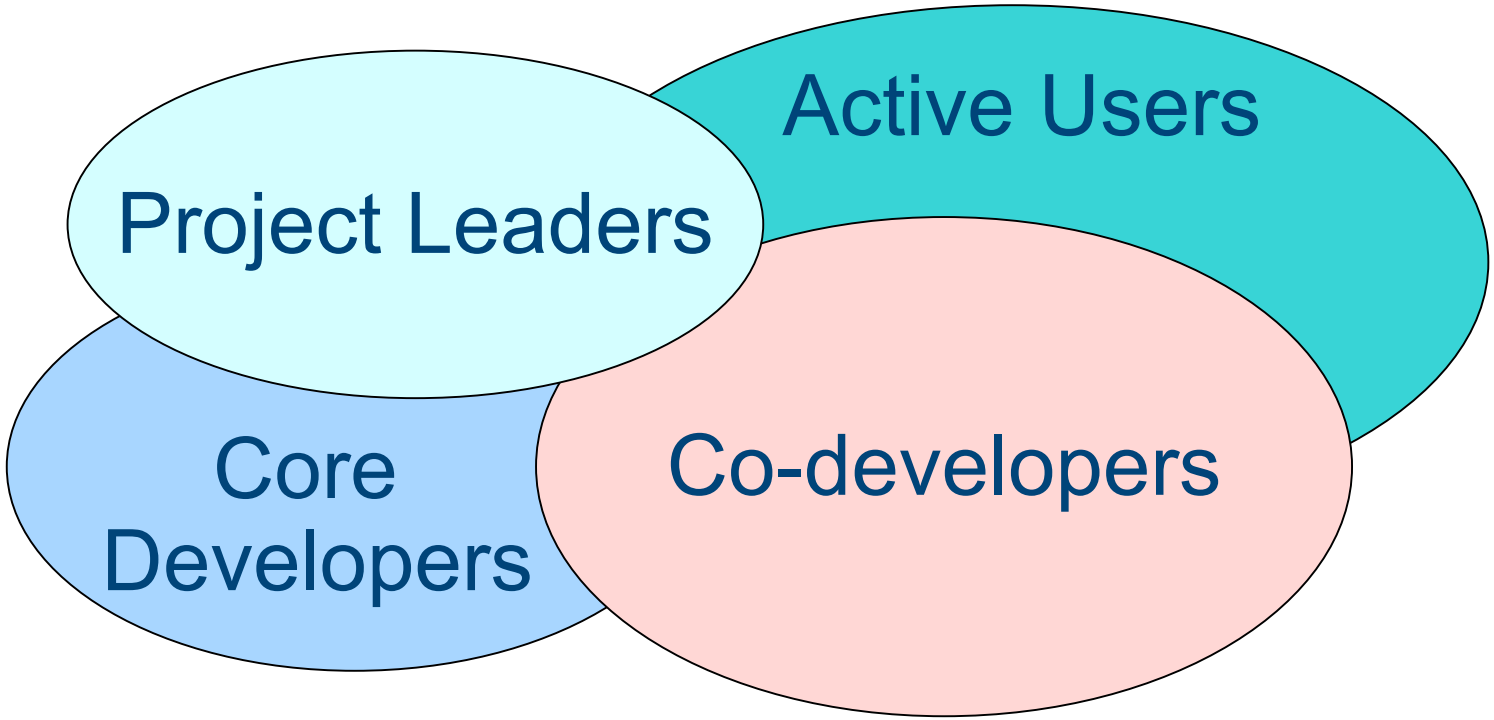
PREVIOUS RESEARCH (cont.)

- Roles Classification
  - Nakakoji K., Yamamoto Y., Kishida K., Ye Y., "Evolution Patterns of Open-source Software Systems and Communities", Proceedings of The International Workshop on Principles of Software Evolution, Orlando Florida, May 19-20, 2002.
  - Xu N., “An Exploratory Study of Open Source Software Based on Public Project Archives”, Thesis, The John Molson School of Business, Concordia University, Canada, 2003.

