

# The Computer Experiment in Computational Social Science

Greg Madey

Yongqin Gao

Computer Science & Engineering  
University of Notre Dame

*<http://www.nd.edu/~gmadey>*

**Eighth Annual Swarm Users/Researchers Conference**

University of Michigan  
Ann Arbor, Michigan USA

May 9-11, 2004

*This research was partially supported by the US National Science Foundation, CISE/IIS-Digital Society & Technology, under Grant No. 0222829*

# Outline

- Background
- The epistemological questions
- Example research question
- Simulation
- Computer experiments
- Discussion

# Background

- Two NFS projects using agent-based simulation
  - 1) Molecules and microbes as agents
  - 2) Free/Open Source Software developers as agents
- Primarily scientific investigations — with IT tool building and simulation support
- How do you justify the use of simulation?
  - From a philosophy of science perspective (not engineering) what do simulation results tell us?

# Why Agent-based Approach for Molecules



# The Epistemological Questions

- How do we come to know social science knowledge?
- What do we (or should we) accept as support for proposition in social science research?
  - Often “real” experiments are not possible
    - Only one real history
    - Ethical issues
- What role can simulation play in answering the above?
- Does simulation have a role beyond “fishing expeditions”?
  - Simulation just discovers phenomenon for “real experiments”?

# Classical Scientific Method

1. Observe the world
  - a) Identify a puzzling phenomenon
2. Generate a falsifiable hypothesis (K. Popper)
3. Design and conduct an experiment with the goal of disproving the hypothesis
  - a) If the experiment “fails”, then the hypothesis is accepted (until replaced)
  - b) If the experiment “succeeds”, then reject hypothesis, but additional insight into the phenomenon may be obtained and steps 2-3 repeated
4. Then add to the body of theory
  - a) A new axiom/law
  - b) A new model
  - c) Then derive new deductions or model conclusions

(Note: Realism vs Instrumentalism)

# 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 [ ]

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

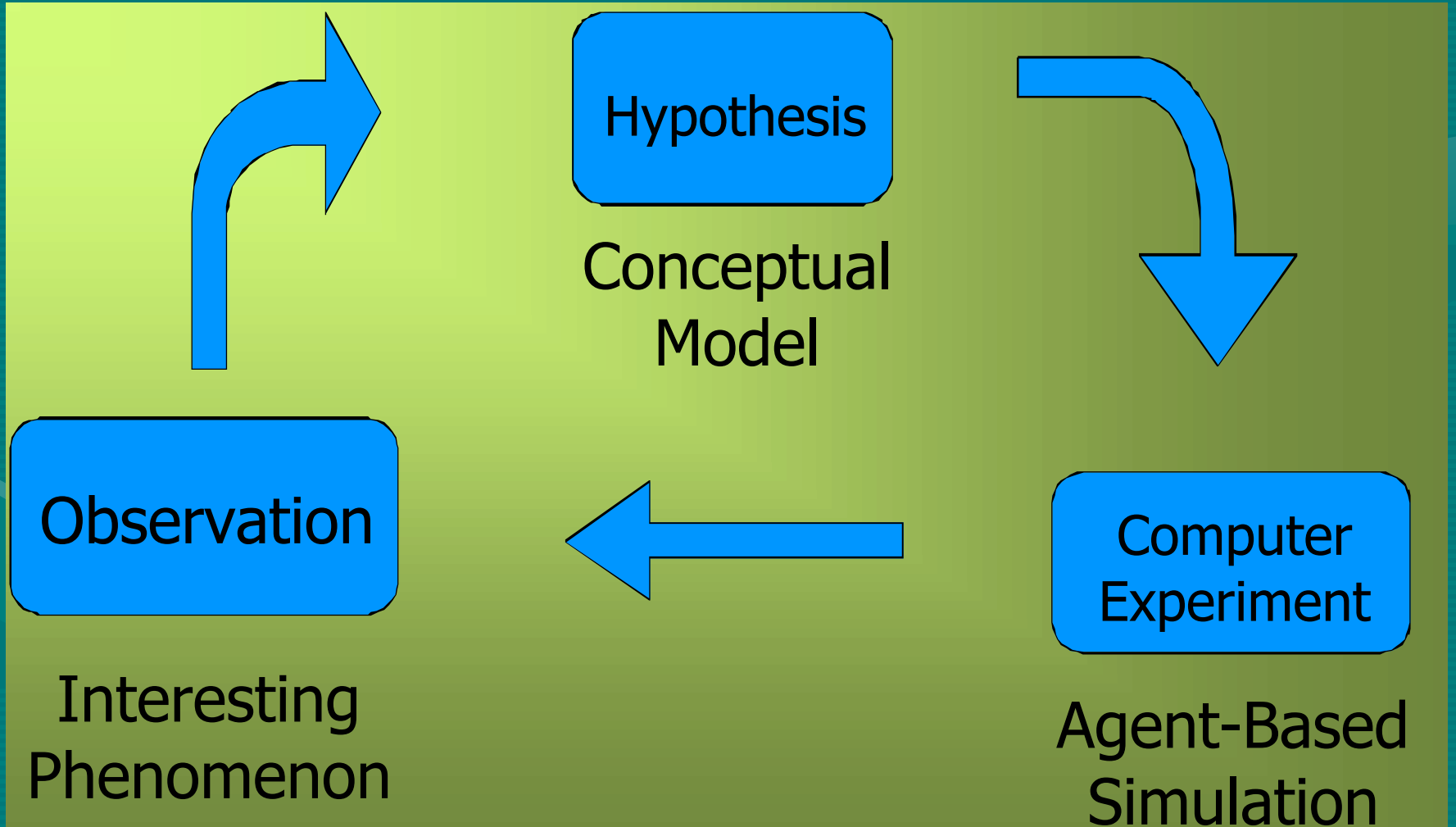
ARTICLE FIELD SPONSORED BY **STARBUCKS.COM**

