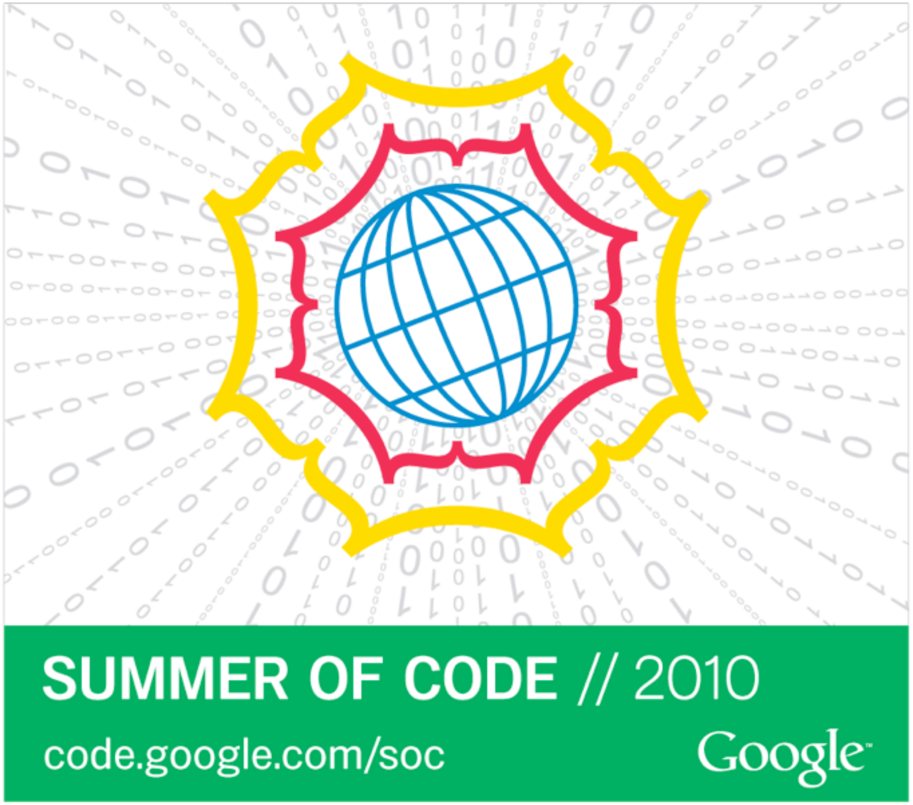
Hierarchy in Meritocracy:

Community Building and Code Production in the ASF

Oscar Castañeda

Student Delft University of Technology

This talk started with a project proposal ...



Overview

- Institutions in open source.
- Modeling behavior.
- Measuring behavior.

What are institutions?

- Rules that underlie the behavior of individuals
  - Allow for reflection at a collective level
  - Institutions can be engineered
  - But also have a natural dimension
    - » (Selznick, 1984)

What are institutions?

- Meritocracy
  - Can be interpreted as a rule:
    - *‘The more you do the more you are allowed to do.’*
  - Underlies the behavior of Apache developers

Why are institutions important?

- They can be used to distinguish between open source communities
  - ASF vs. Google Code or SourceForge
  - ASF vs. Python SF, Eclipse SF

Why are institutions important?

- Useful in decision-making
  - Graduation of an incubator project
  - Assigning roles
  - Delimiting the boundaries of an open source community

Why are institutions important?

- Delimiting the boundaries of an open source community ...
  - – Individuals co-author source code files
  - – The resulting network delimits the community
  - – Literally: community over code

Why are institutions important?

- Delimiting the boundaries of an open source community ...
  - Individuals co-author source code files
  - The resulting network delimits the community
  - ~~Literally~~: community through code

Modeling behavior

- Useful to gain a deeper understanding
  - How are communities organized?
    - e.g. Are there sub-communities?
  - How does behavior influence code production?
    - Aspects?

Modeling behavior

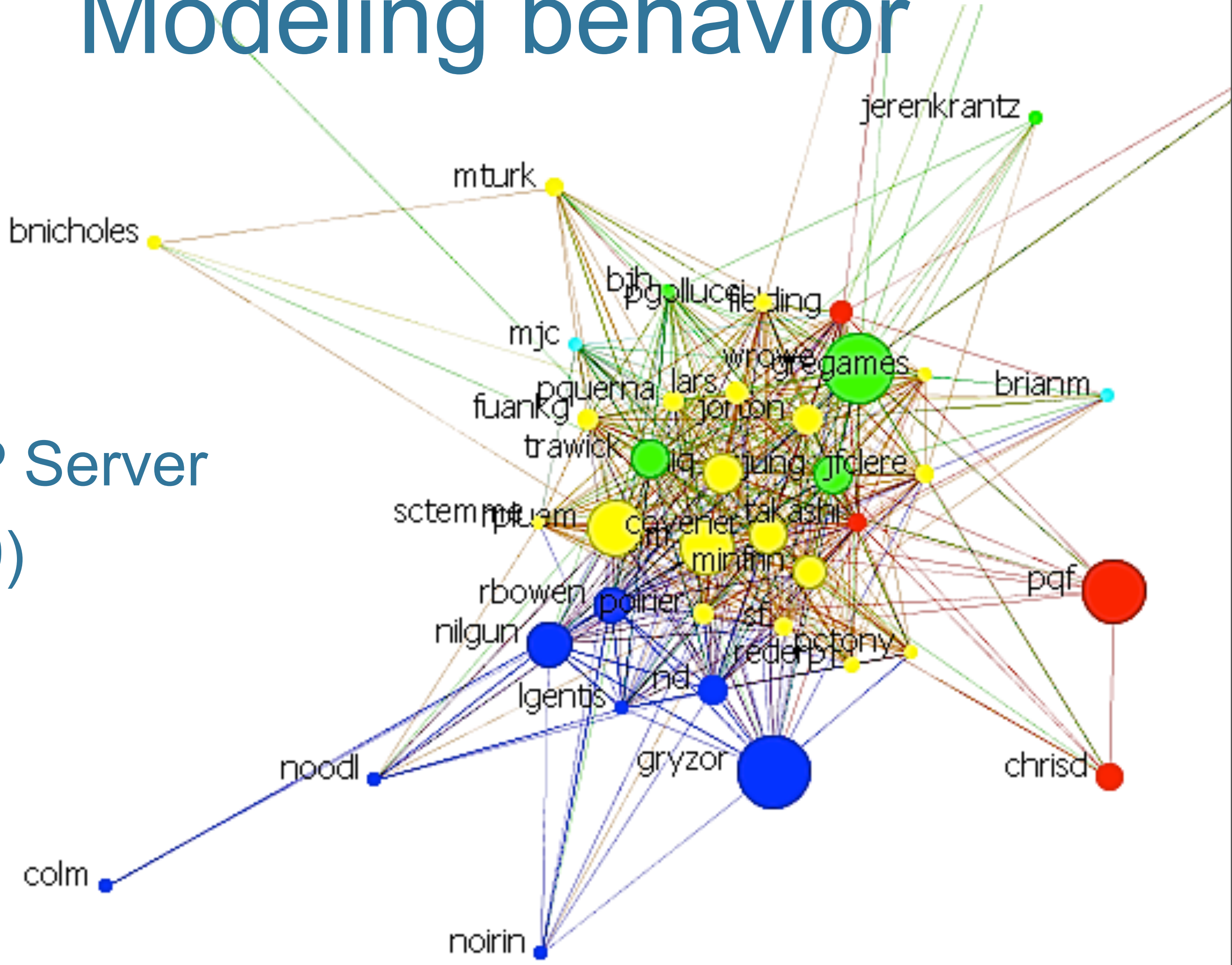
- File co-authorship
  - Social network
  - Different dimensions of institutions
  - Network-level measures

