

Modeling and Simulation of the Open Source Software Community

Yongqin Gao
Greg Madey

Computer Science and Engineering,
University of Notre Dame



Supported in part by
the National Science Foundation – Digital Science & Technology

Outline



- Overview
- Modeling
- Agent-based Simulation Experiments
 - Adapted ER model
 - BA model
 - BA model with constant fitness
 - BA model with dynamic fitness
- Conclusions

Overview (about OSS)



- What is OSS?
 - Free to use, free to distribute
 - Unlimited usage
 - Source code freely available and modifiable
- Potential advantages over commercial software?
 - Higher quality
 - Faster development
 - Lower cost
 - Transparent

Open Source Software (OSS)

Linux



GNU

Savannah



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



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



Overview (about our research)



- Our goal
 - Understanding the OSS phenomenon using modeling and simulation
 - Agent-based simulation hypothesis-testing (in silico experiments)
- Approach
 - SourceForge is the source of our empirical data
 - Model as one of four types of social network
 - Use simulation to test hypotheses implied by the model

The Computer Experiment



The New York Times

Editorials/Op-Ed

March 4, 2003

- HOME
- JOB MARKET
- REAL ESTATE
- AUTOS
- NEWS

- International
- National
- Washington
- Business
- Technology
- Science
- Health
- Sports
- New York Region
- Education
- Weather
- Obituaries
- NYT Front Page
- Corrections

- OPINION
- Editorials/Op-Ed
- Columns
- Readers' Opinions

- FEATURES
- Arts
- Books
- Movies
- Travel
- NYC Guide
- Dining & Wine
- Home & Garden

SEARCH [Go to Advanced Search/Archive](#)

Past 30 Days

MEMBER CENTER

Welcome, [gmadey](#)

The Real Scientific Hero of 1953

By STEVEN STROGATZ

THACA, N.Y.

Last week newspapers and magazines devoted tens of thousands of words to the 50th anniversary of the discovery of the chemical structure of DNA. While James D. Watson and Francis Crick certainly deserved a good party, there was no mention of another scientific feat that also turned 50 this year — one whose ramifications may ultimately turn out to be as profound as those of the double helix.

In 1953, Enrico Fermi and two of his colleagues at Los Alamos Scientific Laboratory, John Pasta and Stanislaw Ulam, invented the concept of a "computer experiment." Suddenly the computer became a telescope for the mind, a way of exploring inaccessible processes like the collision of black holes or the frenzied dance of subatomic particles — phenomena that are too large or too fast to be visualized by traditional experiments, and too complex to be handled by pencil-and-paper mathematics. The computer experiment offered a third way of doing science. Over the past 50 years, it has helped scientists to see the invisible and imagine the inconceivable.

- E-Mail This Article
- Printer-Friendly Format
- Most E-Mailed Articles

ARTICLE FIELD
SPONSORED BY

STARBUCKS.COM

TIMES NEWS TRACKER

Topics

Fermi, Enrico

DNA (Deoxyribonucleic Acid)

Science and Technology

Create Your Own | Manage Alerts
Take a Tour

[Sign Up for Newsletters](#)

Alerts

Create

Create

Create

Overview (about SF)