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



OSS SOCIAL NETWORK

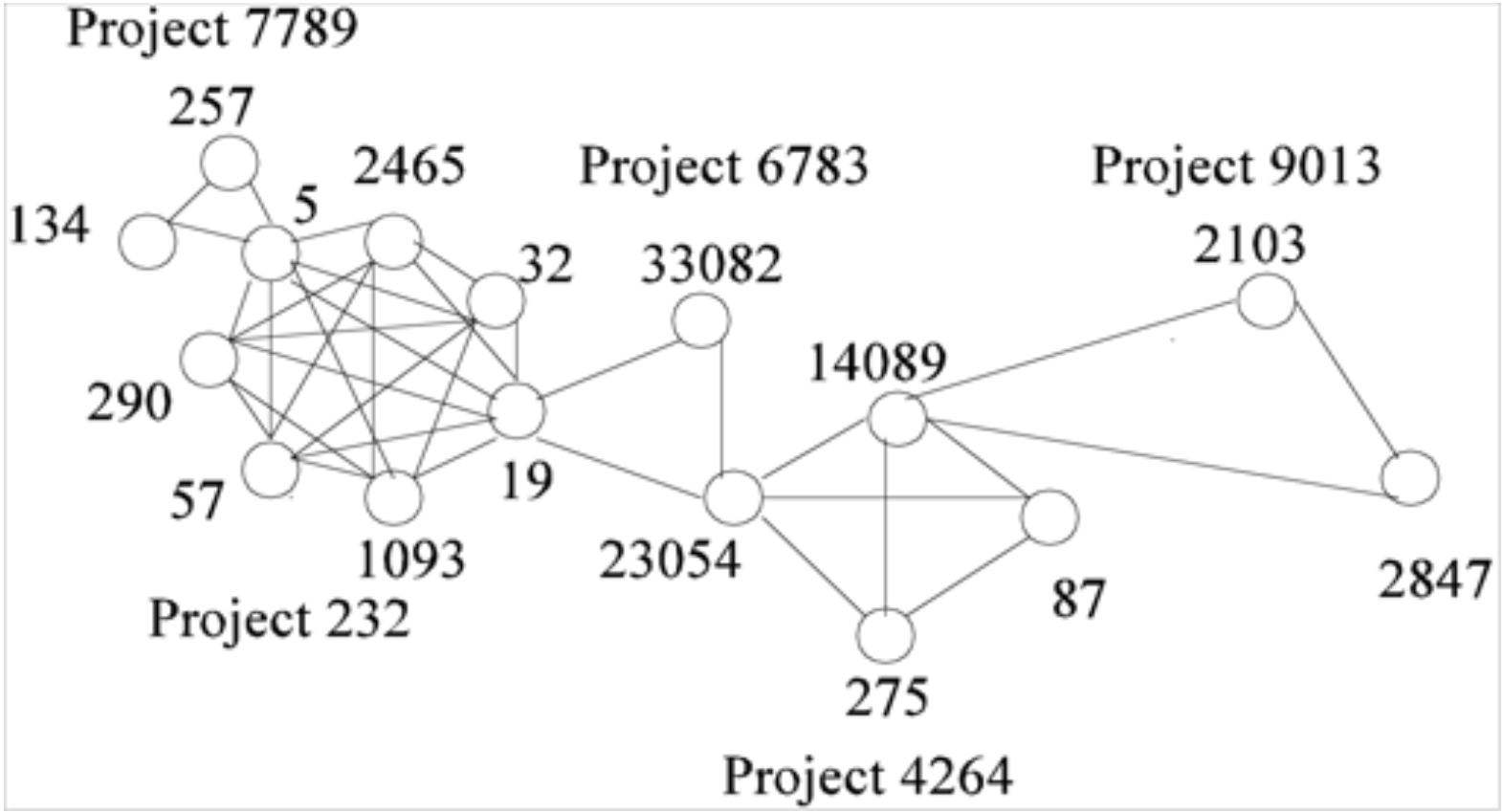
- Two Entities: developer (community member) & project
- Project-Developer Network (bipartite graph)
  - Nodes: project, developer
  - Edges: developers in a project are connected to that project