Modeling behavior

- How is the network constructed?
  - – Original author always gets incoming links
  - – Subsequent authors only get incoming links from later co-authors

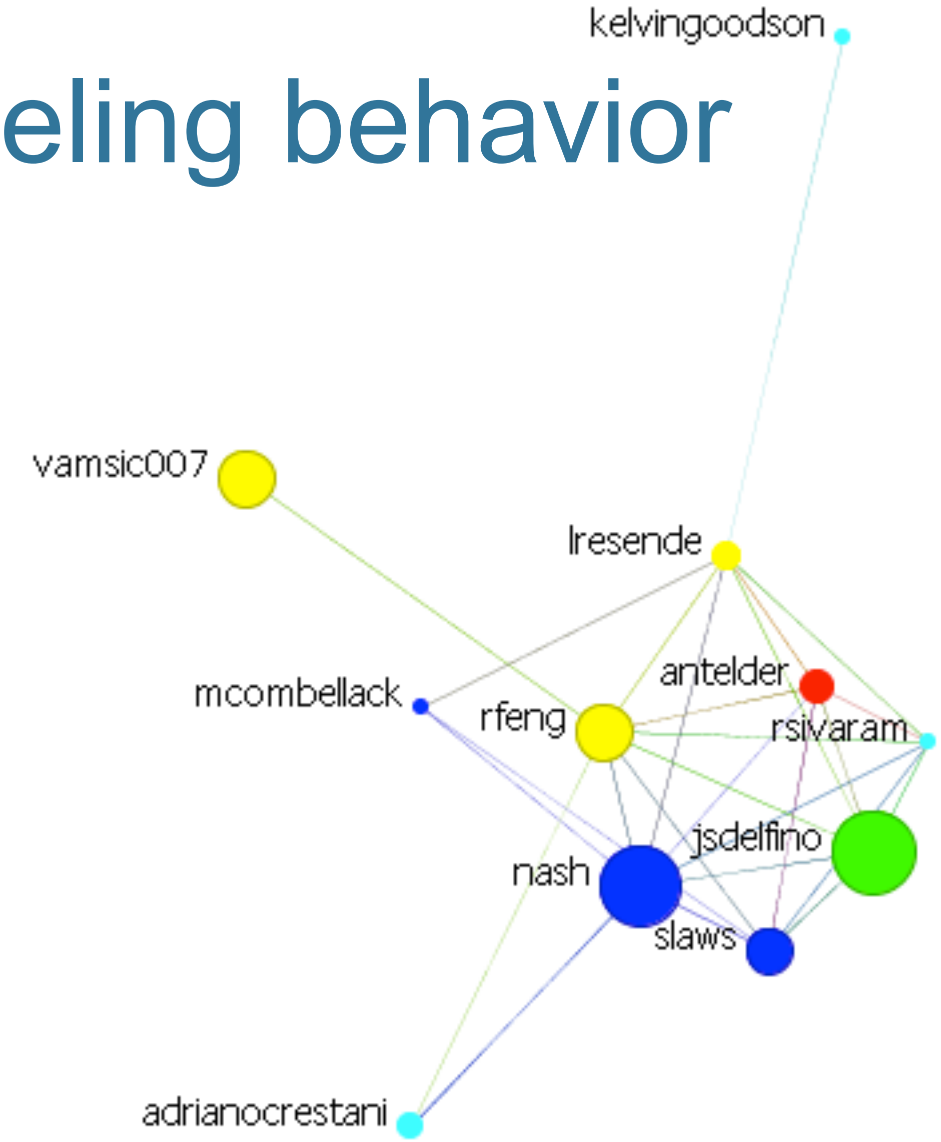
Modeling behavior

HTTP Server
(2009)