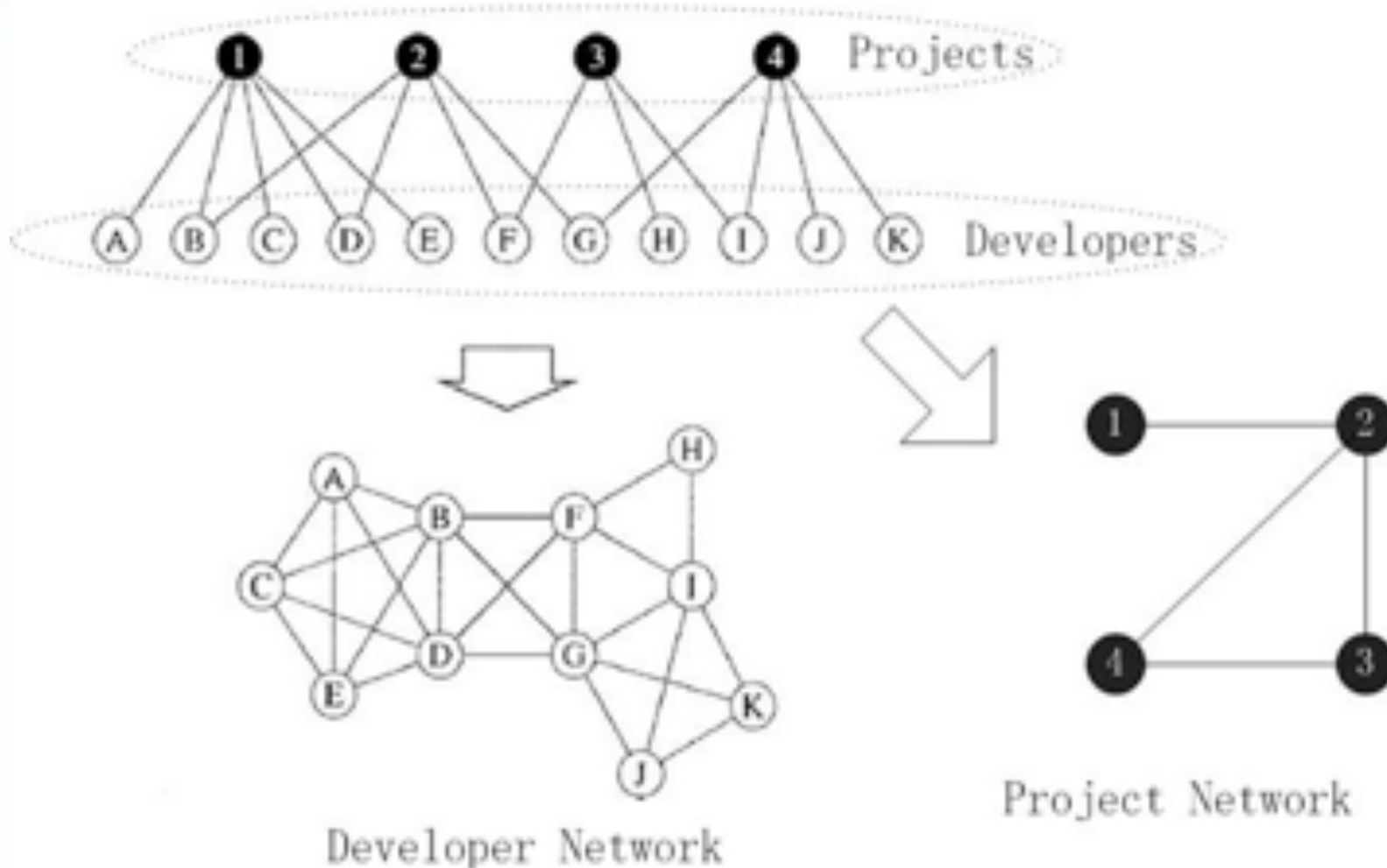
- VA Software
- Part of OSDN
- Started 12/1999
- Collaboration tools
- 97,950 Projects
- over 186,000 Developers
- 1,041,585 Registered Users

Modeling as Collaboration Network



- What is a collaboration network?
 - A social network representing the collaborating relationships.
 - Movie actor network and scientist collaboration network
- Difference of SourceForge collaboration network
 - Link detachment
 - Virtual collaboration
 - Voluntary
 - Global
- Bipartite property of collaboration networks

Collaboration network - bipartite



Adapted from Newman, Strogatz and Watts, 2001

Network Statistics Inspected



- Diameter
- Average degree
- Clustering coefficient
- Degree distribution
- Fitness and life cycle

Terminology



- Diameter
 - Average length of shortest paths between all pairs of vertices
- Degree
 - The count of edges connected to given vertex
- Average degree
 - Average of the degrees of all vertices in the network
- Cluster
 - The connected components of the network
- Clustering coefficient (CC)
 - CC_i : Fraction representing the number of links actually present relative to the total possible number of links among the vertices in its neighborhood.
 - CC: average of all CC_i in a network
- Degree distribution
 - The distribution of degrees throughout a network