PROJECT-DEVELOPER NETWORK

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

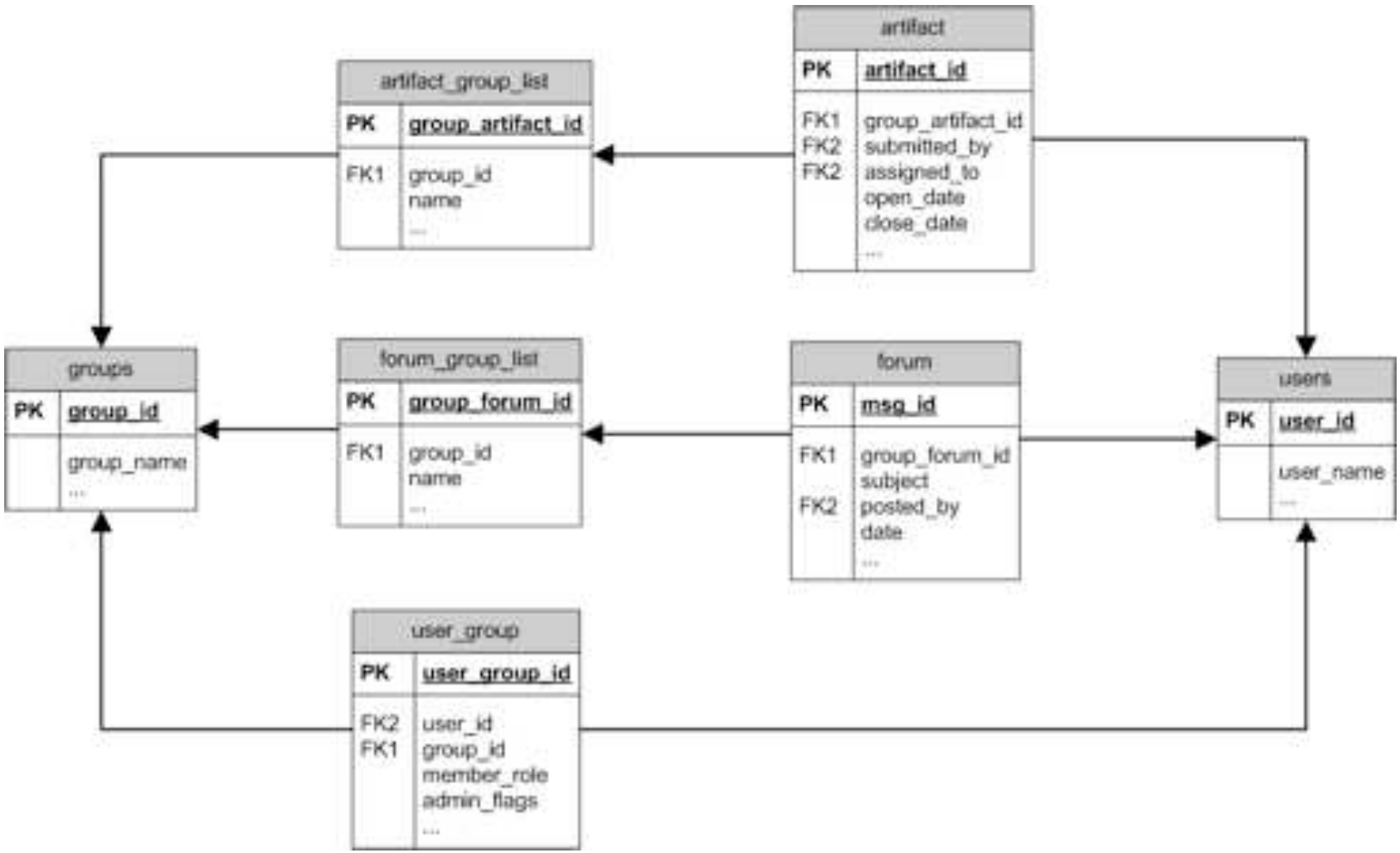
OSS DEVELOPER NETWORK



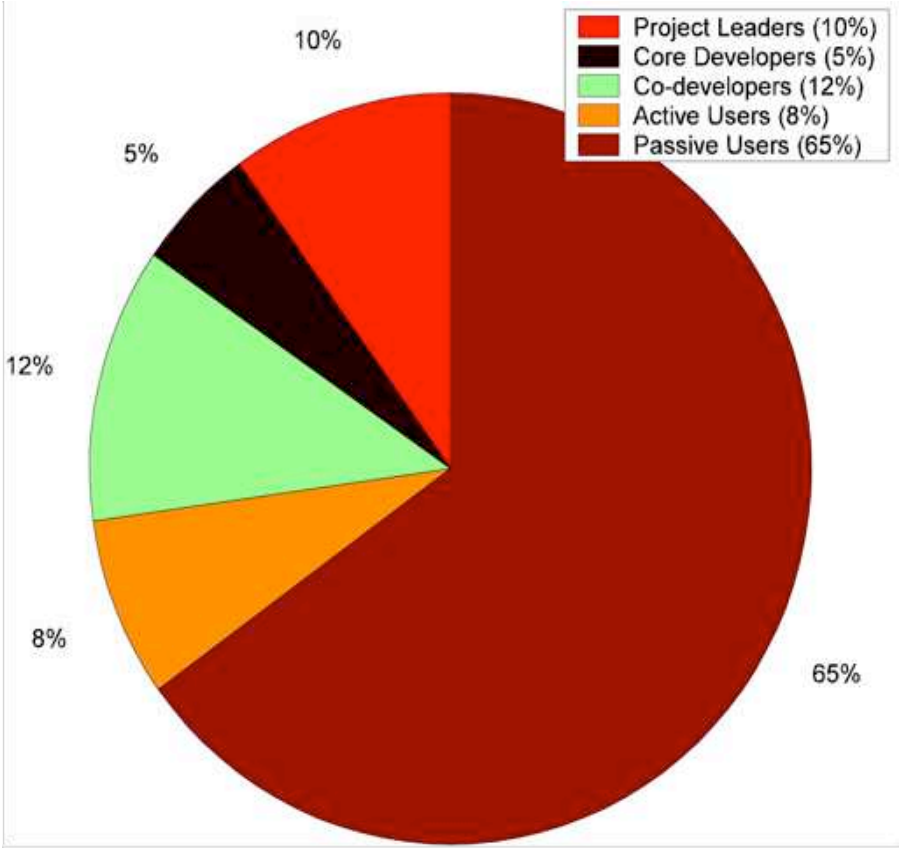
DATA COLLECTION & EXTRACTION

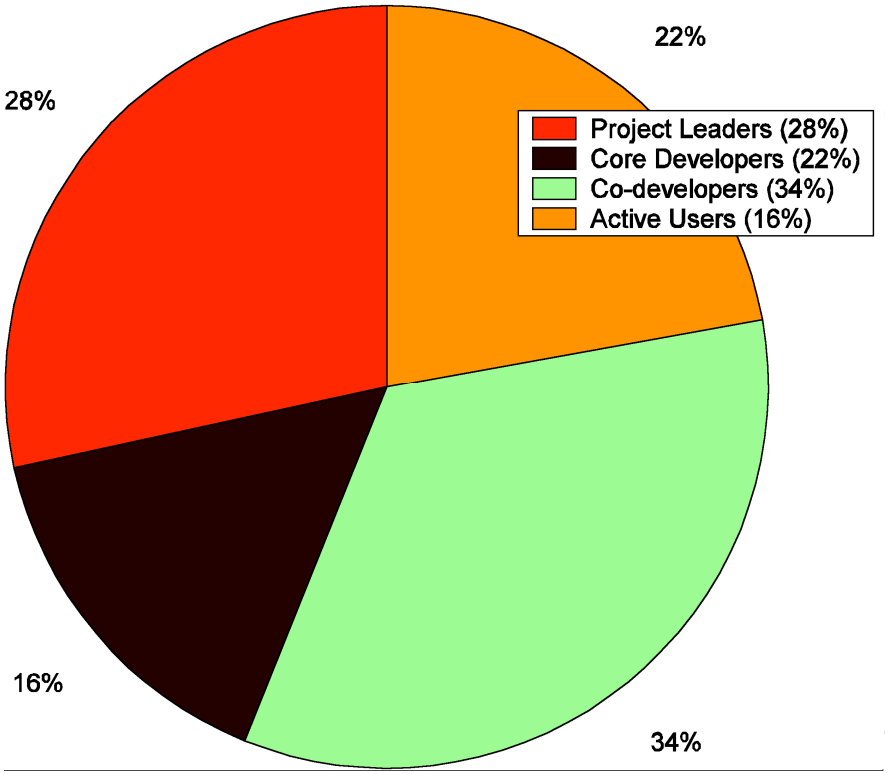
- Data Source: SourceForge 2003 database dump, plus earlier “web-mined” data
- SourgeForge.net approaching:
  - 100,000 registered projects
  - 1,000,000 registered users
- Project Leaders & Core Developers
  - Identification explicitly stored in data dump
- Co-developers & Active Users
  - Identification 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





MEMBER DISTRIBUTION

Project Size	Project Count	Project Leaders	Core Developers	Co-developers	Active Users
≤ 88	64847	47.8%	20.6%	19.8%	11.8%
< 88 ≥ 279	193	2.1%	5.7%	60.3%	31.7%
< 279	70	0.9%	2.7%	55.8%	40.6%