### TIMES NEWS TRACKER

Topics	Alerts
<a href="#">Fermi, Enrico</a>	<input type="button" value="Create"/>
<a href="#">DNA (Deoxyribonucleic Acid)</a>	<input type="button" value="Create"/>
<a href="#">Science and Technology</a>	<input type="button" value="Create"/>

[Create Your Own](#) | [Manage Alerts](#)  
[Take a Tour](#)  
[Sign Up for Newsletters](#)

# Agent-Based Simulation as a Component of the Scientific Method



SOURCEFORGE™  
net

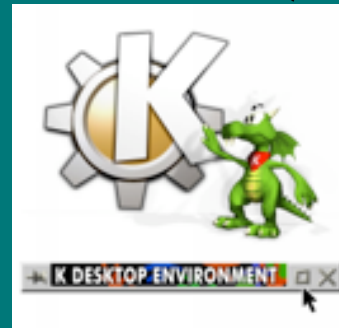
  
**OpenOffice.org 1.0**  
The Open Source Office Suite

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

# GNU Open Source Software (OSS) Linux

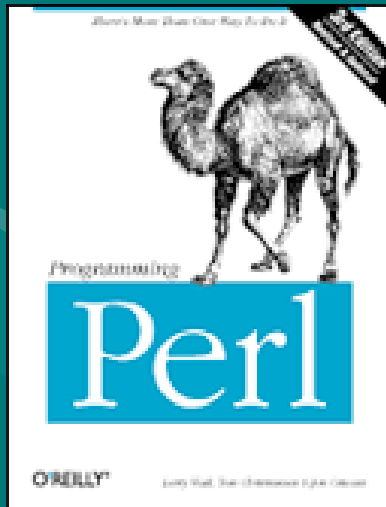


- Free ...
  - to view source
  - to modify
  - to share
  - of cost



**Savannah**

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



**RePast**



mozilla.org

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

**Python**

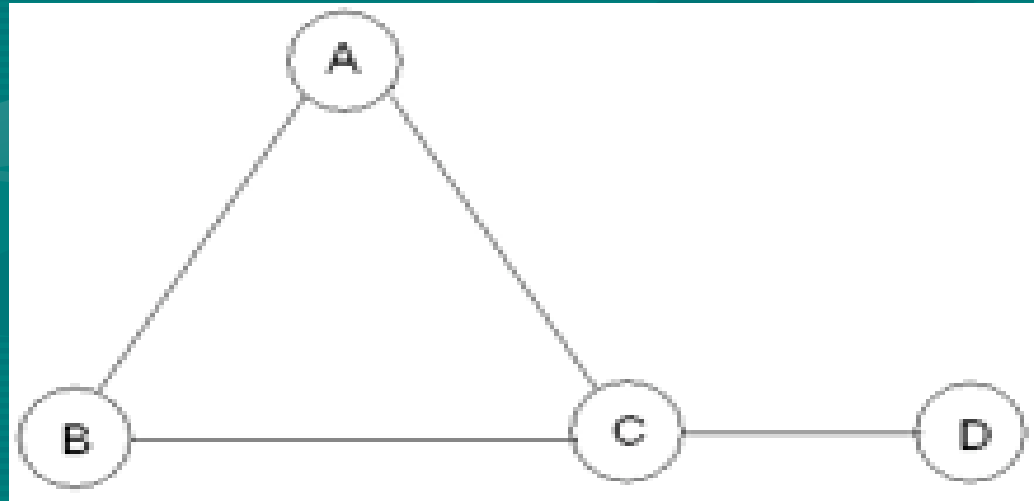


# Example: F/OSS Study

- Online data
  - Screen scraping
  - Database dumps
- Modeling
  - Social network theory
  - Evolutionary assumptions
- Simulation
  - Verification and validation
  - Computer experiments
- Variation of Classical Scientific Method

# Collaborative Social Networks

- Research-paper co-authorship, small world phenomenon, e.g., Erdos number (Barabasi 2001, Newman 2001)
- Movie actors, small world phenomenon, e.g., Kevin Bacon number (Watts 1999, 2003)
- Interlocking corporate directorships
- Terrorist Networks
- Open-source software developers (Madey et al, AMCIS 2002)
- Collaborators are nodes in a graph, and collaborative relationship are the edges of the graph => a framework to model data/phenomenon



# SourceForge



- VA Software
- Part of OSDN
- Started 12/1999
- Collaboration tools
- 70,000 Projects
- 90,000 Developers
- 800,00 Registered Users