Agent-based Modeling

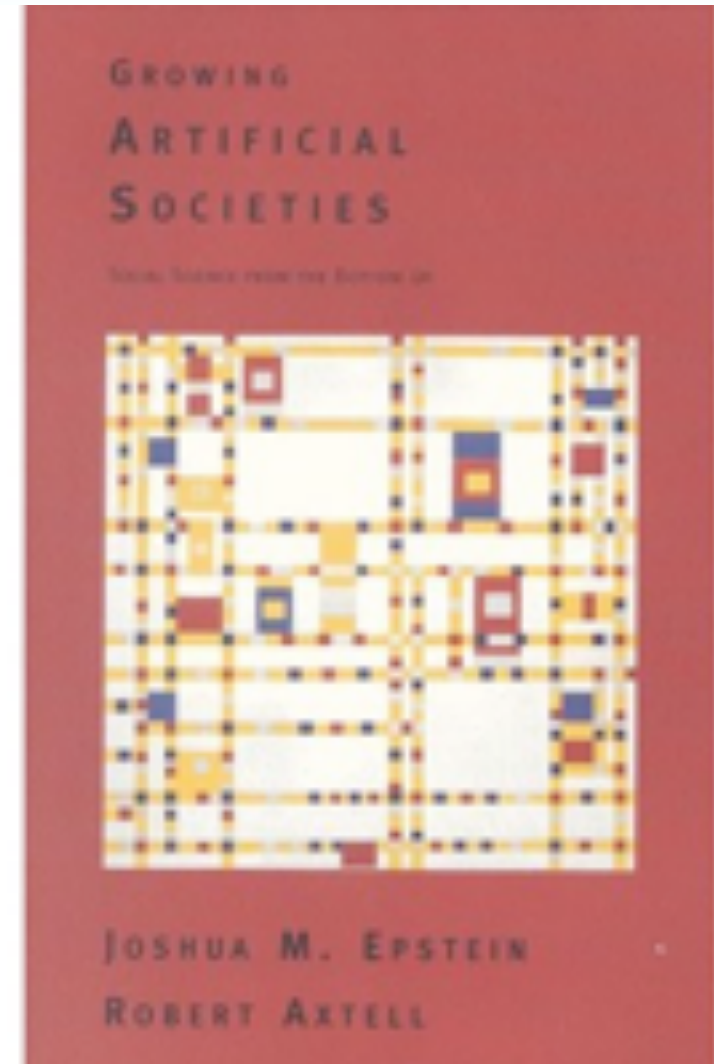


- EBM vs. ABM
 - Heterogeneous individuals
 - Complex network
- Experience environment
 - Hardware: computer cluster
 - Software:
 - Simulation toolkits: Swarm
 - Database: Oracle
 - Language: Java, PL/SQL

Model of SourceForge



- ABM based on bipartite graph
- Model description
 - Agent: developer
 - Behaviors: Create, join, abandon and idle
 - Preference: developer's and project's
 - Fitness
- Four models / four hypotheses
 - ER, BA, BA with constant fitness and BA with dynamic fitness
- Comparison of observed and simulated social networks