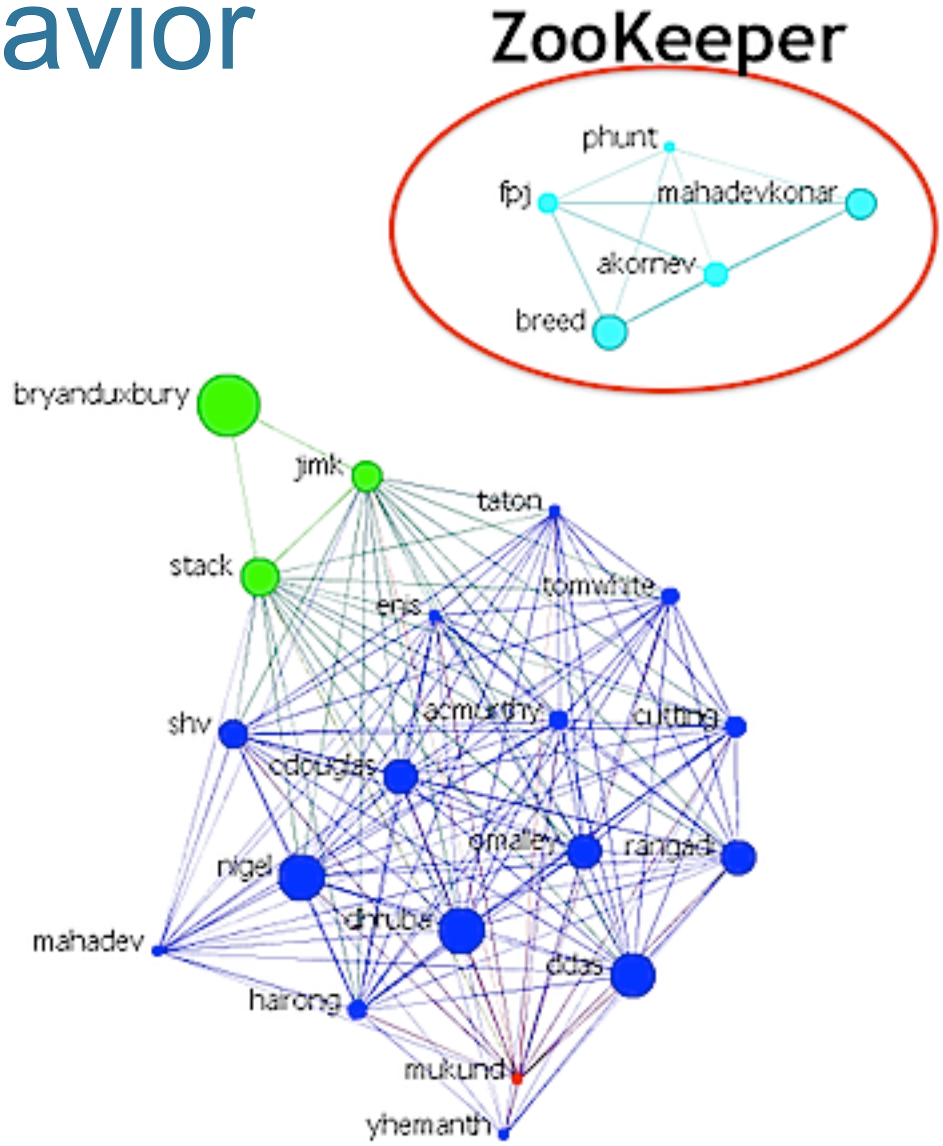
Modeling behavior

Tuscany
(2009)



Modeling behavior

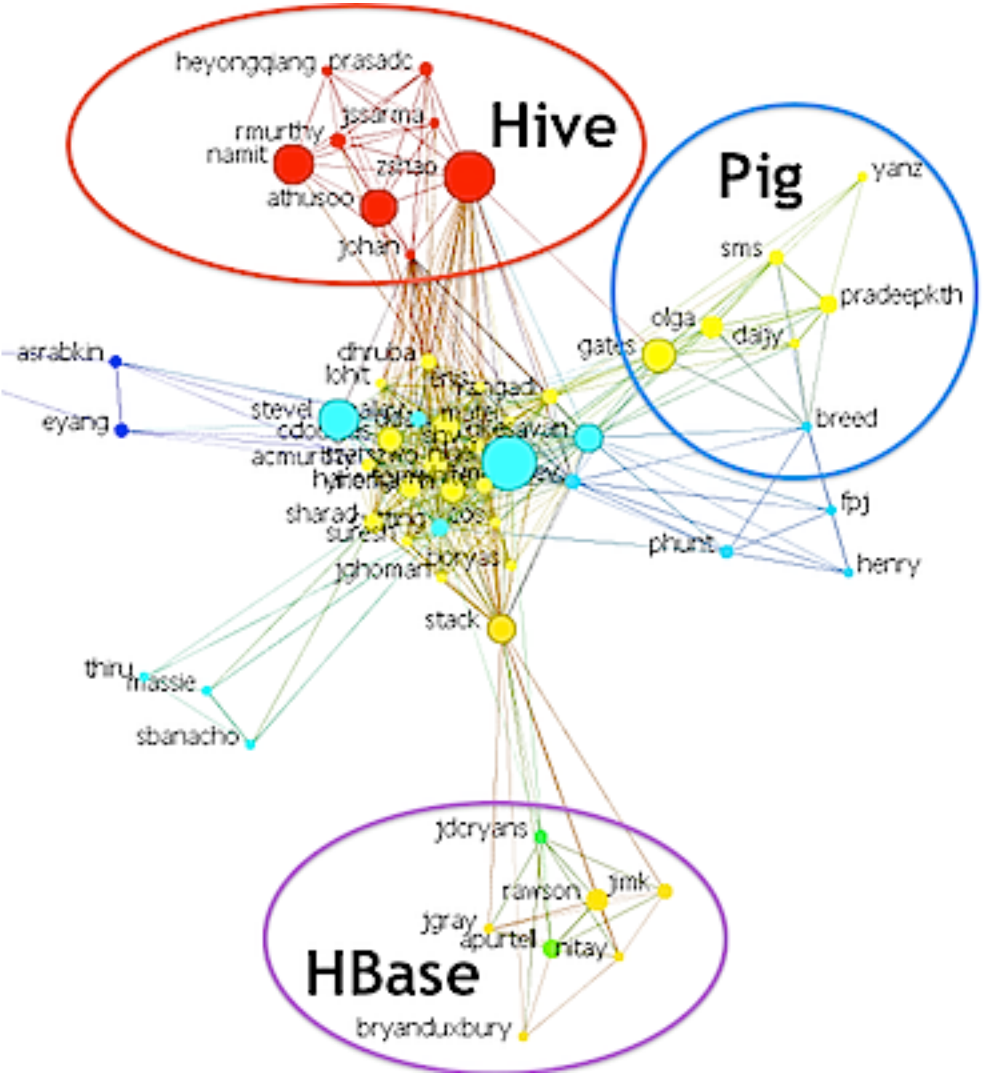
Hadoop
(2008)