# Observations

- Web mining
- Web crawler (scripts)
  - Python
  - Perl
  - AWK
  - Sed
- Monthly
- Since Jan 2001
- ProjectID
- DeveloperID
- Almost 2 million records
- Relational database

PROJ	DEVELOPER
8001	dev378
8001	dev8975
8001	dev9972
8002	dev27650
8005	dev31351
8006	dev12509
8007	dev19395
8007	dev4622
8007	dev35611
8008	dev8975

# F/OSS Developers - Collaboration Social Network

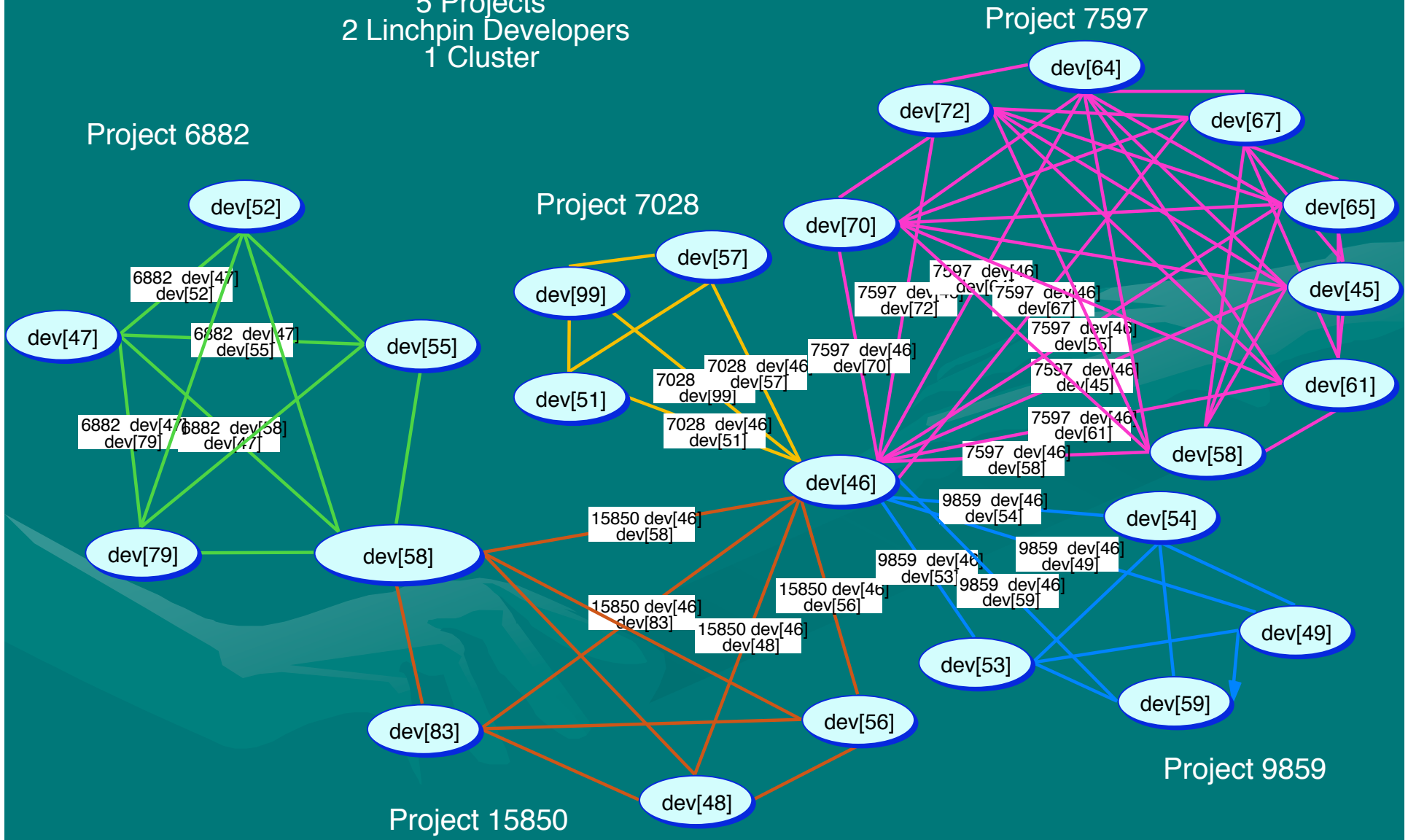
Developers are nodes / Projects are links