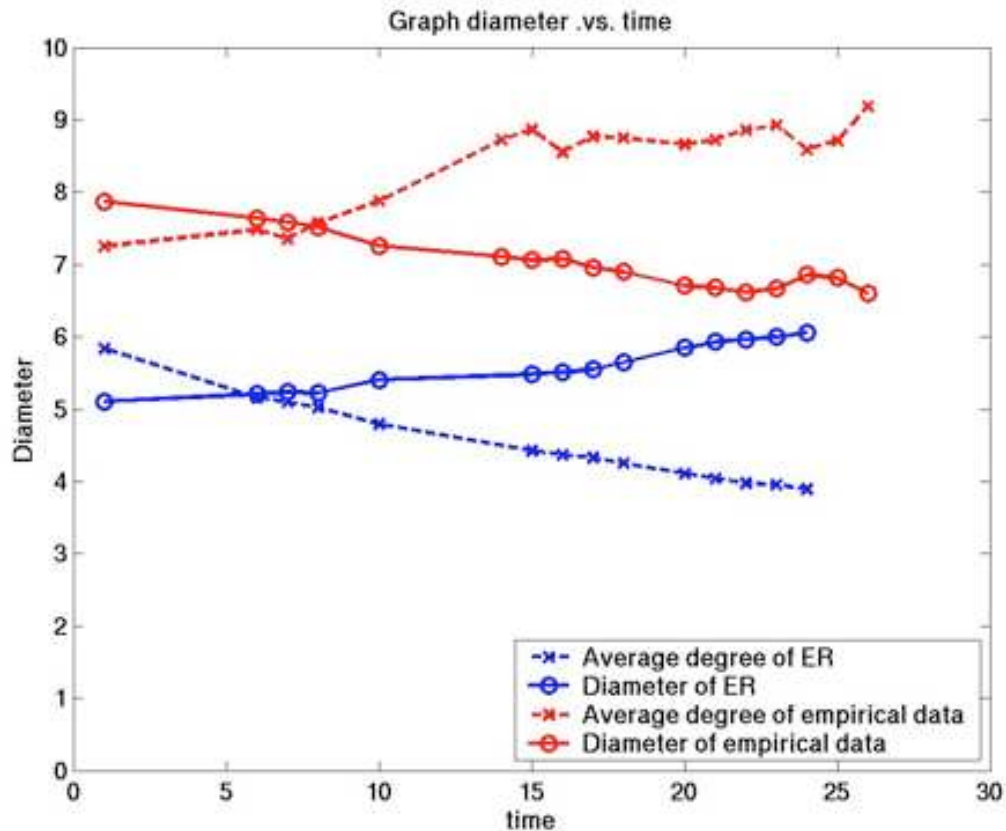


Models/Hypotheses



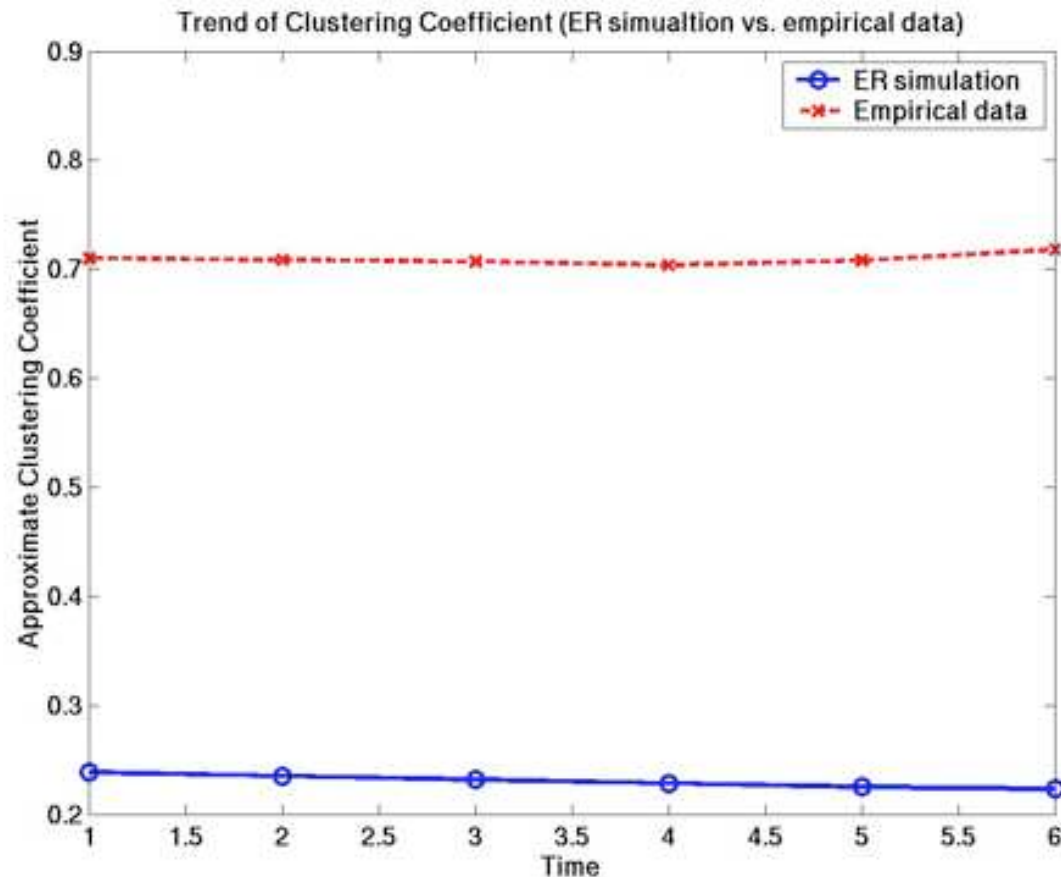
- **Hypothesis One: The developer network is a random network (modified ER model)**
 - This is the “all-projects-are-created-equal” model
- **Hypothesis Two: The developer network is a scale-free network (BA model)**
 - This is the “rich-get-richer” model.
- **Hypothesis Three: The developer network is a scale-free network with fitness (BA model with fitness)**
 - This is the “not-all-projects-are-born-equal” model.
- **Hypothesis Four: The developer network is a scale-free network with dynamic fitness (BA model with dynamic fitness)**
 - This is the “projects-have-a-life-cycle” model.

