

# Towards Understanding: A Study of the SourceForge.net Community using Modeling and Simulation

Yongqin Gao

Greg Madey

Computer Science & Engineering

University of Notre Dame

ADS '07 - SpringSim '07 - SCS

Norfolk, VA - March 27, 2007

# Outline

- Introduction & Background
- Research data description
- Our scientific methodology
- Experimental results
  - Hypothesis/Model I
  - Hypothesis/Model II
  - Hypothesis/Model III
- Summary and discussion

## Continuation of “Future Work” from:

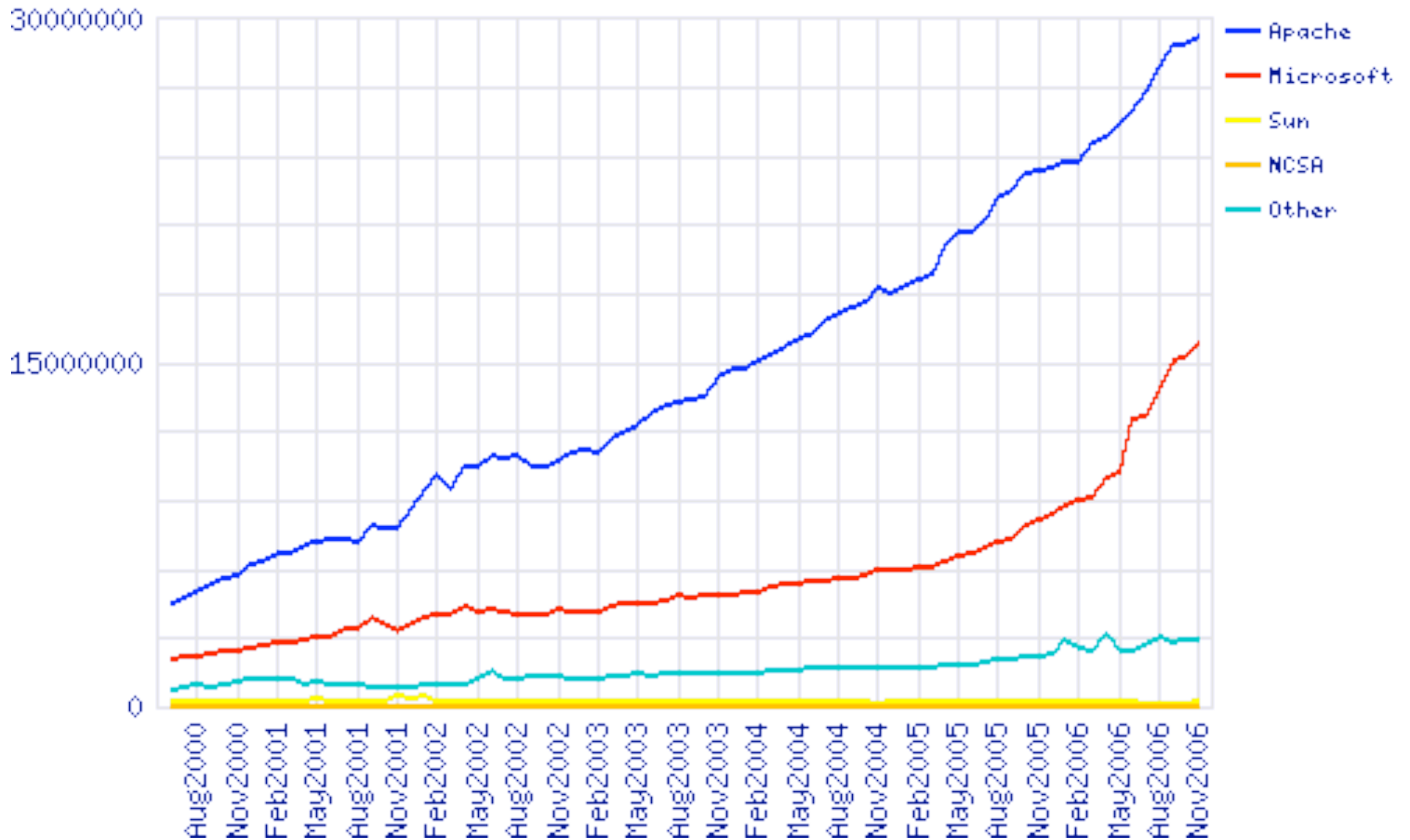
- Yongqin Gao, Greg Madey, Vince Freeh, "Modeling and Simulation of the Open Source Software Community", *Agent-Directed Simulation Conference*, San Diego, CA, April 2005
  - Chapter from Masters Thesis (Gao, 2003)
- This presentation/paper describes a continuation of the research
  - Chapter in PhD Dissertation (Gao, 2007)
  - <http://www.nd.edu/~oss/Papers/papers.html>

# Background (OSS)

- What is OSS?
  - Free to use, modify and distribute
  - Source code available and modifiable
- Potential advantages over commercial software
  - High quality
  - Fast development
  - Low cost
- Why study OSS?
  - Software engineering — new development and coordination methods
  - Open content — model for other forms of open, shared collaboration
  - Complexity — successful example of self-organization/emergence
  - Economic motivations, virtual teams, organizational behavior, patent and intellectual property, etc.

Evidence of adoption and popularity is Apache-->

# Number of Active Apache Hosts



Source: <http://news.netcraft.com/>

SOURCEFORGE<sup>™</sup>  
net

  
OpenOffice.org  
The Open Source Office Suite

<http://www.freebsd.org>  
**FreeBSD**  
FreeBSD: The Power To Serve

mozilla.org

# Open Source Software (OSS) Linux



Savannah



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



K DESKTOP ENVIRONMENT



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

MySQL



RePast



The Apache Software Foundation  
<http://www.apache.org/>

python


# Research Data



**SourceForge.net**  
Create, Participate, Evaluate

Registered Projects: 144,548 Registered Users: 1,545,019

SourceForge is Hiring 

Project of the Month   
Zenoss Core

- SourceForge.net community
  - The largest OSS development community
  - Over 144,000 registered projects
  - Over 1,545,000 registered users
- SourceForge.net Research Archive
  - <http://zerlot.cse.nd.edu/>
  - <http://www.nd.edu/~oss/Data/>
  - 500 GB of data
  - Open to scholarly researchers