24 Developers

5 Projects

2 Linchpin Developers

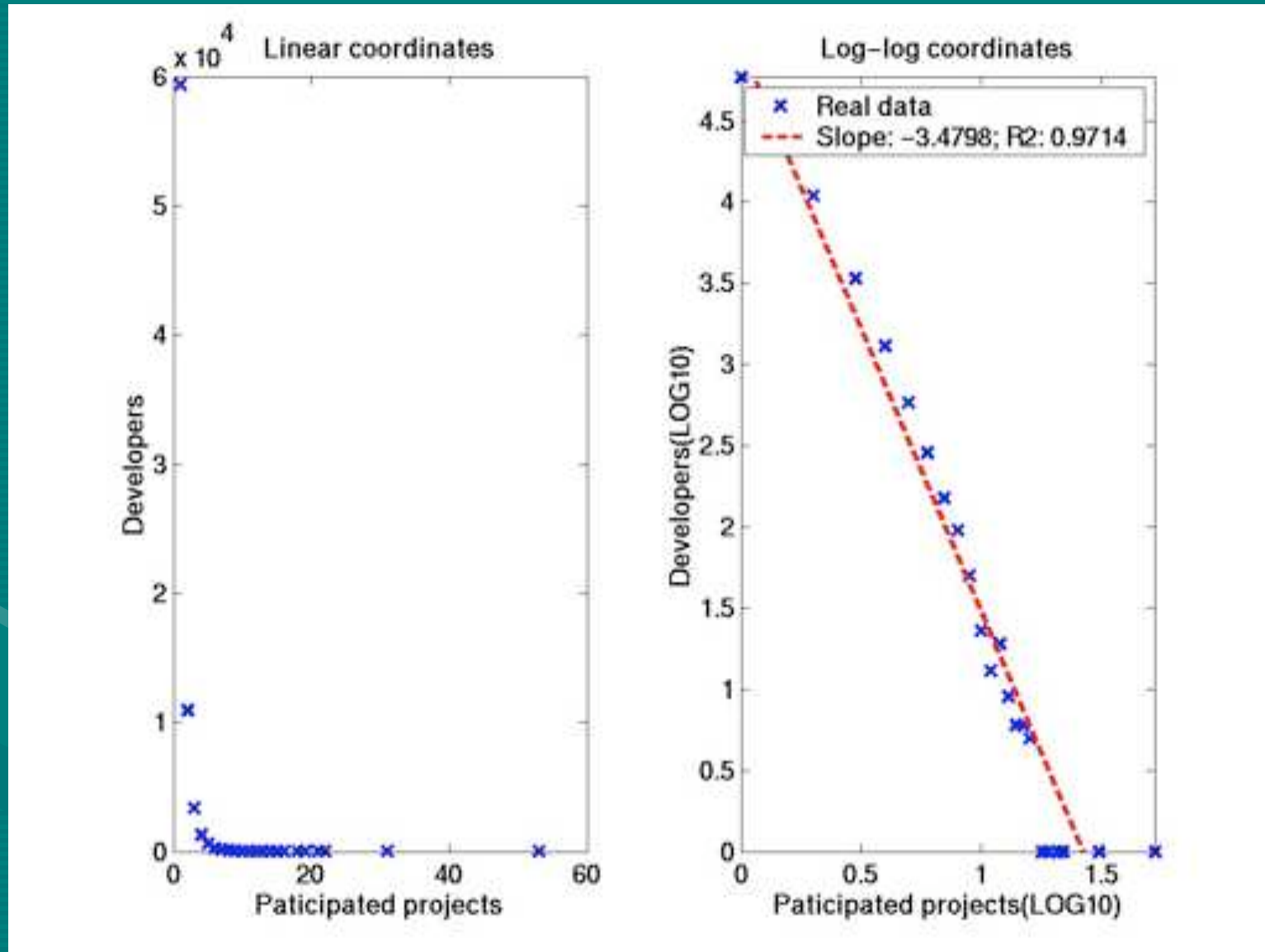
1 Cluster



# Topological Analysis of the Data

- Statistics inspected
  - Diameter
  - Average degree
  - Clustering coefficient
  - Degree distribution
  - Cluster size distribution
  - Relative size of major cluster
  - Fitness and life cycle
- Evolution of these statistics
- Dual networks
  - developer network and project network