TOPOLOGICAL PROPERTIES

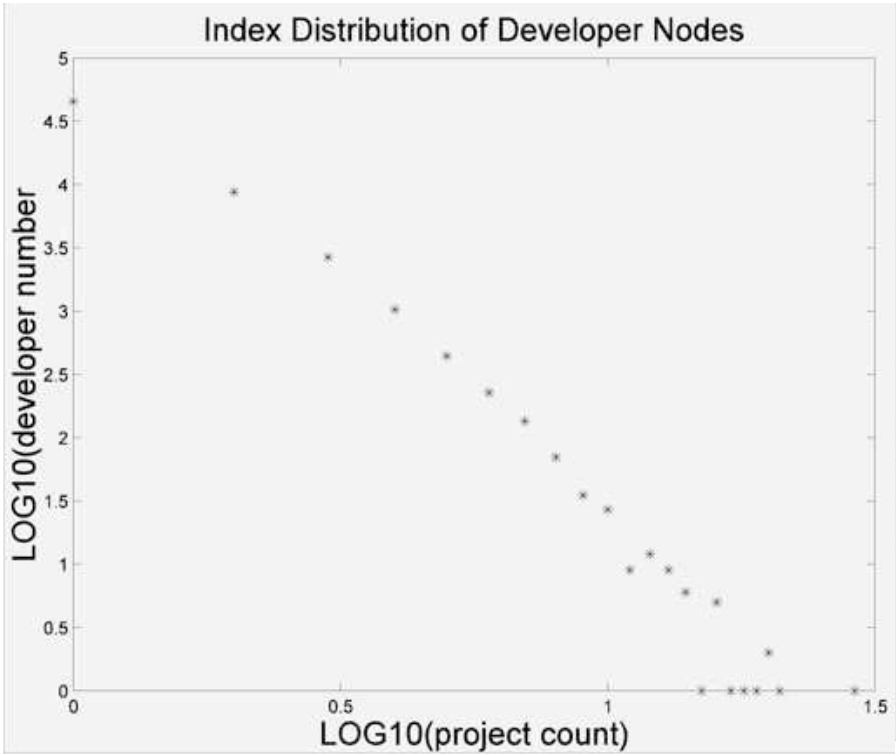
- Degree Distribution
  - The total number of links connected to a node
  - Relative frequency of each index value
  - Power law distribution
- Diameter
  - The maximum longest shortest-path
  - The average longest shortest path
- Cluster
  - A social network consists of connected nodes
- Clustering Coefficient
  - The ratio of the number of links to the total possible number of links among its neighbors

FOUR SETS

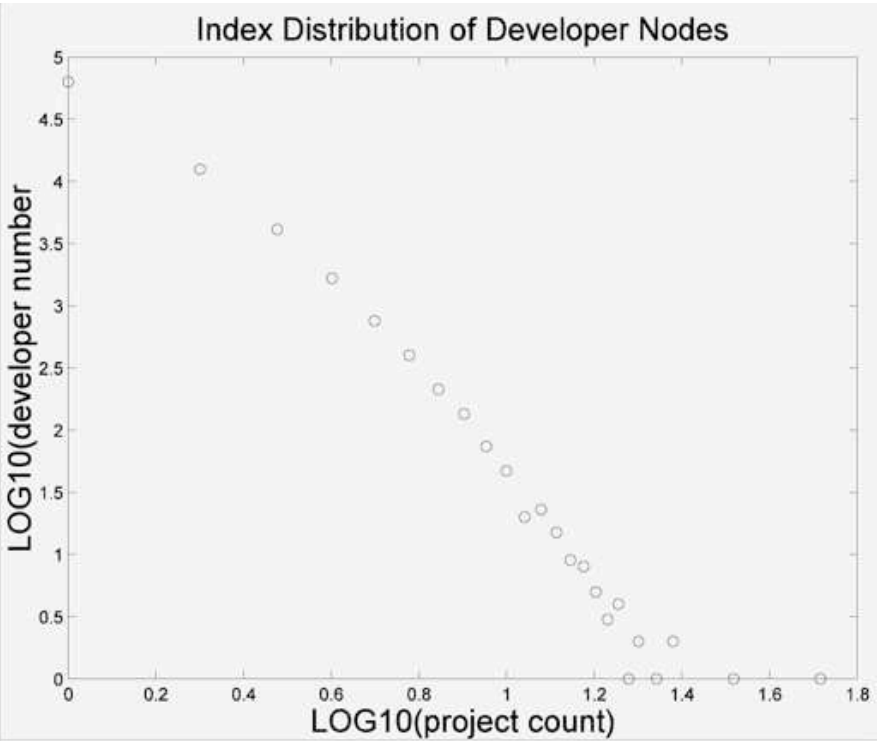
- Subset **A** = { project leaders}
- **Subset B = {project leaders} U {core developers}**
- Subset **C** = {project leaders} U {core developers} U {co-developers}
- Subset **D** = {project leaders} U {core developers} U {co-developers} U {active users}