<http://sourceforge.net/>

March 27, 2007

SOURCEFORGE®  
net

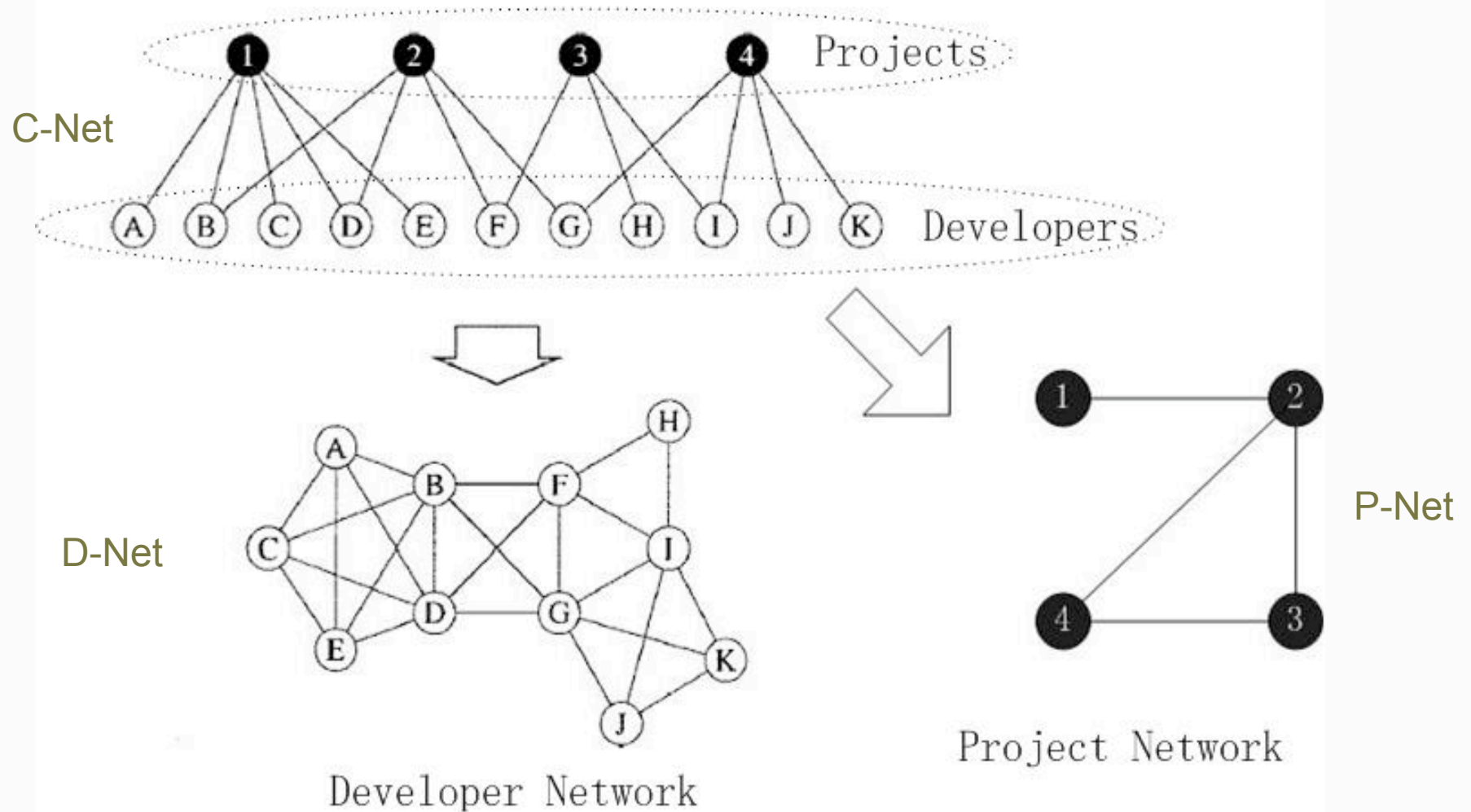
Create Participate Evaluate

# Collaboration Networks



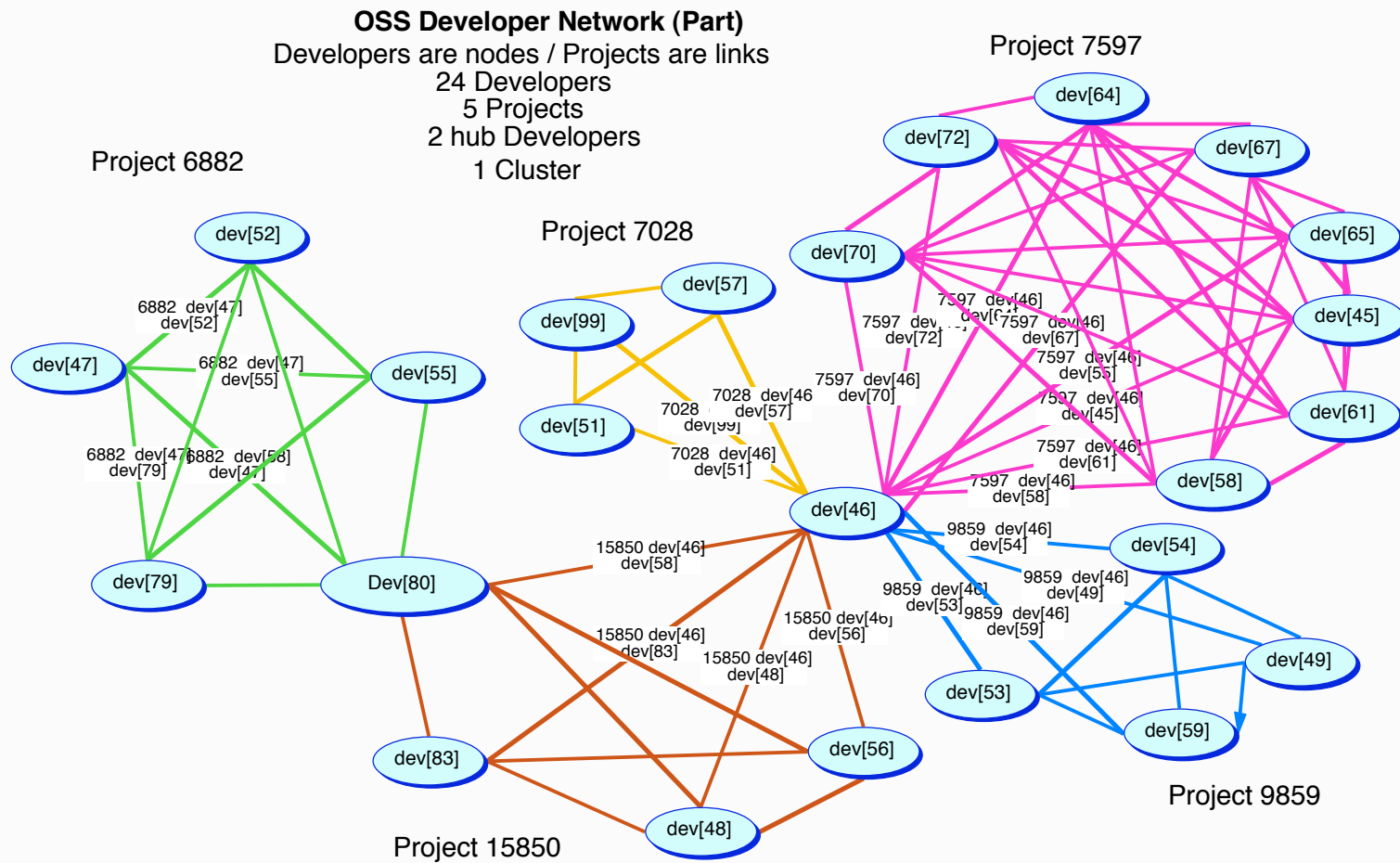
- What is a collaboration network?
  - A social network representing the collaborating relationships
  - Movie actor network
    - Kevin Bacon number
  - Research paper authorship network
    - Erdős number in mathematics
  - Open source software developers/projects
- Differences in the SourceForge collaboration network
  - Link detachment
  - Virtual collaboration
  - Open source software
- Bipartite/unipartite properties of collaboration networks