# Degree Distribution: Developers



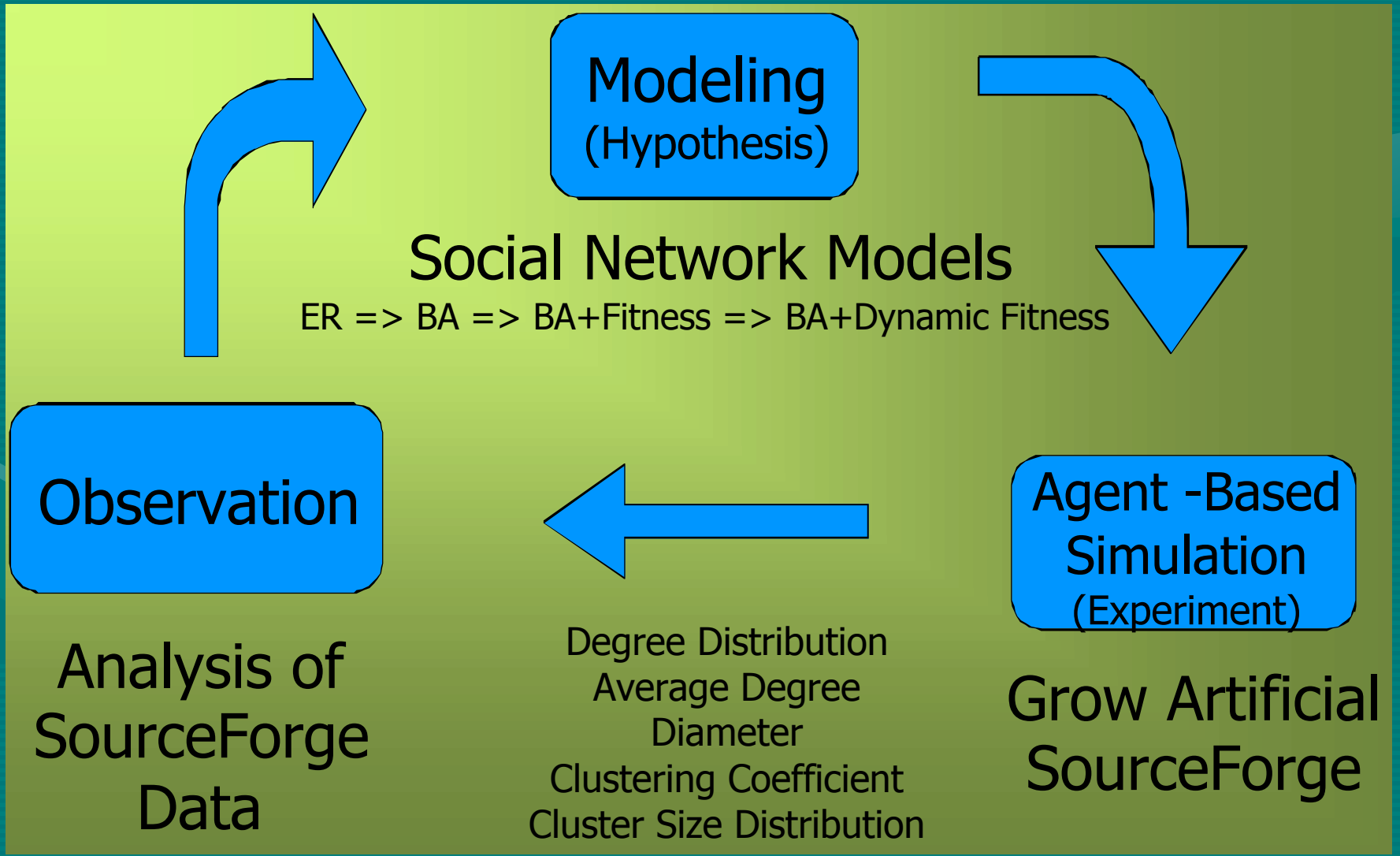
# An Example Research Question

- What processes can explain the evolution of the developer social networks?
  - Randomly growing network (Erdos-Reyni, 1960)?
  - Evolving network with preferential attachment (Barabasi-Albert, 1999)?
  - Evolving network with preferential attachment and fitness (Barabasi-Albert, 2001)?
  - Evolving network with preferential attachment and fitness (Madey et al, 2003)?
- Can we use the computer experiment to test (falsify?) hypothesis about possible processes in the formation of the F/OSS developer network

# Computer Experiments

- Agent-based simulations
- Java programs using Swarm class libraries
  - Validation (docking) exercises using Java/Repast
- Grow artificial SourceForge's (Epstein & Axtell, 1996)
  - Parameterized with observed data, e.g., developer behaviors
    - Join rates
    - New project additions
    - Leave projects
  - Evaluation of multiple models (hypotheses)
- Verification/falsification (simulation and hypothesis)
  - Ensemble averages of time series data
  - Distributions
  - Chi-squared tests
  - t-Tests
  - Kolmogorov-Smirnov tests

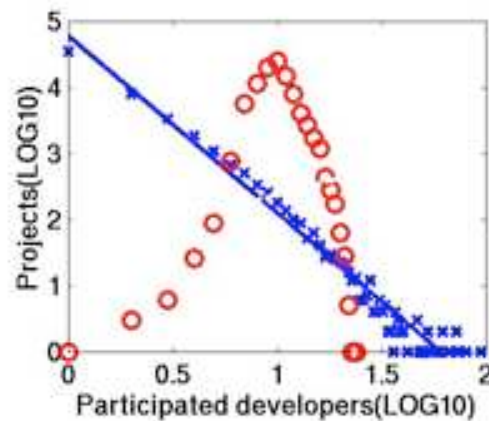
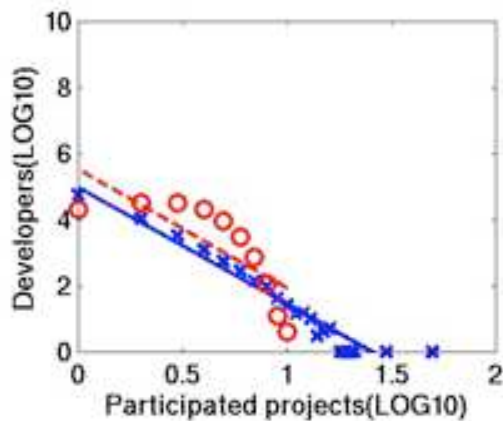
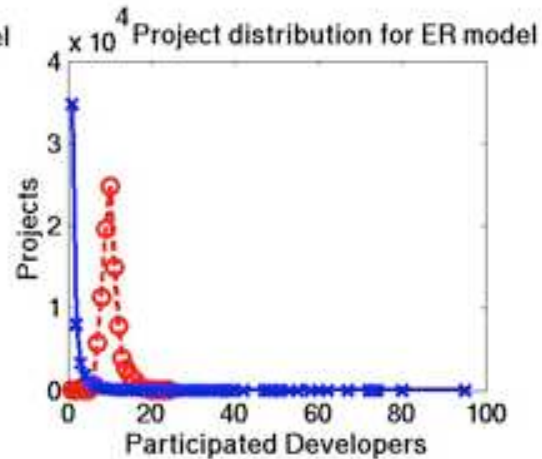
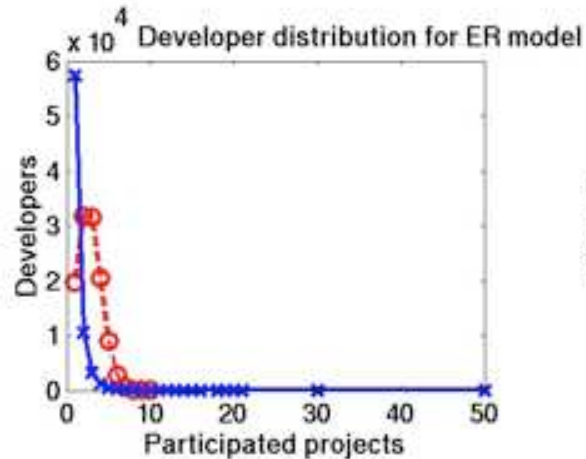
# Cycles of Modeling & Simulation