Prior work looked at Subset **B** only

DEVELOPER DEGREE DISTRIBUTION

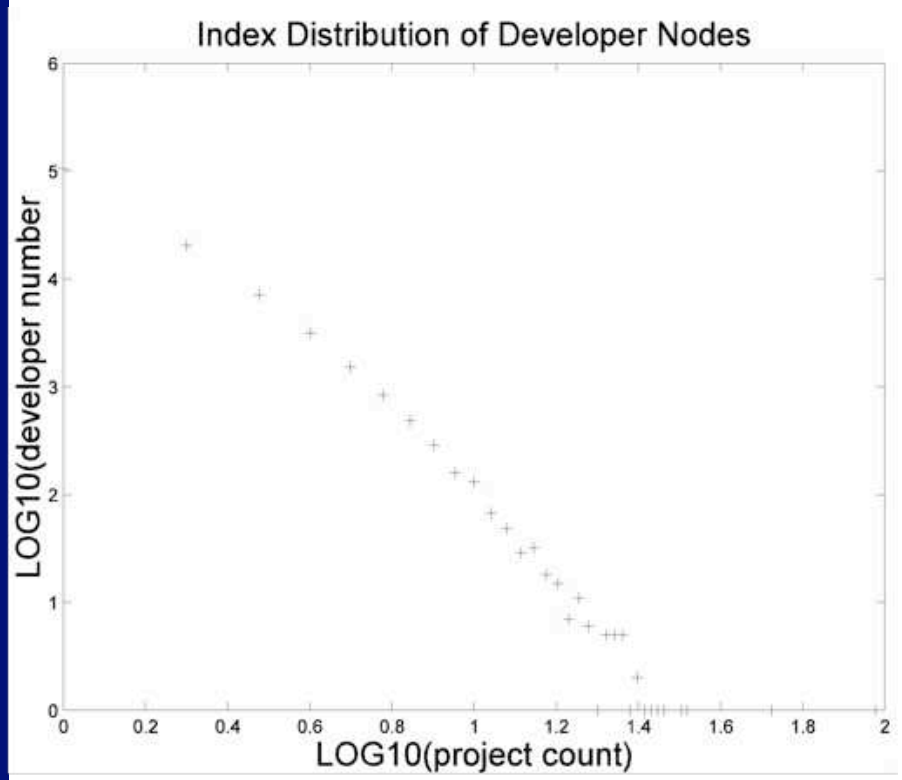


Subset A

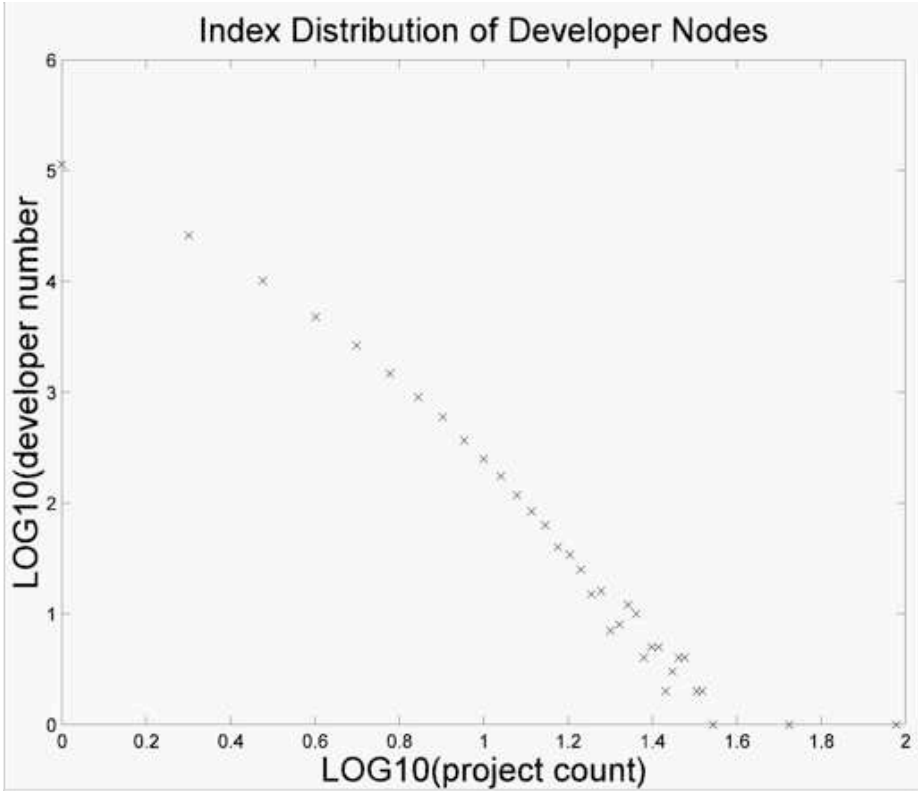


Subset B

DEVELOPER DEGREE DISTRIBUTION



Subset C



Subset D

REGRESSION PARAMETERS

		A	B	C	D
Project-Network	R-squared	0.9396	0.9704	0.6905	0.7221
	Slope	-3.5841	-2.6968	-1.3020	-1.2220
Developer-Network	R-squared	0.9870	0.9846	0.9469	0.9830
	Slope	-3.3747	-3.4676	-3.7793	-3.2743