# Collaboration Networks Bipartite and Unipartite

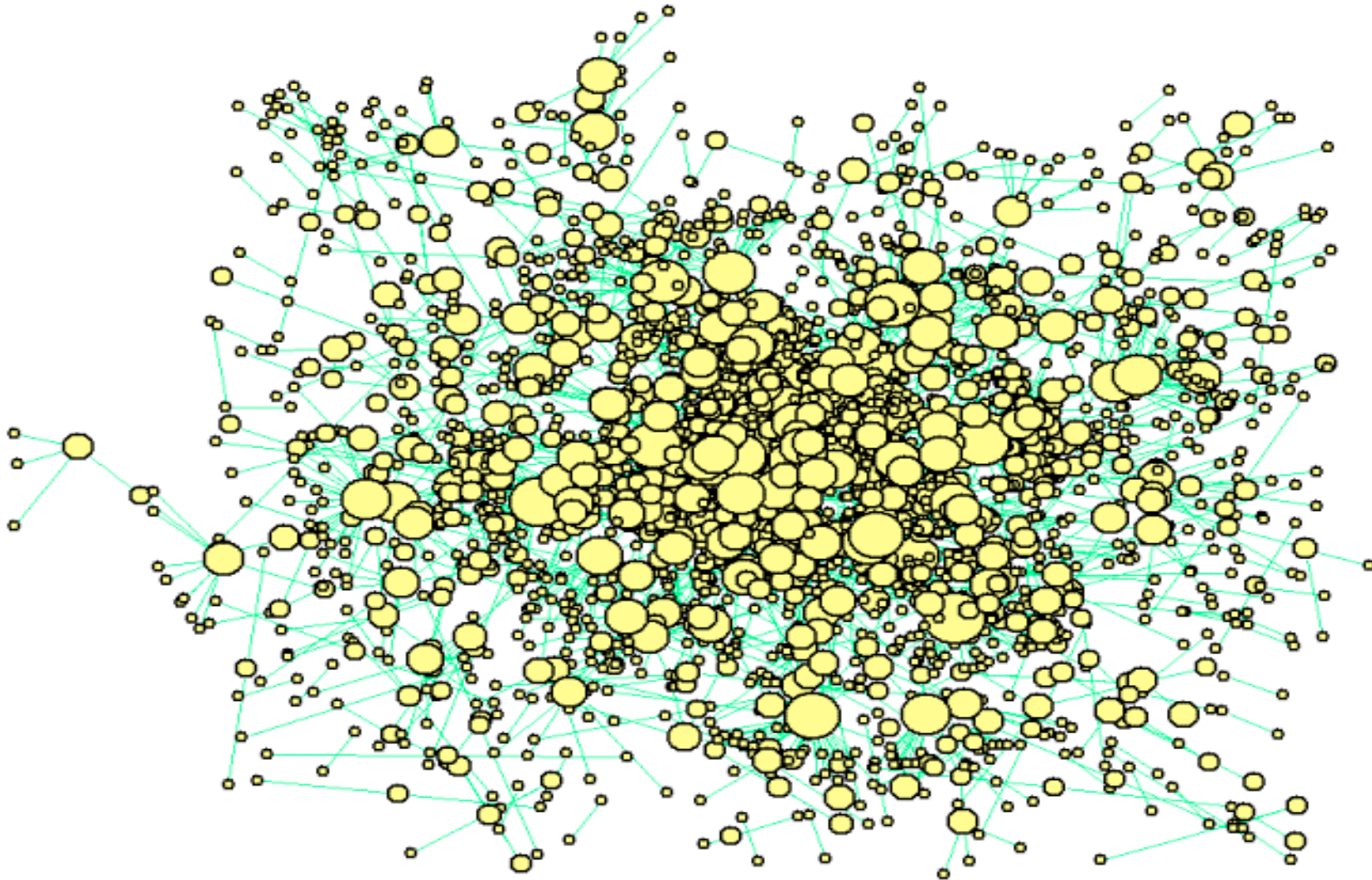


Adapted from Newman, Strogatz and Watts, 2001

# SourceForge Developer Collaboration Network (a cluster)



# Another Cluster



# Research Data Description

- Our Data Set
  - 27 monthly dumps between Jan 2003 and Mar 2007
  - Every dump has about 100 tables
  - Largest table has up to 30 million records
- Experimental Environment
  - Dual Xeon 3.06GHz, 4G memory, 2T storage
  - Linux 2.4.21-40.ELsmp with PostgreSQL 8.1
  - Swarm and R

# 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
  - Fashion & Style

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

Past 30 Days

GO TO **MEMBER CENTER**

LOG OUT

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 TOOLS  
SPONSORED BY

STARBUCKS.COM

### TIMES NEWS TRACKER

Topics

[Fermi, Enrico](#)

[DNA \(Deoxyribonucleic Acid\)](#)

[Science and Technology](#)

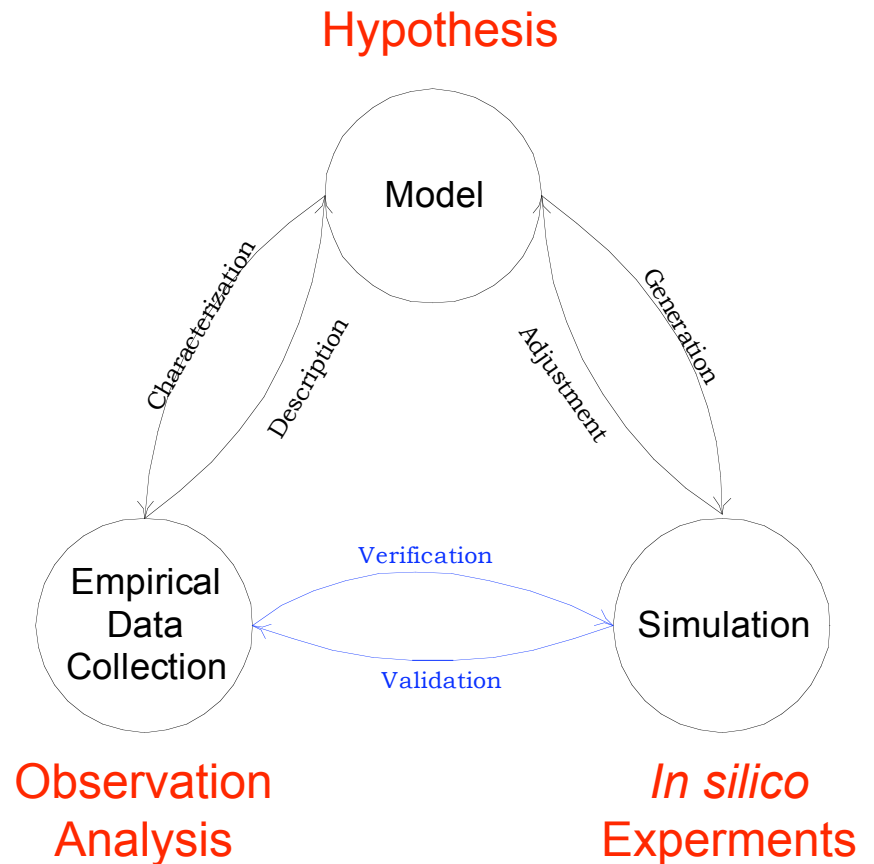
[Create Your Own](#) | [Manage Alerts](#)  
[Take a Tour](#)

[Sign Up for Newsletters](#)