# Model for SourceForge

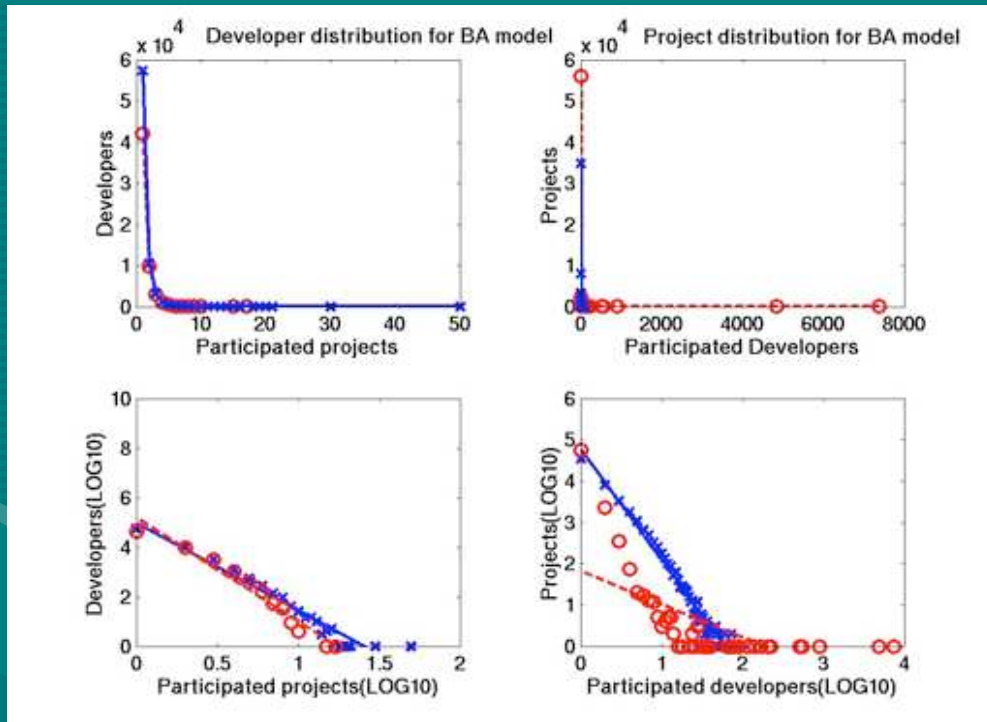
- ABM — collaborative social network
- Model description
  - Agent: developer
  - Behaviors: Create, join, abandon and idle
  - Preference: developer's and project's
  - Fitness
- Four models in iterations
  - ER, BA, BA with constant fitness and BA with dynamic fitness
- Comparison of empirical and simulated data