DEGREE DISTRIBUTION

- 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

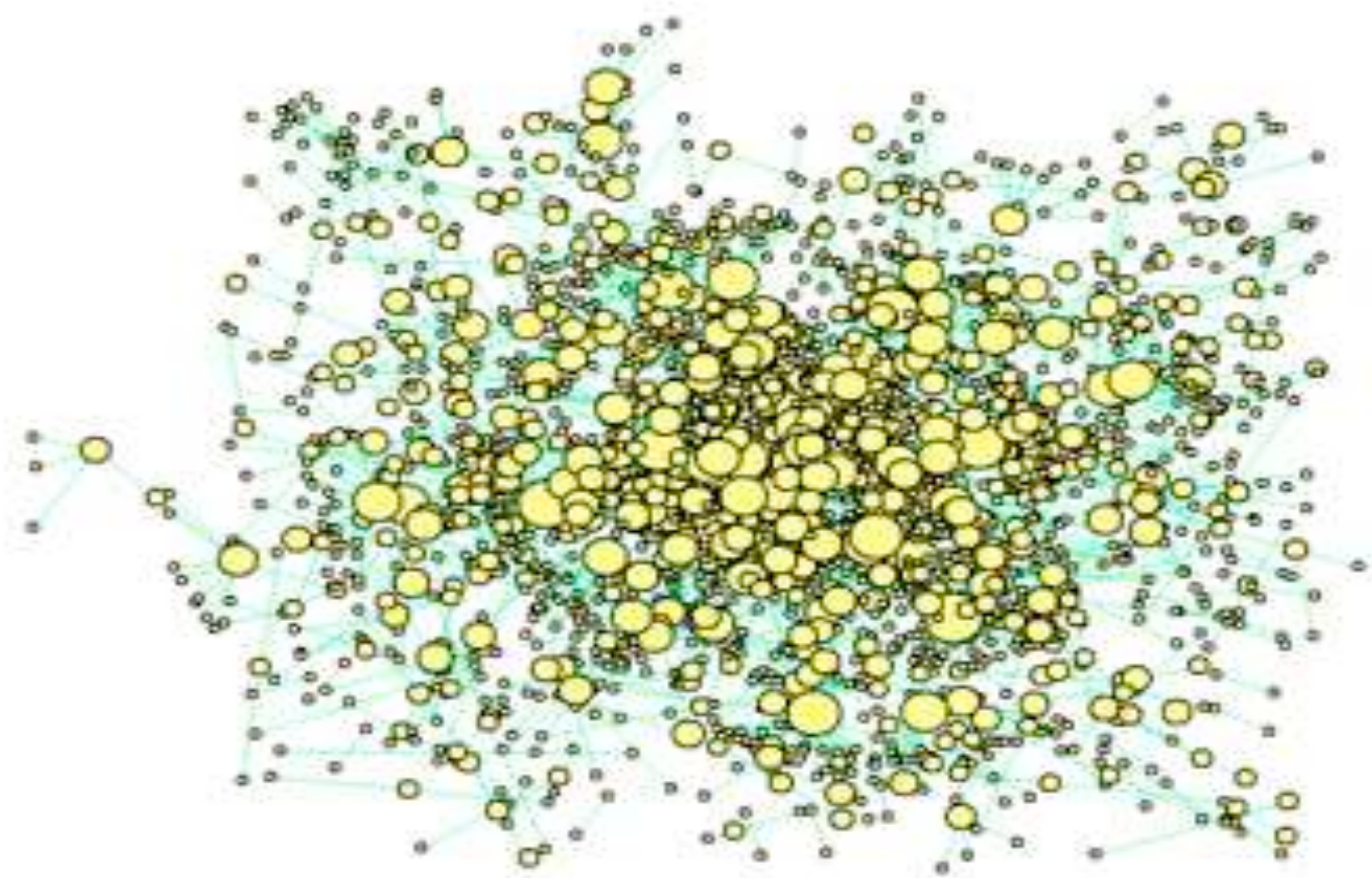
- Diameter length
  - Subset A — N/A
  - Subset B — 10.2 out of 83118 members
  - Subset C — 2.7 out of 139570 members
  - Subset D — 2.7 out of 161691 members
- Project leader network is highly disconnected
- The degree of separation is significantly decreased with the participation of co-developers and active users.