Alerts

# Scientific Methodology

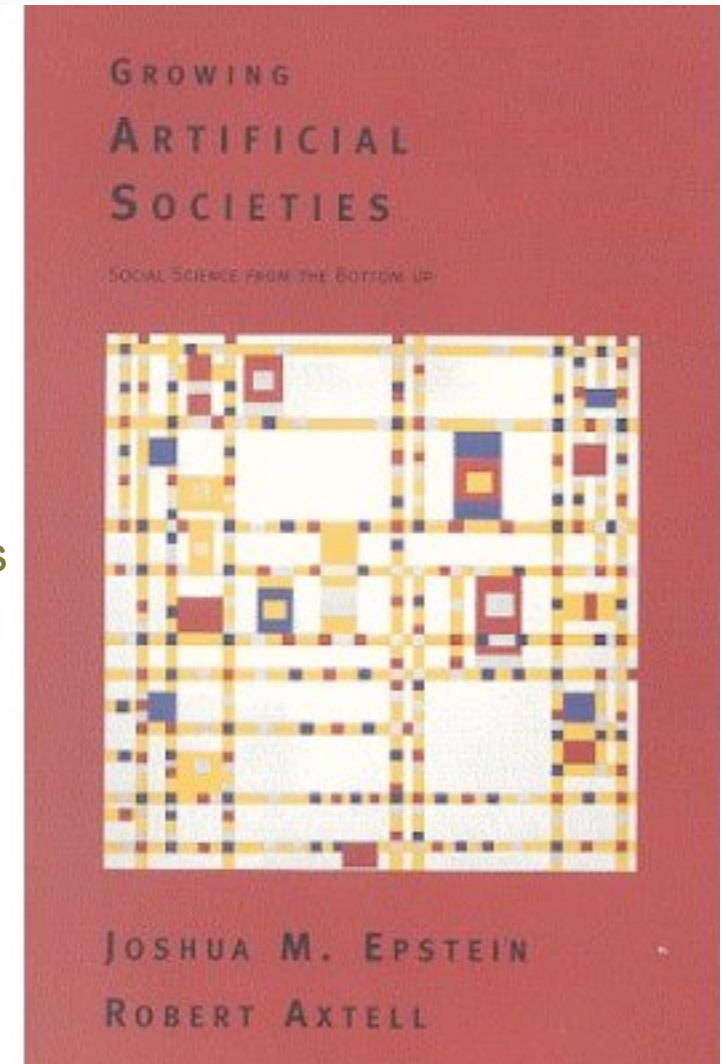
- Iterative simulation method
  - Empirical dataset
  - Model
  - Simulation
- Verification and validation
  - More measures
  - More methods
- Analogous to the development of engineering simulations



# Model of SourceForge.net



- ABM based on unipartite graph
- Grow Artificial SourceForge.net's to evaluate hypotheses about evolution of real-world SourceForge.net
- Model description
  - Agents: developers with randomized characteristics
  - Behaviors: create, join, abandon and idle
  - Projects: have attractiveness / characteristics
  - Developers: have preferences / characteristics
- Previous: Four models / hypotheses
  - ER, BA, BA with constant fitness and BA with dynamic fitness
- New: Three models / hypotheses
- Comparison of observed and simulated social networks
  - Social network properties
  - Measures of graph (network) characteristics



# Summary: Previous Study

(Gao, Freeh & Madey, ADS 2005)



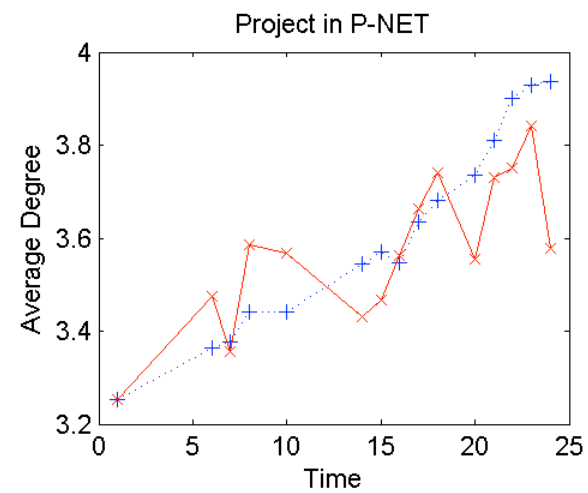
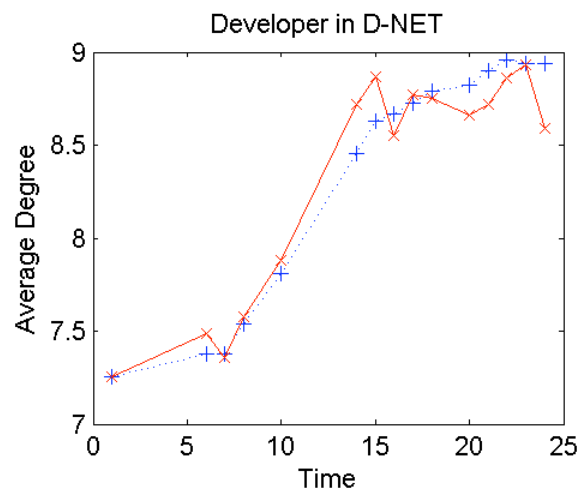
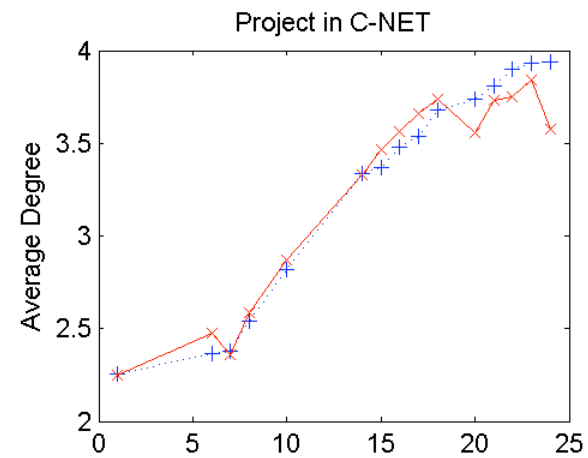
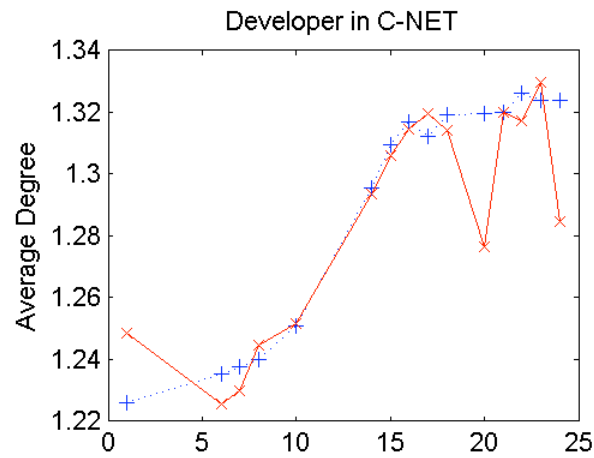
| Model                    | Parameter              | expected pattern         | observed pattern         |
|--------------------------|------------------------|--------------------------|--------------------------|
| ER                       | Developer distribution | Power law                | Normal                   |
|                          | 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               |
| 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                 |

# Model I

- Description
  - Realistic stochastic procedures.
    - New developer every time step based on Poisson distribution
    - Initial fitness based on log-normal distribution
  - Updated procedure for the weighted project pool (for preferential selection of projects).