ER Model - Diameter



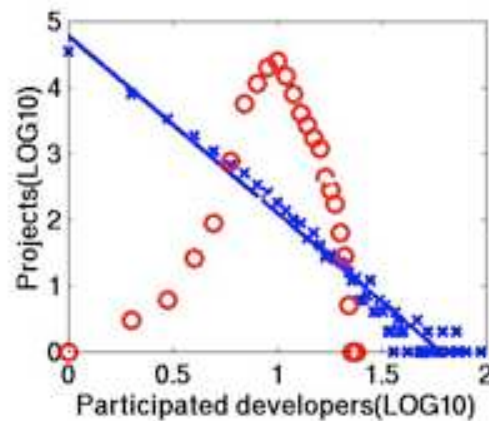
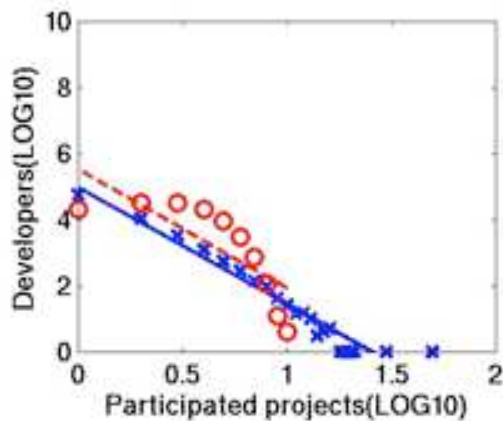
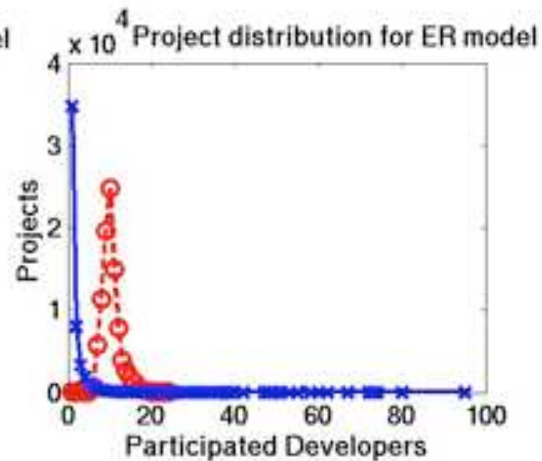
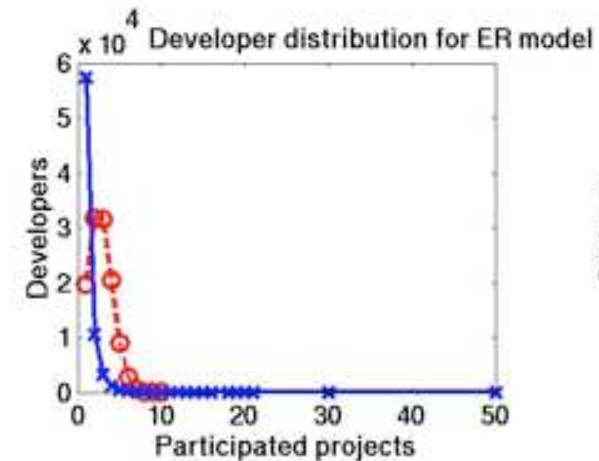
- Average degree is decreasing while it is increasing in empirical data
- Diameter is increasing while it is decreasing in empirical data

ER Model – Clustering Coefficient



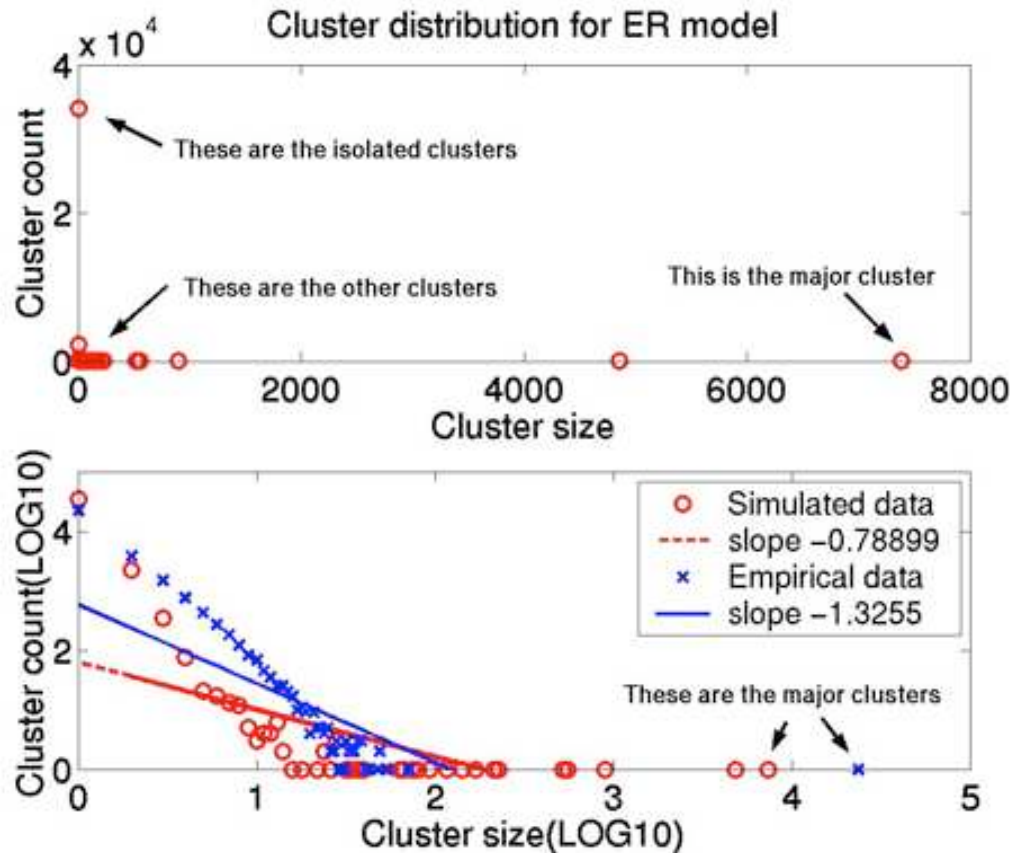
- Clustering coefficient is relatively low under 0.3 while it is around 0.7 in empirical data.