# ER Model – Degree Distribution



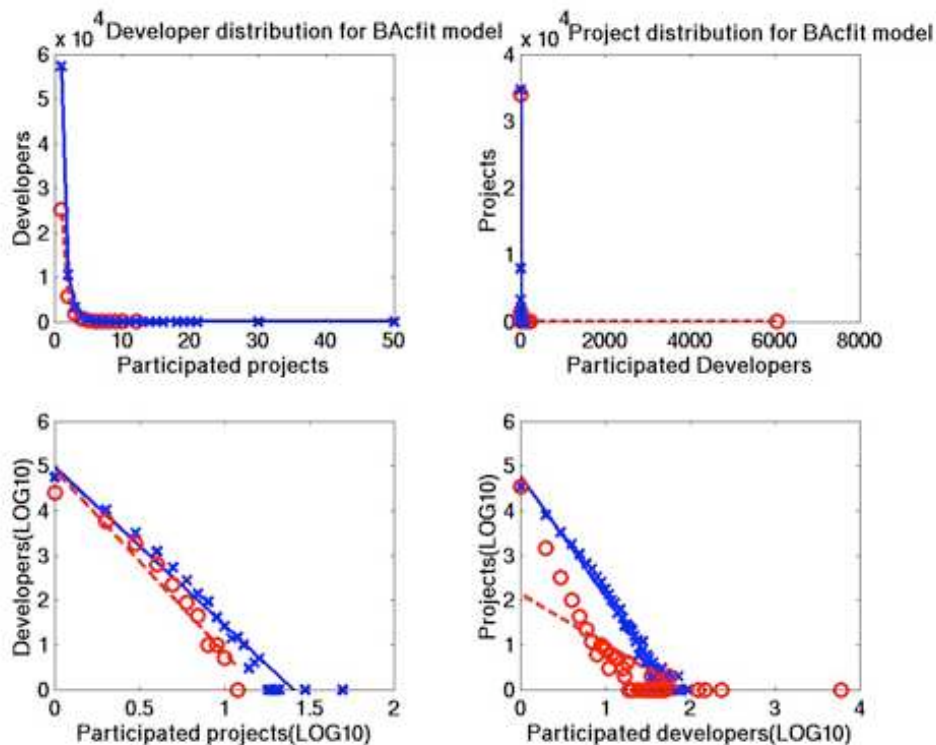
- Degree distribution is normal distribution while it is power law in empirical data
- **Fit Fails!**

# BA Model – Degree Distribution



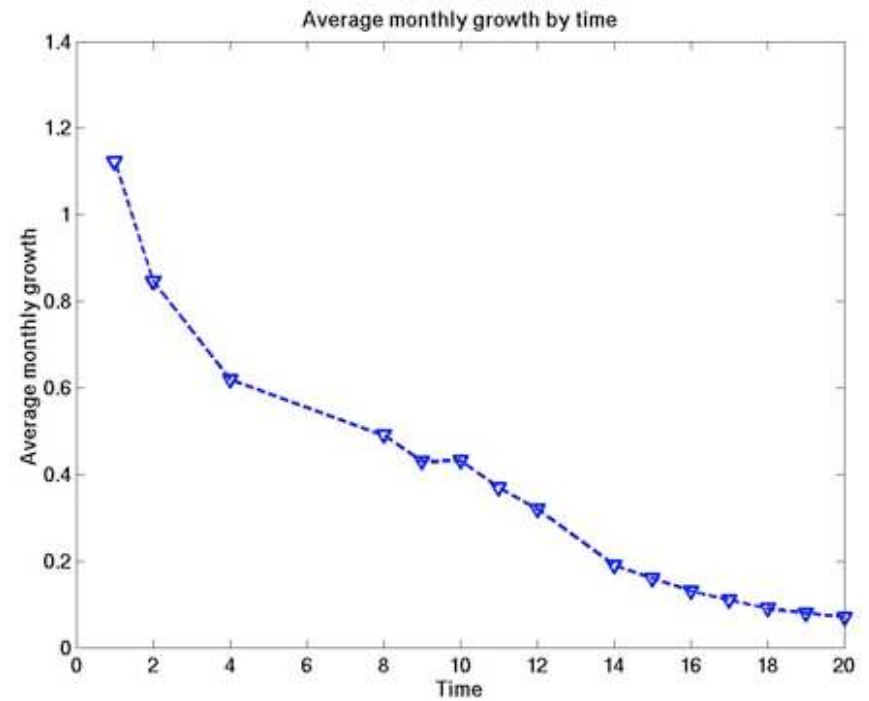
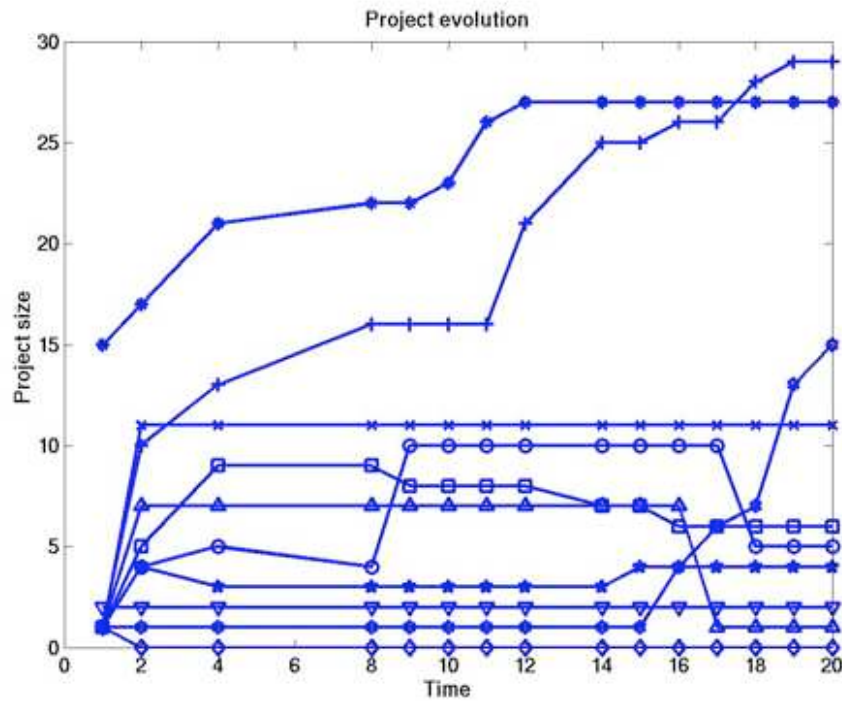
- Power laws in degree distributions, similar to empirical data (o for simulated data and x for empirical data).
- For developer distribution: simulated data has  $R^2$  as 0.9798 and empirical data has  $R^2$  as 0.9714.
- For project distribution: simulated data has  $R^2$  as 0.6650 and empirical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