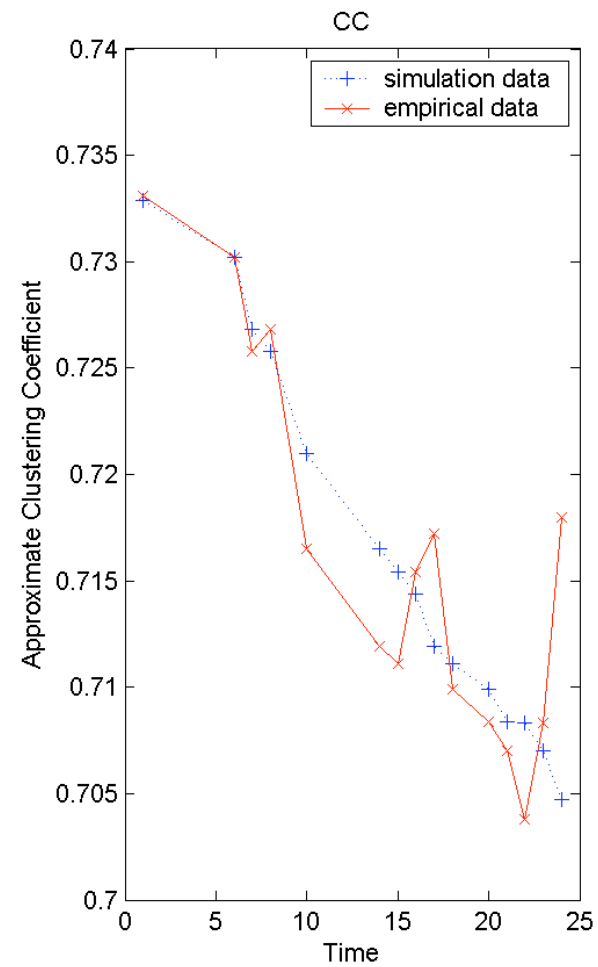
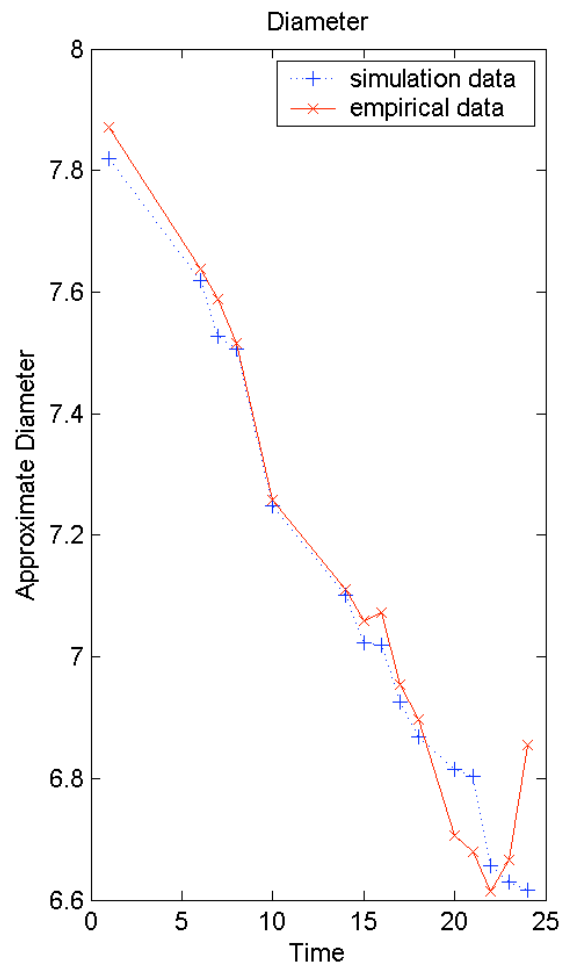
# Results: Model I

- Average degrees



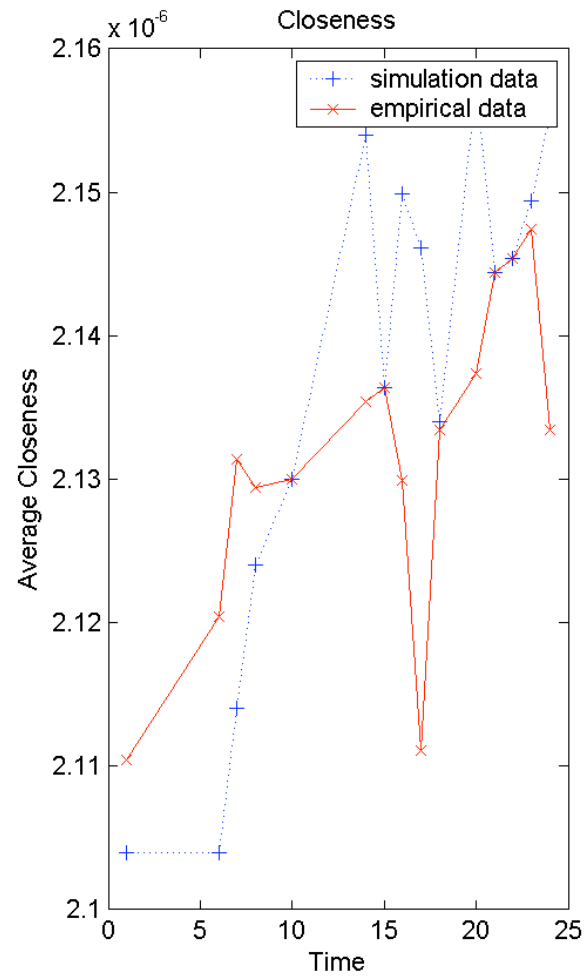
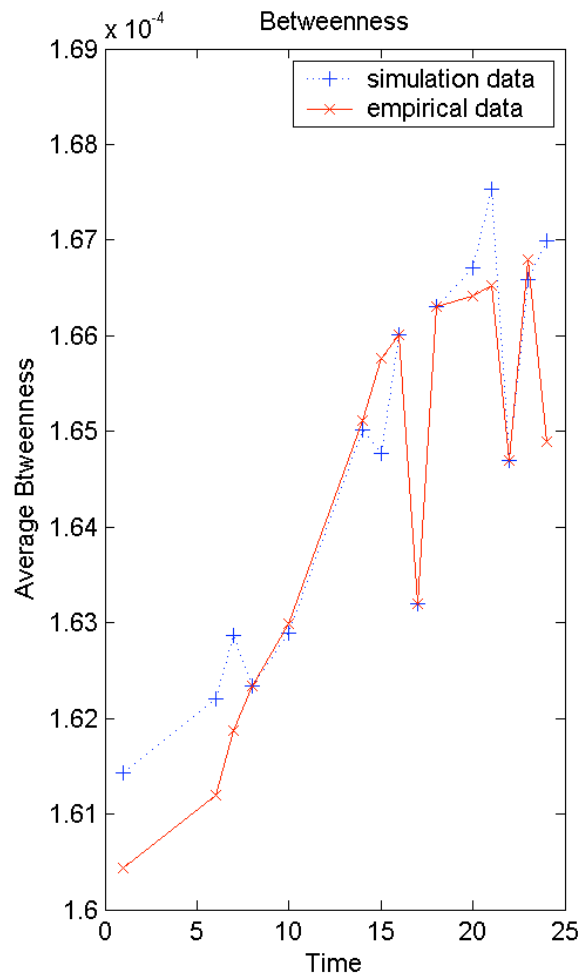
# Results: Model I