ER Model – Degree Distribution



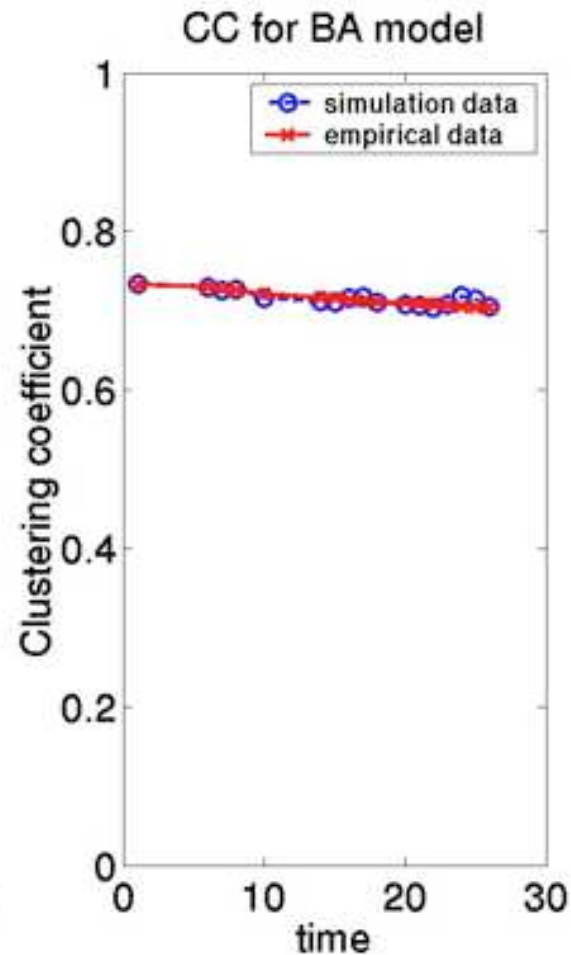
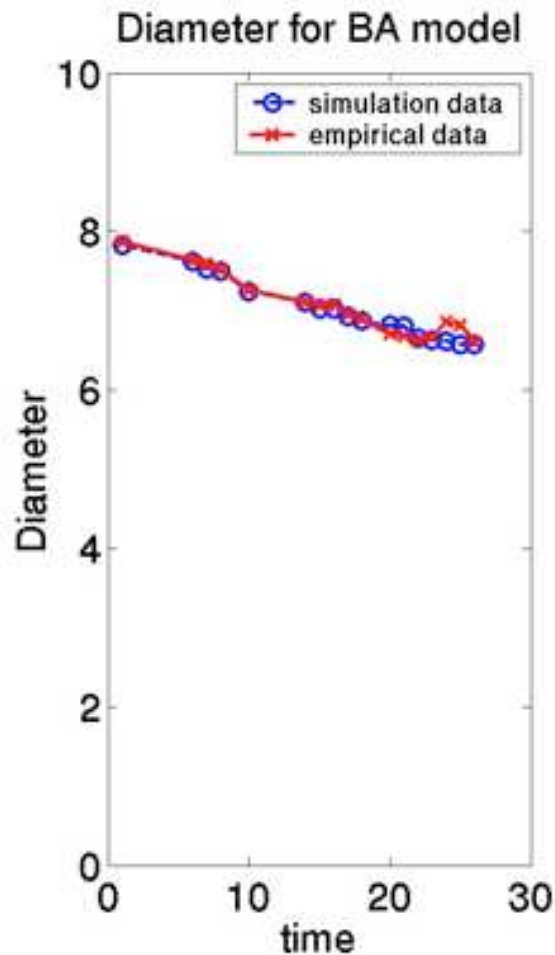
- Degree distribution is normal distribution while it is power law in empirical data