# BA Model with Constant Fitness

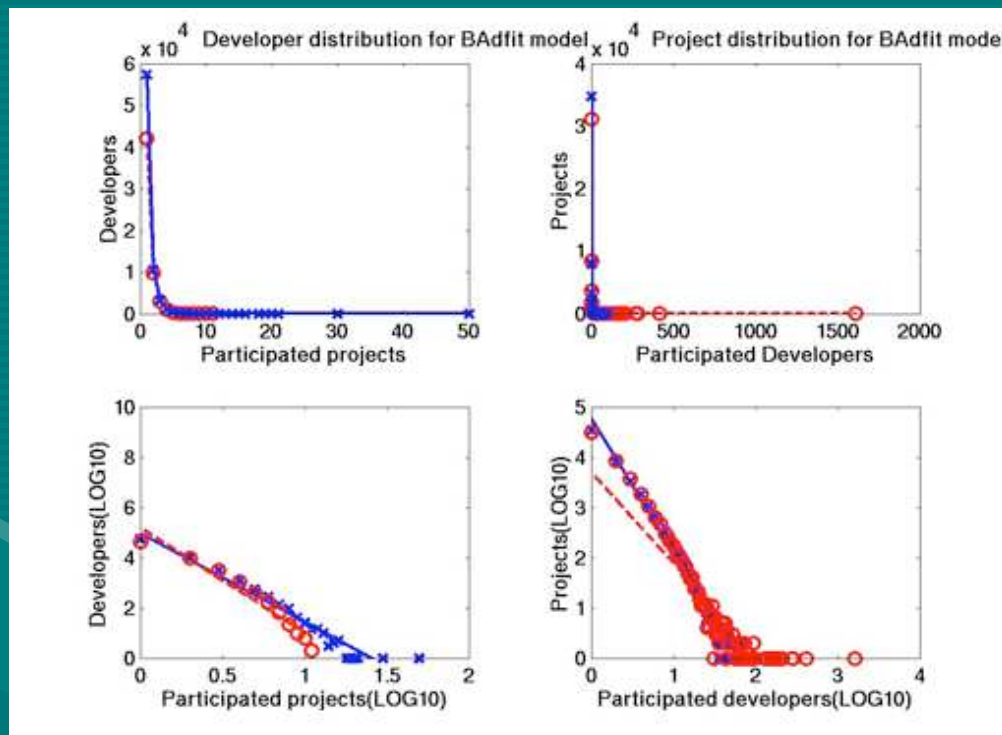


- Power laws in degree distributions, similar to empirical data (o for simulated data and x for empirical data).
- For developer distribution: simulated data has  $R^2$  as 0.9742 and empirical data has  $R^2$  as 0.9714.
- For project distribution: simulated data has  $R^2$  as 0.7253 and empirical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

# Discovery: Project Life Cycle



# 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.
- **Somewhat better fit!**

# Models of the F/OSS Social Network (Alternative Hypotheses)

- General model features
  - Agents are nodes on a graph (developers or projects)
  - Behaviors: Create, join, abandon and idle
  - Edges are relationships (joint project participation)
  - Growth of network: random or types of preferential attachment, formation of clusters
  - Fitness
  - Network attributes: diameter, average degree, degree distribution, clustering coefficient
- Four specific models
  - ER (random graph) - (1960)
  - BA (preferential attachment) - (1999)
  - BA ( + constant fitness) - (2001)
  - BA ( + dynamic fitness) - (2003)

# Discussion

- Is simulation better for falsification, but weaker at confirmation of hypotheses?
- Under what conditions can simulation results be accepted as confirmation of a hypothesis?
  - Need more validation/verification of simulations
    - Confidence in results
    - Case of computer proofs (four color problem in mathematics)
    - Need for open source/open data
      - For replication of results?
      - For docking and model-2-model comparisons
- Or is the real value of the simulation for “fishing around” for developing new hypotheses? Discovery?
  - Hidden relationships/rules-of-operations
  - Hidden features of components
  - Black-box, grey-box, white-box models
  - Discovery by reverse engineering

# Summary

- Why Agent-Based Modeling and Simulation?
  - Can be used as components of the Scientific Method
  - A research approach for studying socio-technical systems
- Case study: F/OSS - Collaboration Social Networks
  - SourceForge conceptual models: ER, BA, BA with constant fitness and BA with dynamic fitness.
  - Simulations
    - Computer experiments rejected some and confirmed plausibility of one hypothesis
    - Provided insight into the phenomenon under study and guided data mining of collected observations
    - Provided focus for additional data collection and “real experiments”.

The background is a solid teal color. In the lower half, there is a faint, semi-transparent illustration of two hands shaking, rendered in a lighter shade of teal. The text "Thank you" is centered in the upper half of the image.

Thank you