Modeling behavior

Hadoop

(2009)



Modeling behavior

- What aspects were modeled?
  - Connectedness
  - Asymmetry
  - Redundancy

Modeling behavior

- Related institutions (from literature)
  - – Collective choice
  - – Conflict resolution
  - – Nested enterprise
    - » (Van Wendel de Joode, 2005)

Modeling behavior

- “If we make a chart of social interactions, of who talks to whom, the clusters of dense interaction in the chart will identify a rather well-defined hierarchic structure. The groupings in this structure may be defined operationally by some measure of frequency of interaction in this sociometric matrix”
  - » Simon (1997), pg. 186

Modeling behavior

- What other aspects were modeled?
  - Clustering
  - Average distance

Modeling behavior

- However, no related institutions
  - Self-organization
- But interesting phenomena
  - Small-world effect
    - High clustering coefficient
    - Small average distance

Small-world effect

- In 1967, Stanley Milgram:
  - Gave letters to 160 random people, each
  - addressed to a stockbroker in Boston,
  - to be delivered by first-name connections.
  - 42 letters delivered
  - 5.5 intermediaries

Washington
Oregon
Montana
North Dakota
Idaho
Minnesota
South Dakota
Wisconsin
Michigan
Vermont
Maine
New Hampshire
New York
Massachusetts
Rhode Island
Connecticut
Wyoming
Nebraska
Iowa
Pennsylvania
New Jersey
Nevada
Utah
Illinois
Ohio
West Virginia
Delaware
Maryland
California
Colorado
Indiana
Kansas
Missouri
Kentucky
Virginia
North Carolina
Tennessee
Arizona
New Mexico
Oklahoma
Arkansas
South Carolina
Georgia
Mississippi
Alabama
Texas
Louisiana
Florida
Hawaii
Alaska

Small-world effect

- Social networks tend to have short average distance between nodes
- Many highly connected nodes
  - local connections
- Some nodes also have:
  - global connections

Small-world effect

- Regular graphs
  - high clustering coefficient, long paths
  - Fully structured
- Random graphs
  - low clustering coefficient, short paths
  - Self-organized
- Small-world graphs
  - high clustering coefficient, short paths
  - Somewhere in between
      - » (Watts and Strogatz, 1998)

Measuring behavior

- Institutionalized behavior
  - Follows rules or norms
- Self-organized behavior
  - Emergent
    - ‘To measure is to know.’ -Lord Kelvin

Measuring behavior

- Sample: ~260 observations
  - Each observation = 1 project / 1 year
  - Dump of ASF Subversion repository
    - All ASF communities from 2004-2009
- Tools
  - Data mining: SVNPlot (version 0.7.0)
  - SNA: *ORA, Gephi

Measuring behavior

- What is SVNPlot?
  - A tool that creates various types of graphs and statistics from SVN logs
  - In 2 steps:
    - 1. Convert Subversion logs to sqlite3 db
    - 2. Query database to produce graphs

Measuring behavior

- Why is SVNPlot better than others?
  - Does not require ‘checked out’ repository
  - Separates data collection and report generation (2 steps).
  - Easy to write your own tools
    - In fact, that was the coding part of my GSoC project
    - Generate networks of file co-authorship from Subversion logs