- Diameter and CC



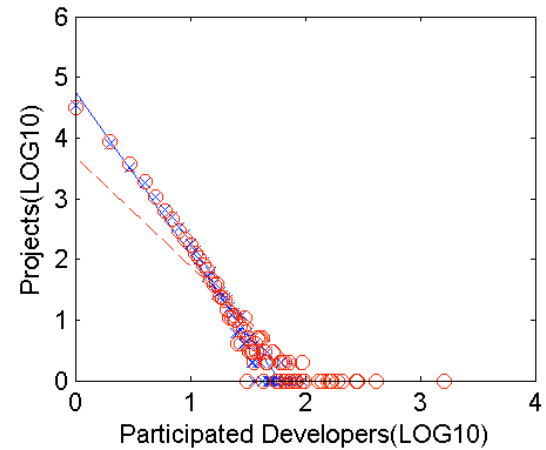
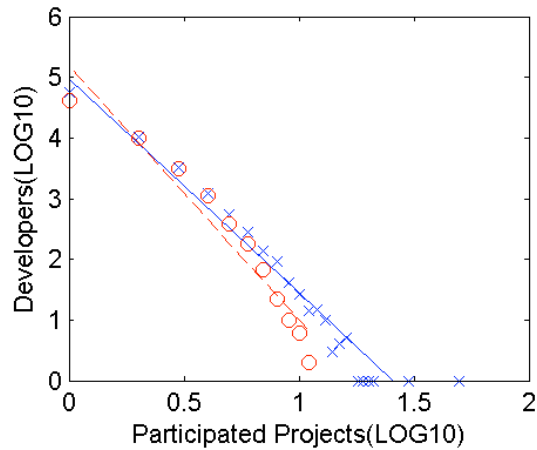
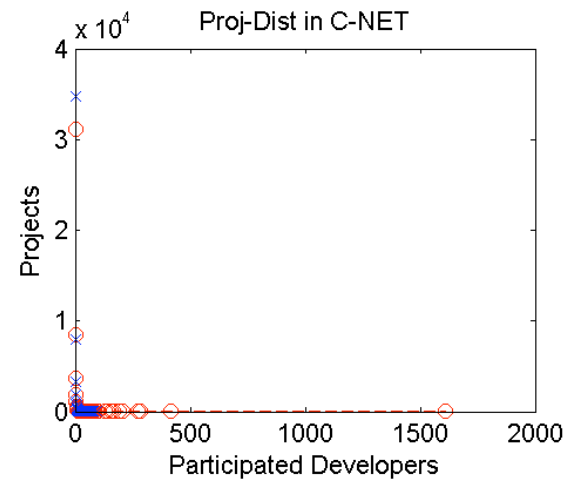
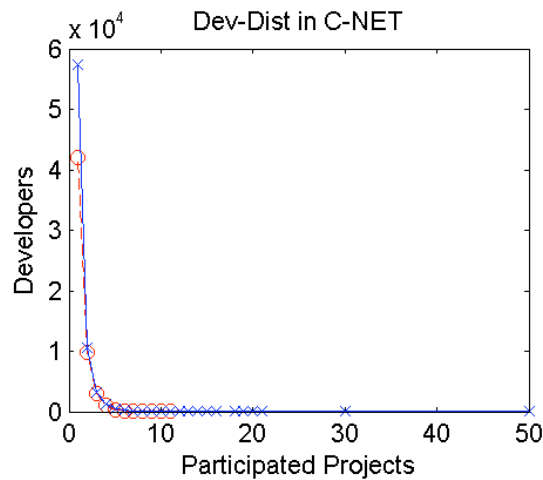
# Results: Model I

- Betweenness and Closeness



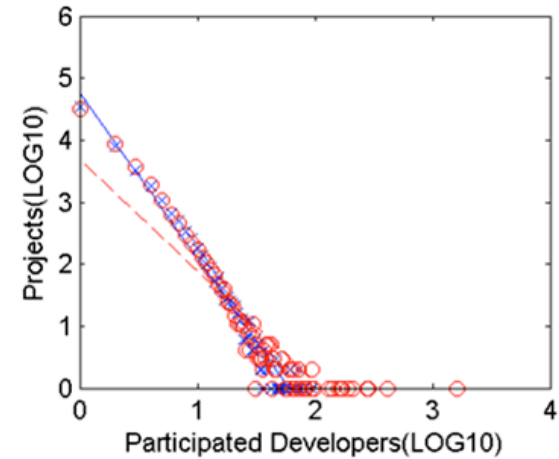
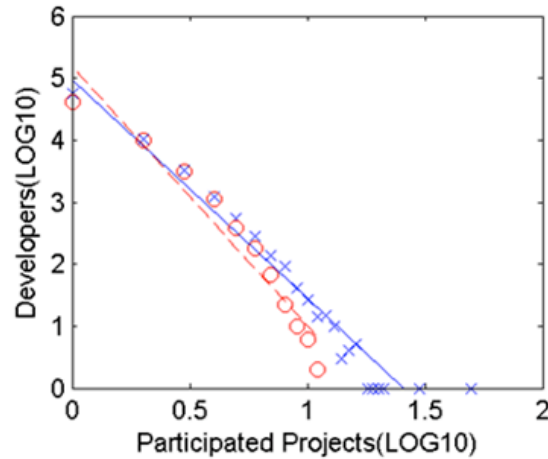
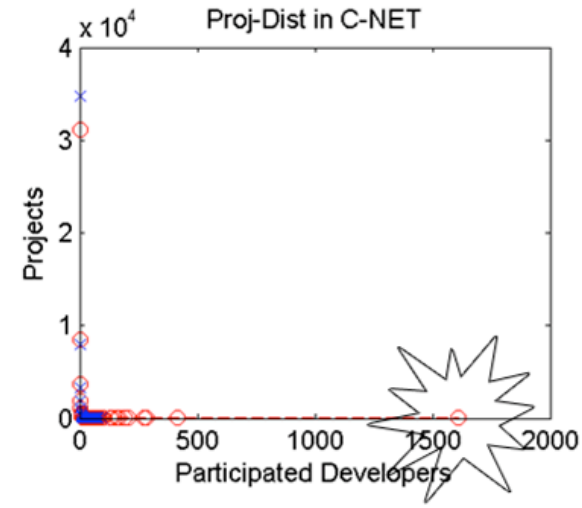
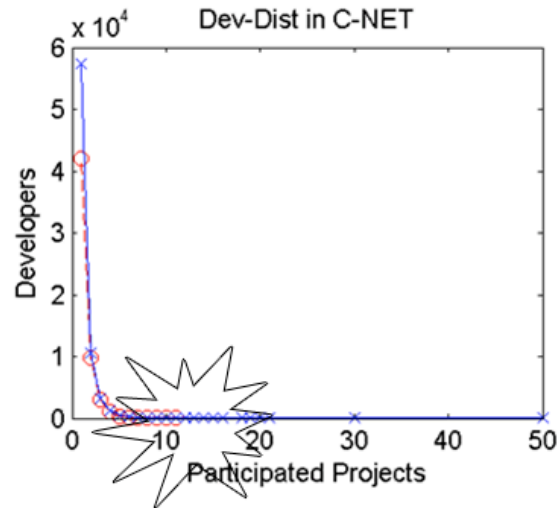
# Results: Model I