CLUSTERS & CLUSTERING COEFFICIENT

Property	A	B	C	D
Largest cluster	737	15091	30794	40175
2nd largest cluster	197	34	20	20
# of clusters	43826	34280	27983	21659
Clustering coefficient	0.8406	0.8078	0.8867	0.8297

- 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

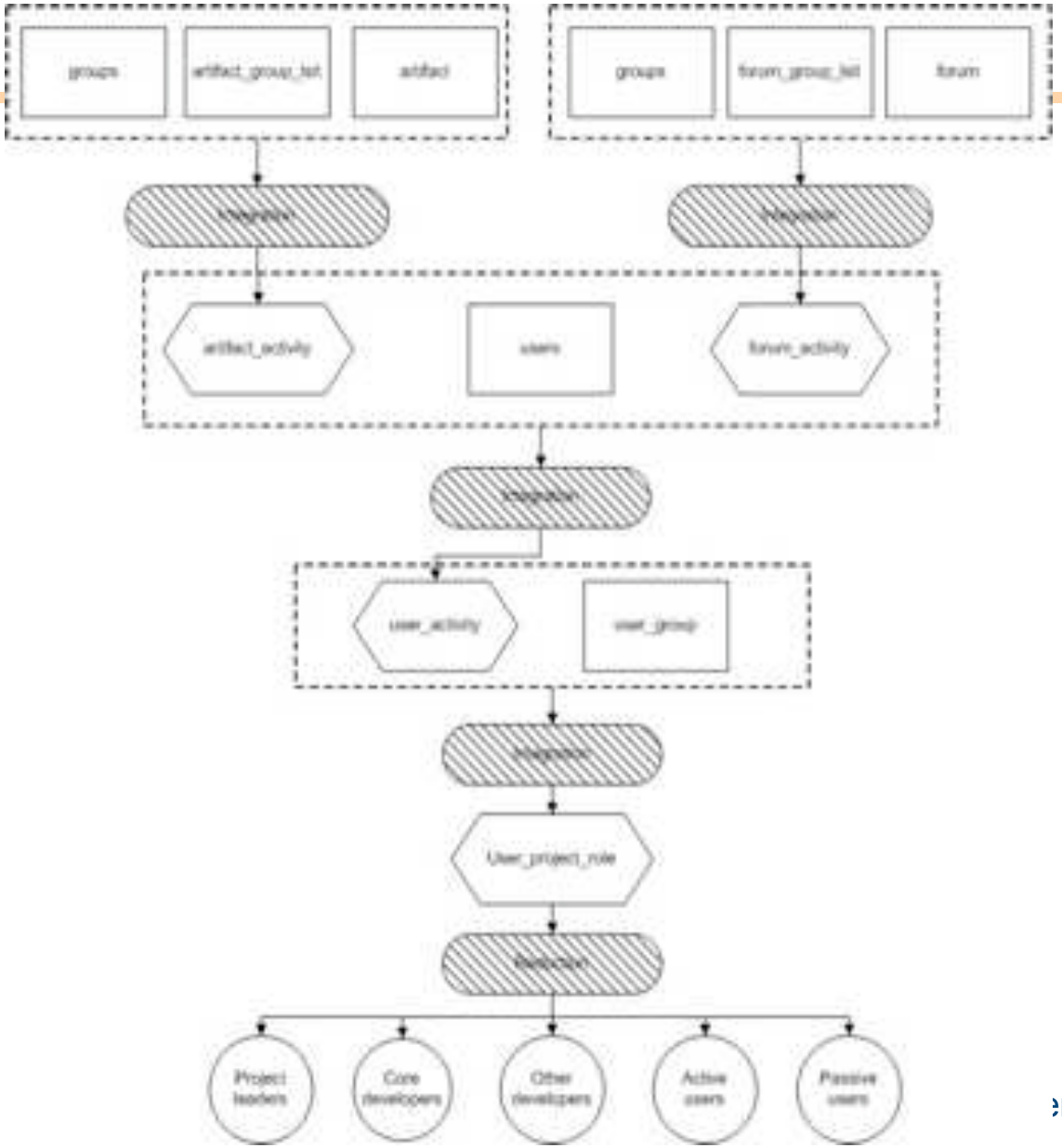


DISCUSSION & FUTURE WORK

- 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

Jin Xu, University of Notre Dame


groups
artifact_group_list
artifact
groups
forum_group_list
forum
artifact_activity
users
forum_activity
user_activity
user_group
User_project_role
Project leaders
Core developers
Other developers
Active users
Passive users