ER Model – Cluster Size Distribution



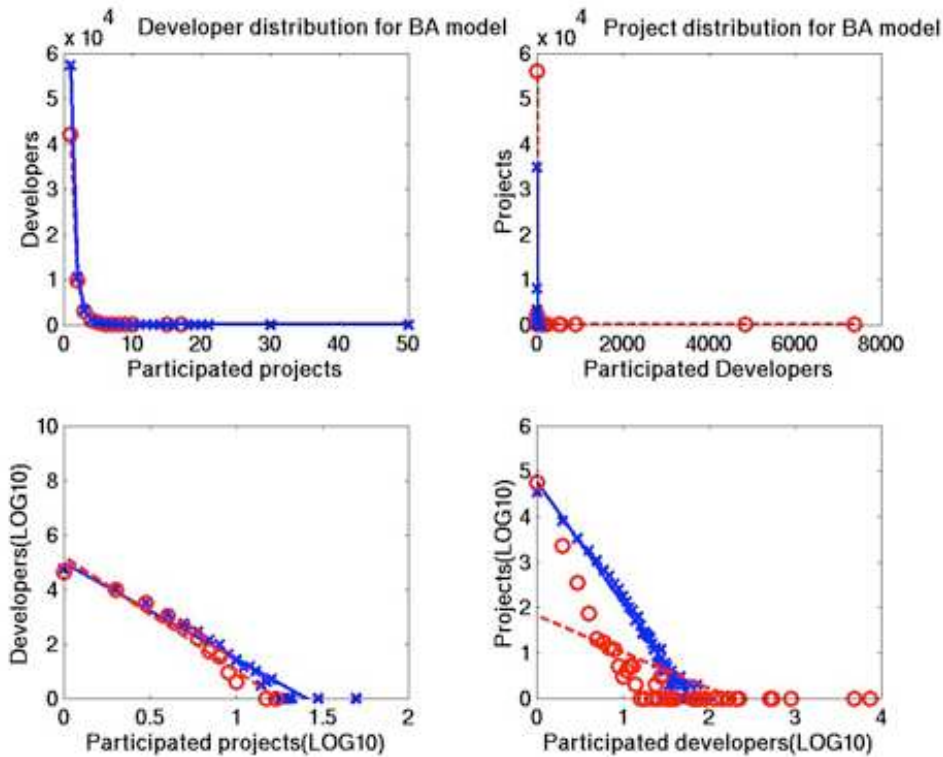
- power law distribution with R^2 as 0.6667 (0.9653 without the major cluster) while R^2 in empirical data is 0.7426 (0.9799 without the major cluster)
- The actual distribution is different from empirical data

BA Model – Diameter and Clustering Coefficient



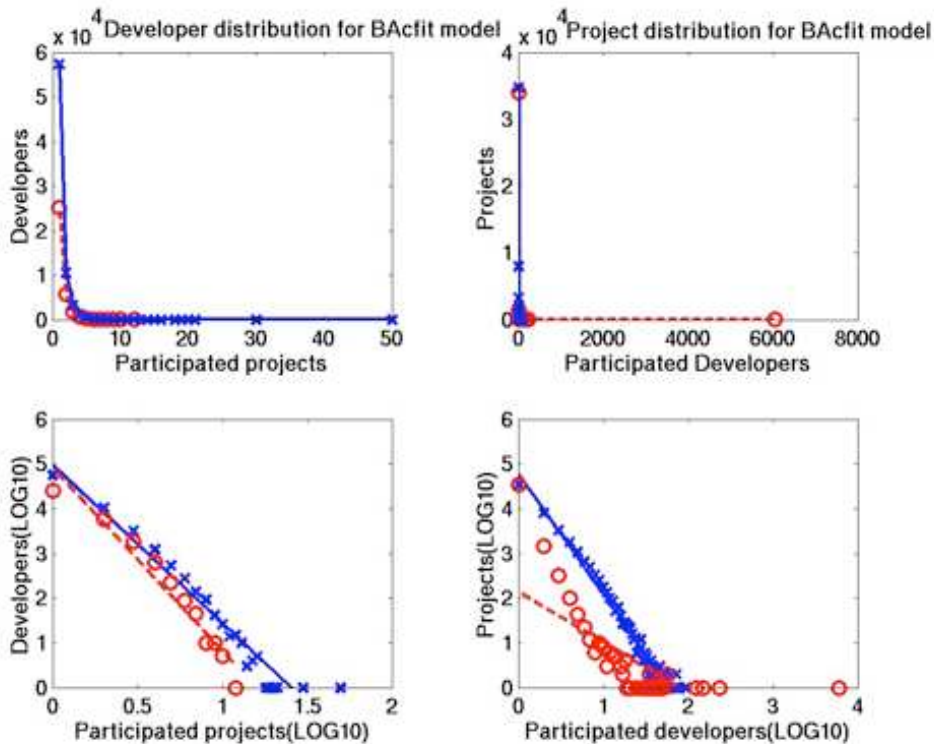
- Small diameter and high clustering coefficient like empirical data
- Diameter and clustering coefficient are both decreasing like empirical data