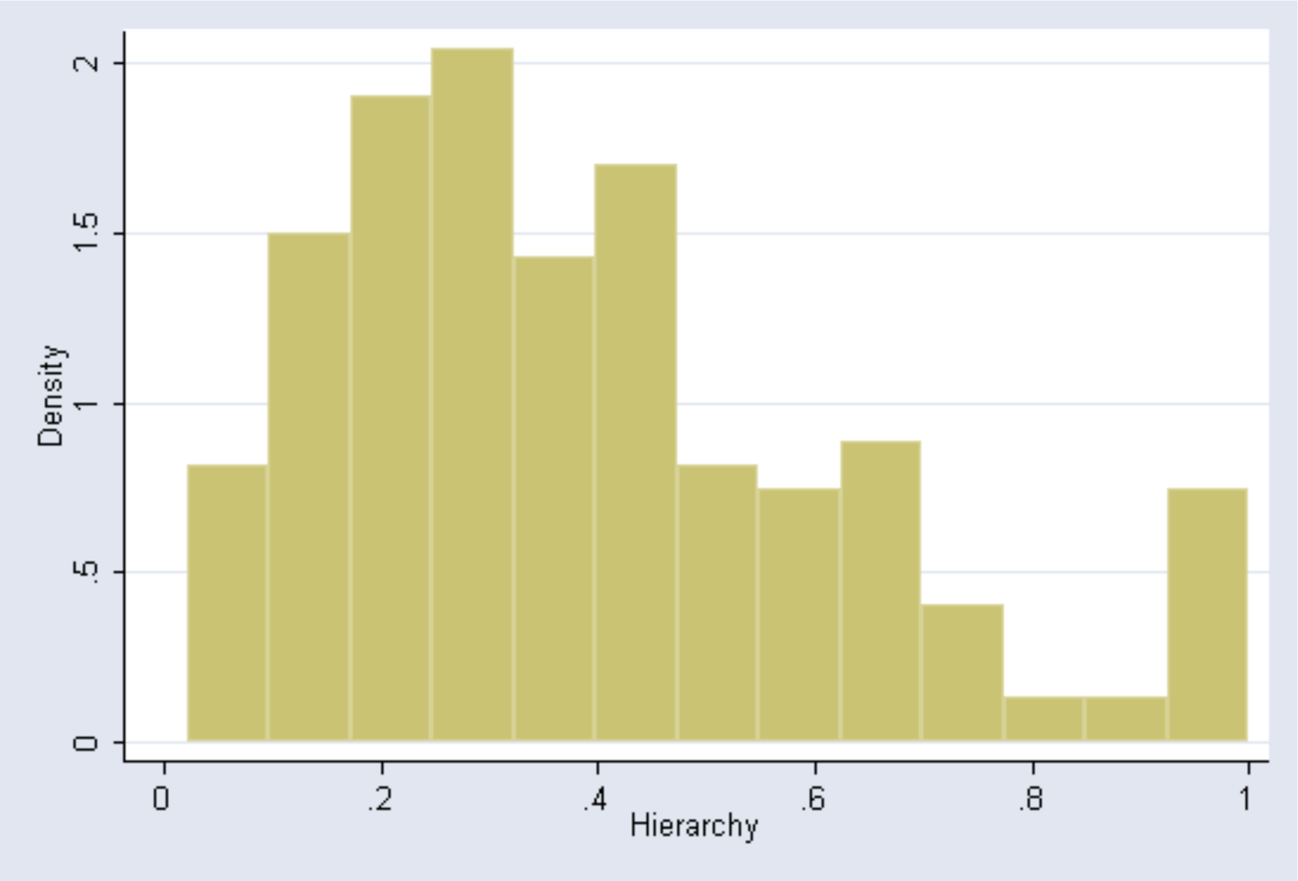
Measuring behavior

- Measures of hierarchy
  - graph hierarchy (asymmetry)
  - graph connectedness (connectedness)
  - graph efficiency (redundancy)

» (Krackhardt, 1994)

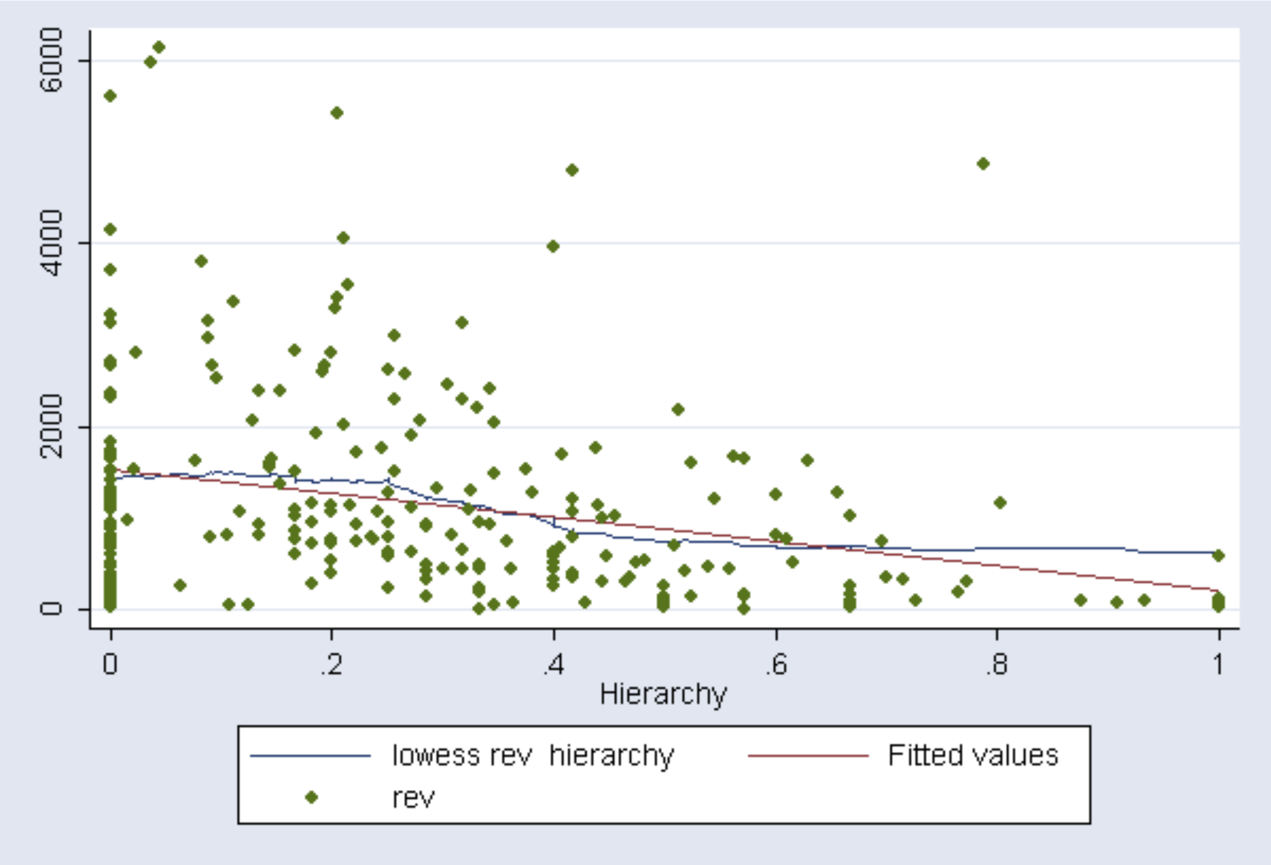
Measuring behavior

- Graph hierarchy (asymmetry)