- Degree Distributions



# Results: Model I

- Problems

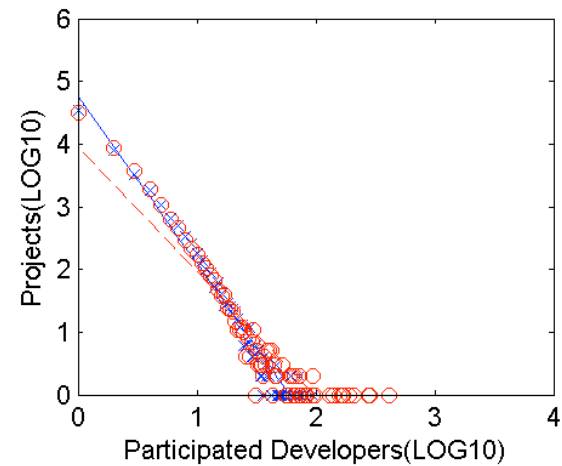
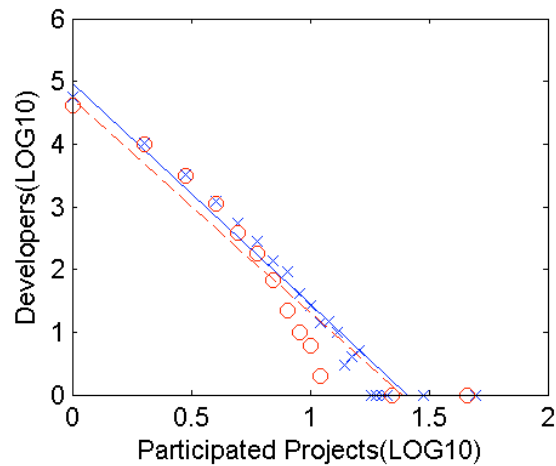
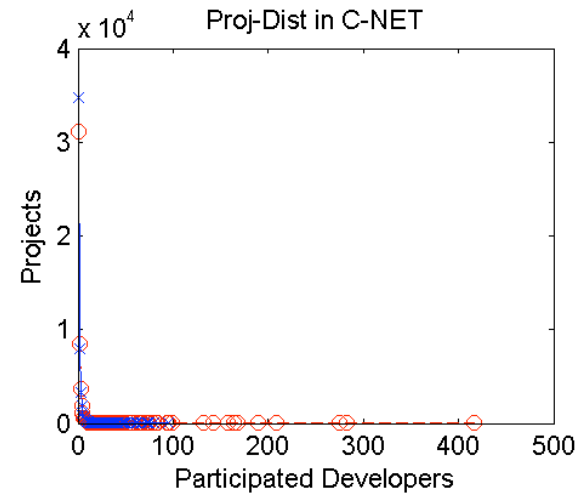
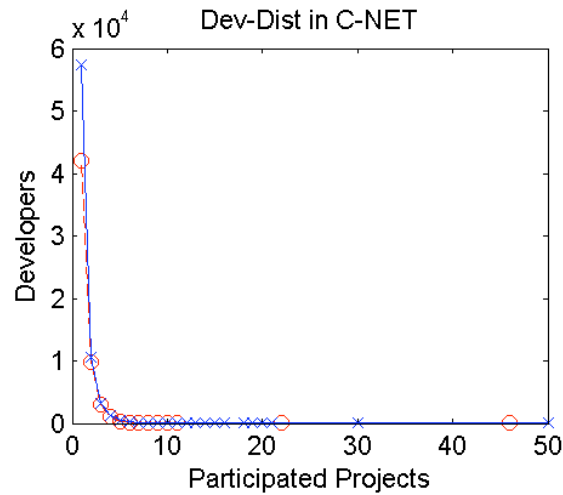


# Model II

- Description
  - New addition: user energy.
  - User energy
    - The “fitness” parameter for the user
    - Every time a new user is created, a energy level is randomly generated for the user
    - Energy level will be used to decide whether a user will take a action or not during every time step.

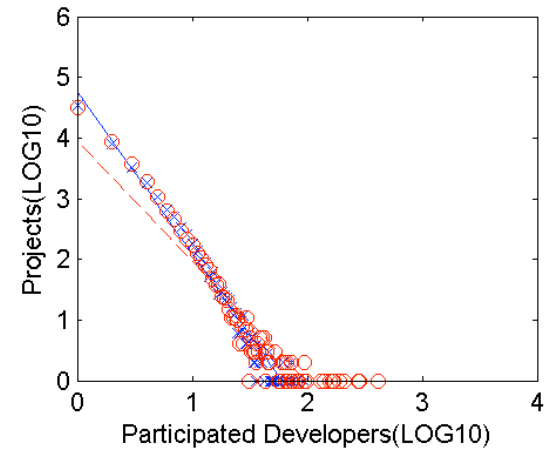
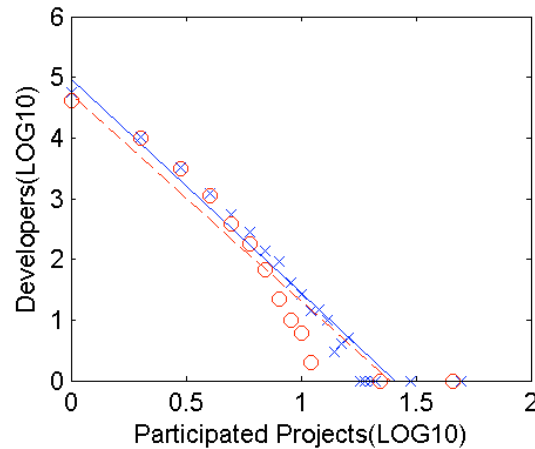
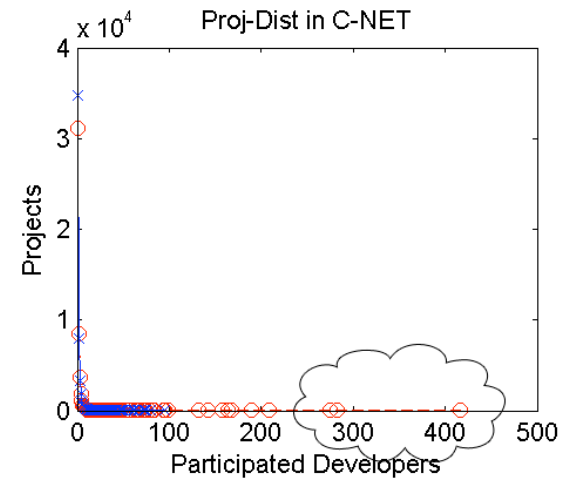
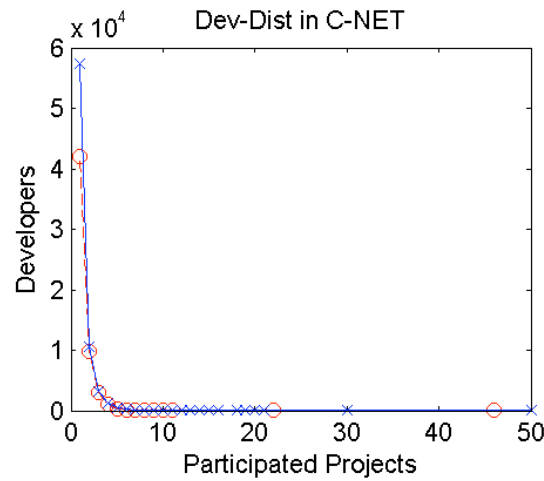
# Results: Model II