BA Model – Degree Distribution



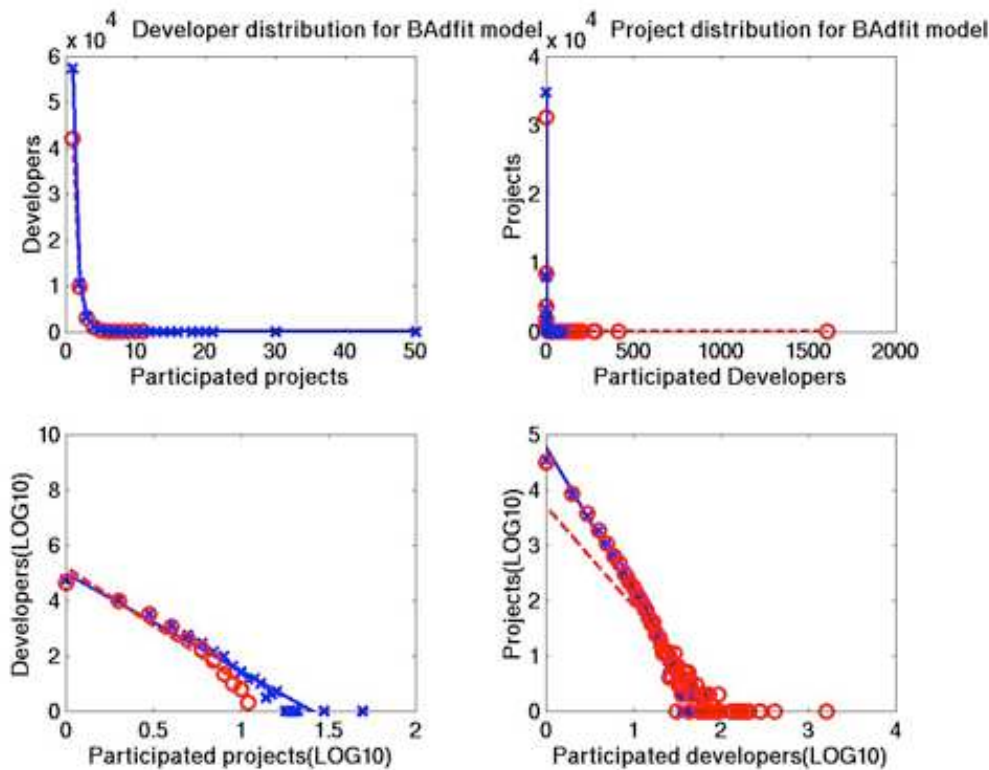
- Power laws in degree distributions, similar to empirical data (o for simulated data and x for empirical data).
- For developer distribution: simulated data has R^2 as 0.9798 and empirical data has R^2 as 0.9714.
- For project distribution: simulated data has R^2 as 0.6650 and empirical data has R^2 as 0.9838.

BA Model with Constant Fitness



- Power laws in degree distributions, similar to empirical data (o for simulated data and x for empirical data).
- For developer distribution: simulated data has R^2 as 0.9742 and empirical data has R^2 as 0.9714.
- For project distribution: simulated data has R^2 as 0.7253 and empirical data has R^2 as 0.9838.

BA Model with Dynamic Fitness



- Power laws in degree distribution, similar to empirical data (o for simulated data and x for empirical data).
- For developer distribution: simulated data has R^2 as 0.9695 and empirical data has R^2 as 0.9714.
- For project distribution: simulated data has R^2 as 0.8051 and empirical data has R^2 as 0.9838.

Idea of Dynamic Fitness



- Intuition: Projects grow and decline with time.
- Statistics: project has life cycle behavior which can not be replicated by BA model with constant fitness, but can be replicated by BA model with dynamic fitness

Summary



Model	Parameter	expected pattern	observed pattern
ER	Developer distribution	Power law	Normal
	Project distribution	Power law	Normal
	Cluster distribution	Power law	Power law
	Average degree	Increasing	Decreasing
	Clustering coefficient	Decreasing (large value)	Decreasing (small value)
	Diameter	Decreasing	Increasing
BA	Developer distribution	Power law	Power law
	Project distribution	Power law	Power law (heavy tail)
	Cluster distribution	Power law	Power law
	Average degree	Increasing	Increasing
	Clustering coefficient	Decreasing (large value)	Decreasing (large value)
	Diameter	Decreasing	Decreasing
	“Young upcomer”	Existing	Not existing
BA with constant fitness	Developer distribution	Power law	Power law
	Project distribution	Power law	Power law (heavy tail)
	Cluster distribution	Power law	Power law
	Average degree	Increasing	Increasing
	Clustering coefficient	Decreasing (large value)	Decreasing (large value)
	Diameter	Decreasing	Decreasing
	“Young upcomer”	Existing	Existing
BA with dynamic fitness	Developer distribution	Power law	Power law
	Project distribution	Power law	Power law (small tail)
	Cluster distribution	Power law	Power law
	Average degree	Increasing	Increasing
	Clustering coefficient	Decreasing (large value)	Decreasing (large value)
	Diameter	Decreasing	Decreasing
	“Young upcomer”	Existing	Existing

Cycles of Modeling & Simulation: Scientific Method



Modeling
(Hypothesis)

Social Network Models

ER => BA => BA+Fitness => BA+Dynamic Fitness

Observation

Analysis of
SourceForge
Data

Agent -Based
Simulation
(Experiment)

Grow Artificial
SourceForge

Degree Distribution
Average Degree
Diameter
Clustering Coefficient
Cluster Size Distribution

Conclusion



- Study of SourceForge collaboration network can help us understanding the OSS community
- Simulation is used to investigate of SourceForge collaboration network.



Thank you