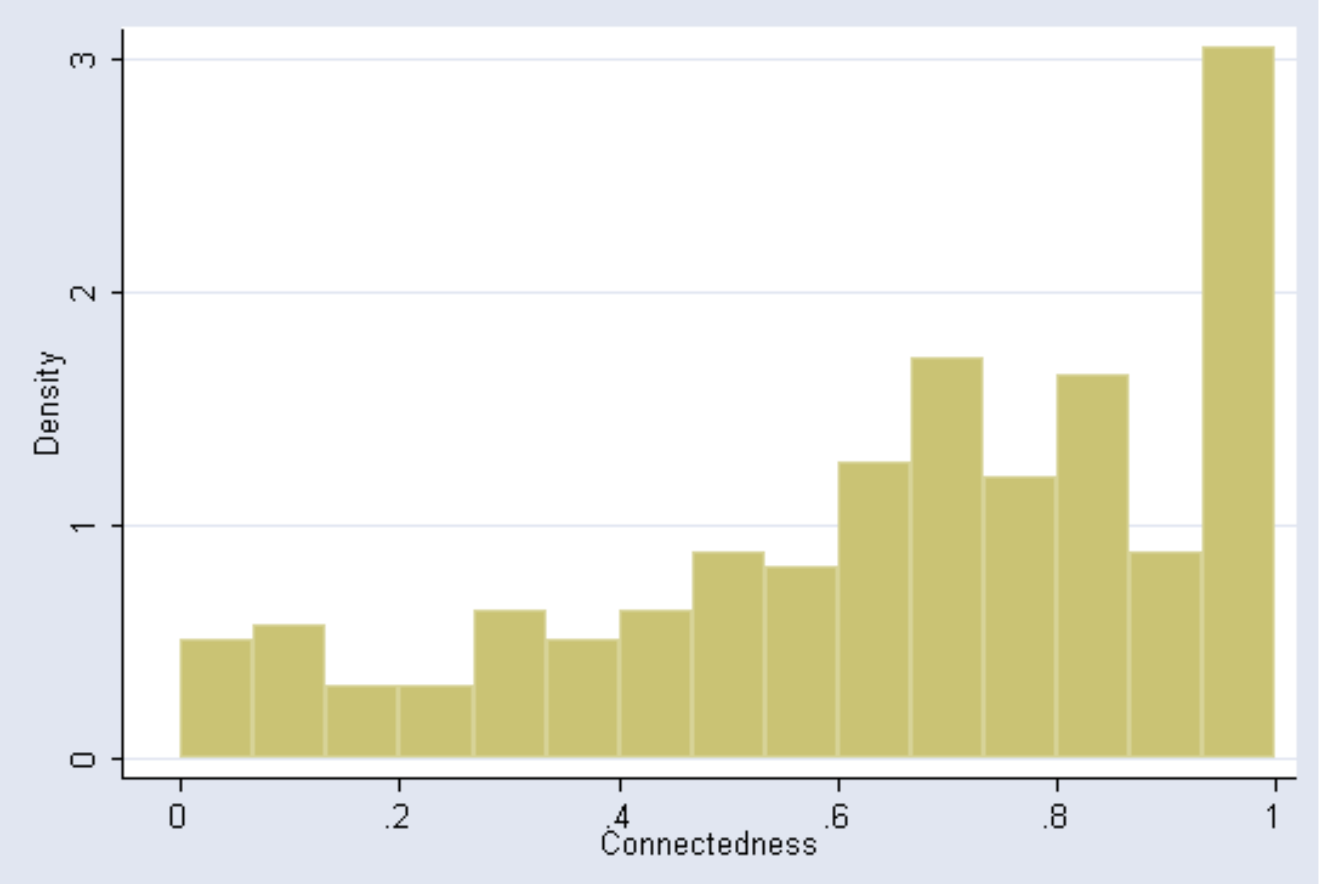
Measuring behavior

- Graph hierarchy (asymmetry)



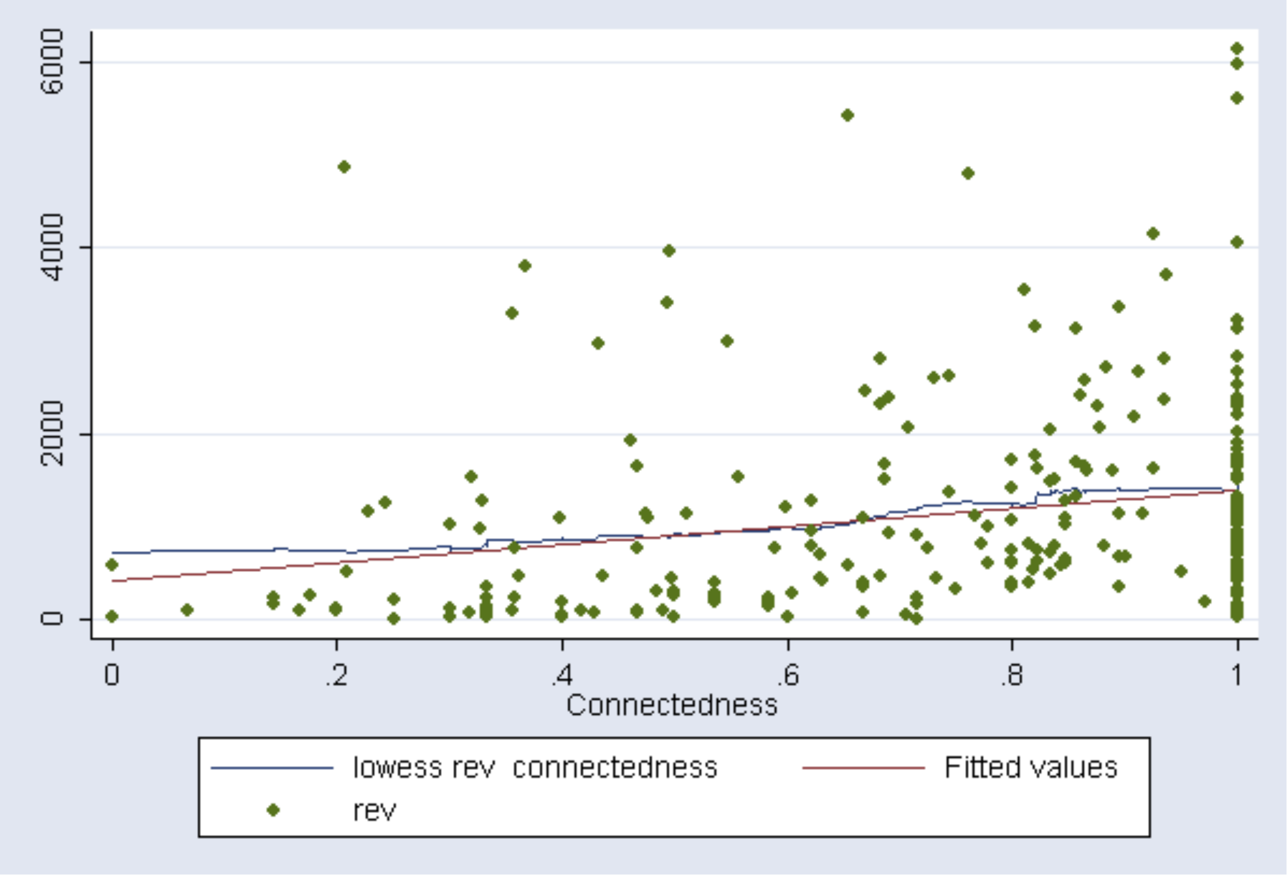
Measuring behavior

- Graph connectedness (connectedness)