- Degree distributions



# Results: Model II

- Better, but still has problems

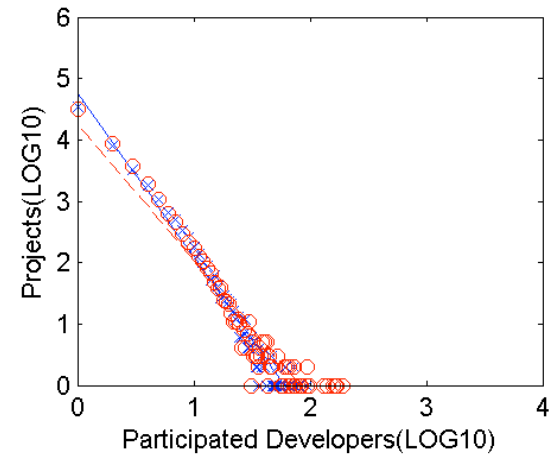
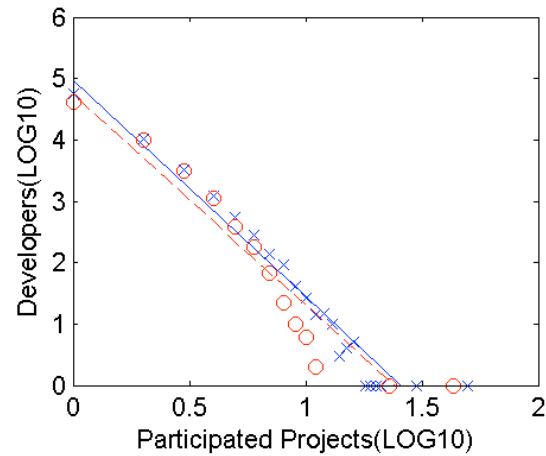
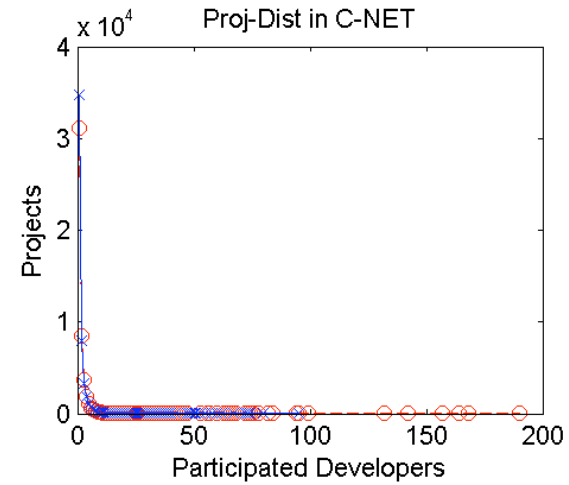
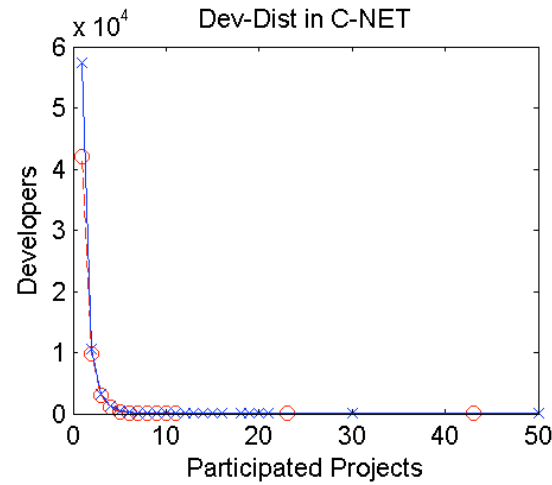


# Model III

- Description
  - New addition: dynamic user energy.
  - Dynamic user energy
    - Decaying with respect to time
    - Self-adjustable according to the roles the user is taking in various projects.

# Results: Model III

- Degree distributions



# Summary

| Models  | Measures               | Patterns in Data              | Simulated Patterns                 |
|---|------------------------|-------------------------------|------------------------------------|
| <b>Model I</b><br><b>(more realistic distributions)</b> | Developer Distribution | <b>Power Law (large tail)</b> | <b>Power Law (small tail)</b>      |
|   | Project Distribution   | <b>Power Law (small tail)</b> | <b>Power Law (large tail)</b>      |
|   | Average Degrees        | Increasing                    | Increasing                         |
|   | Clustering Coefficient | Decreasing                    | Decreasing                         |
|   | Diameter               | Decreasing                    | Decreasing                         |
|   | Average Betweenness    | Decreasing                    | Decreasing                         |
|   | Average Closeness      | Decreasing                    | Decreasing                         |
| <b>Model II</b><br><b>(constant user energy)</b>        | Developer Distribution | Power Law (large tail)        | Power Law (large tail)             |
|   | Project Distribution   | <b>Power Law (small tail)</b> | <b>Power Law (reasonable tail)</b> |
|   | Average Degrees        | Increasing                    | Increasing                         |
|   | Clustering Coefficient | Decreasing                    | Decreasing                         |
|   | Diameter               | Decreasing                    | Decreasing                         |
|   | Average Betweenness    | Decreasing                    | Decreasing                         |
|   | Average Closeness      | Decreasing                    | Decreasing                         |
| <b>Model III</b><br><b>(dynamic user energy)</b>        | Developer Distribution | Power Law (large tail)        | Power Law (large tail)             |
|   | Project Distribution   | Power Law (small tail)        | Power Law (small tail)             |
|   | Average Degrees        | Increasing                    | Increasing                         |
|   | Clustering Coefficient | Decreasing                    | Decreasing                         |
|   | Diameter               | Decreasing                    | Decreasing                         |
|   | Average Betweenness    | Decreasing                    | Decreasing                         |
|   | Average Closeness      | Decreasing                    | Decreasing                         |

# Discussion

- Results/Discussion
  - Expanding the network models for modeling evolving complex networks (more hypotheses and computer experiments)
  - Provided a validated model to simulate the collaboration network at SourceForge.net
  - Demonstration of the use of “computer experiments” for scientific research using agent-based modeling --> analogous to the development of engineering simulations
  - Research approach that can be used to study other OSS communities or similar collaboration networks
  - Demonstrated the use of various network metrics for V&V of agent simulations
  - Resources/references:
    - <http://www.nd.edu/~oss/Papers/papers.html>
    - <http://zerlot.cse.nd.edu/mywiki/>



**Thank you!**

# Related Work

- Related Research:
  - P.J. Kiviat, “Simulation, technology, and the decision process”, *ACM Transactions on Modeling and Computer Simulation*, 1991.
  - R. Albert and A.L. Barabási, “Emergence of scaling in random networks”, *Science*, 1999.
  - J. Epstein R. Axtell, R. Axelrod and M. Cohen, “Aligning simulation models: A case study and results”, *Computational and Mathematical Organization Theory*, 1996.

# Continuation of the Computer Experimental Cycle

- Previous iterated models (ADS 05):
  - Adapted ER Model
  - BA Model
  - BA Model with fitness
  - BA Model with dynamic fitness
- Iterated models in this study (ADS 07)
  - Improved Model Four (Model I)
  - Constant user energy (Model II)
  - Dynamic user energy (Model III)