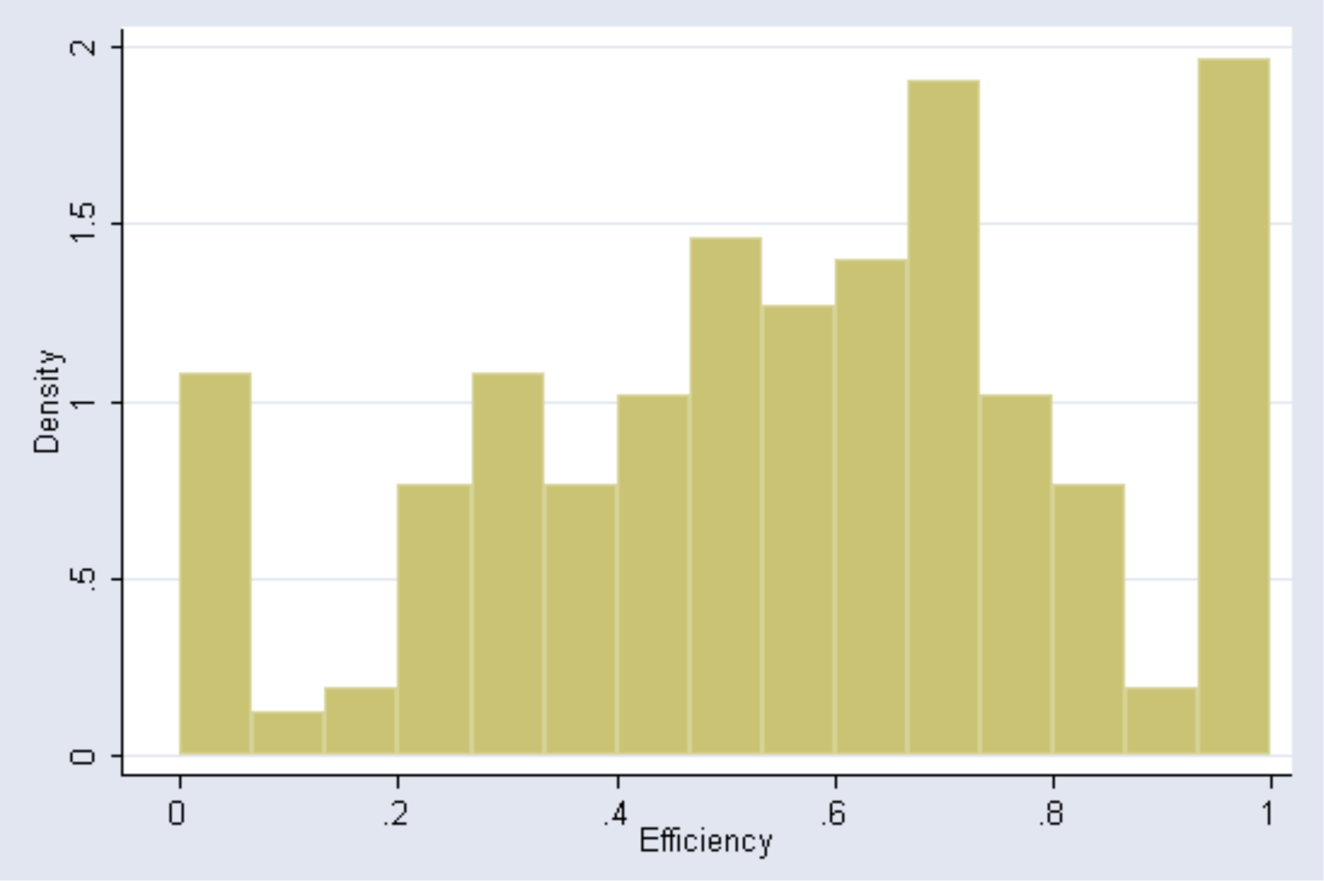
Measuring behavior

- Graph connectedness (connectedness)



Measuring behavior

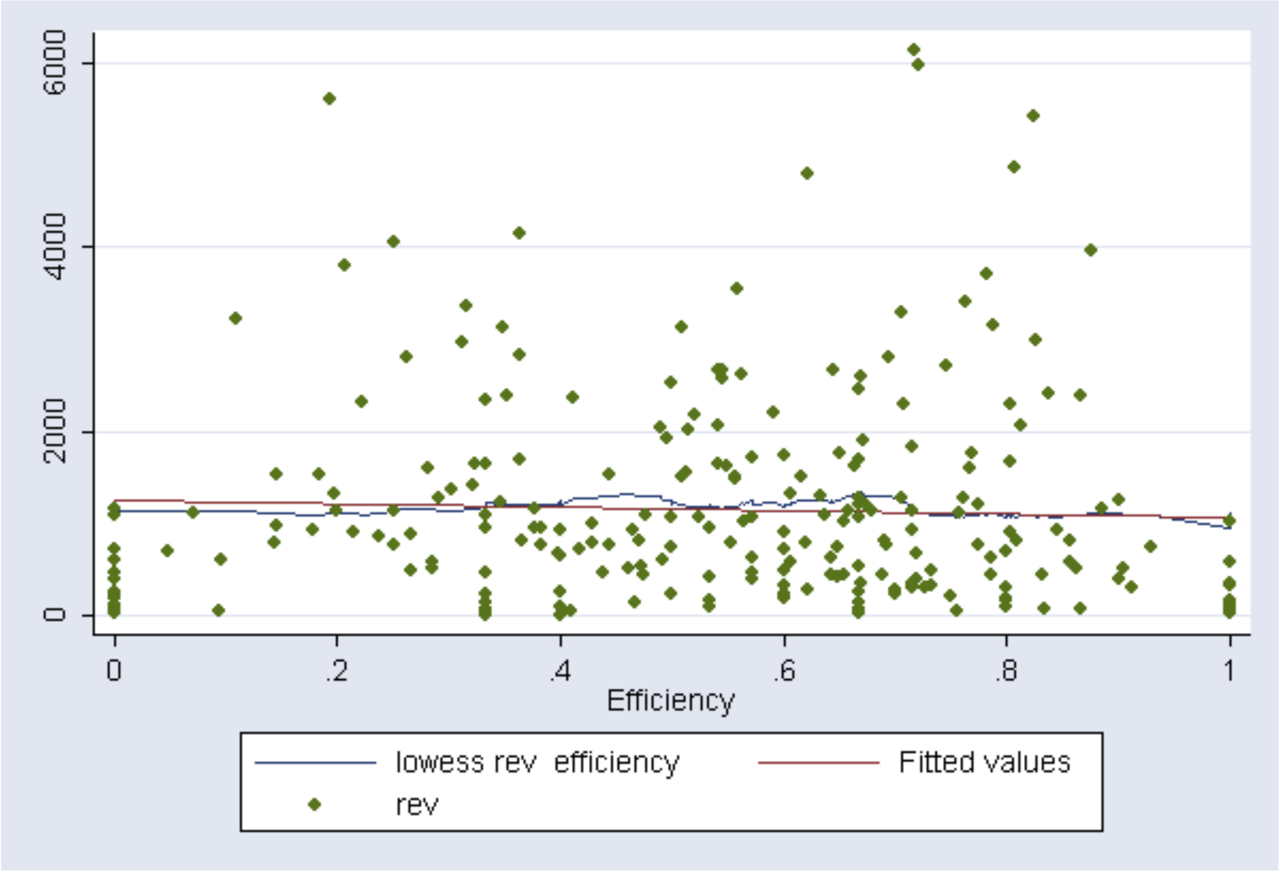
- Graph efficiency (redundancy)



Leading the Wave
of Open Source

Measuring behavior

- Graph efficiency (redundancy)

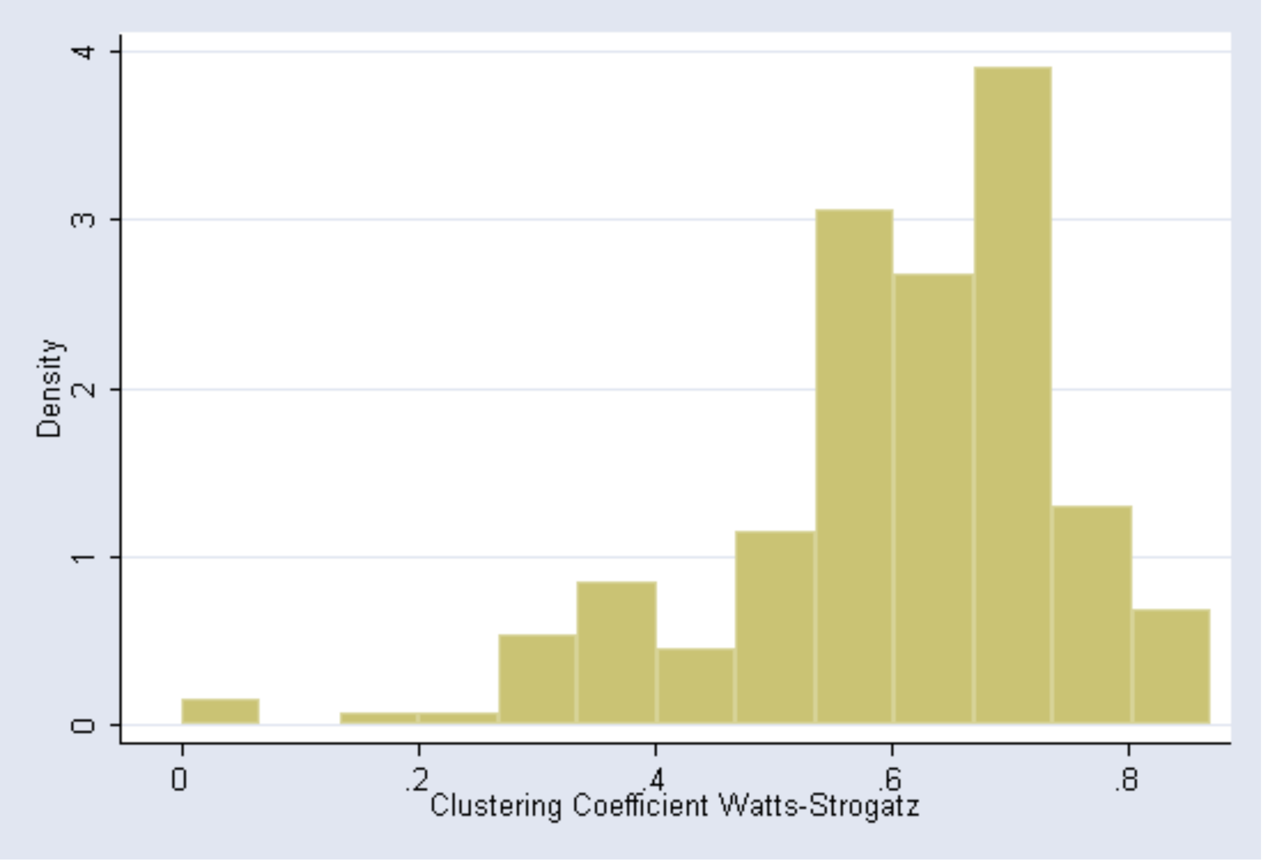


Measuring behavior

- Measures of self-organization
  - clustering coefficient
  - average distance

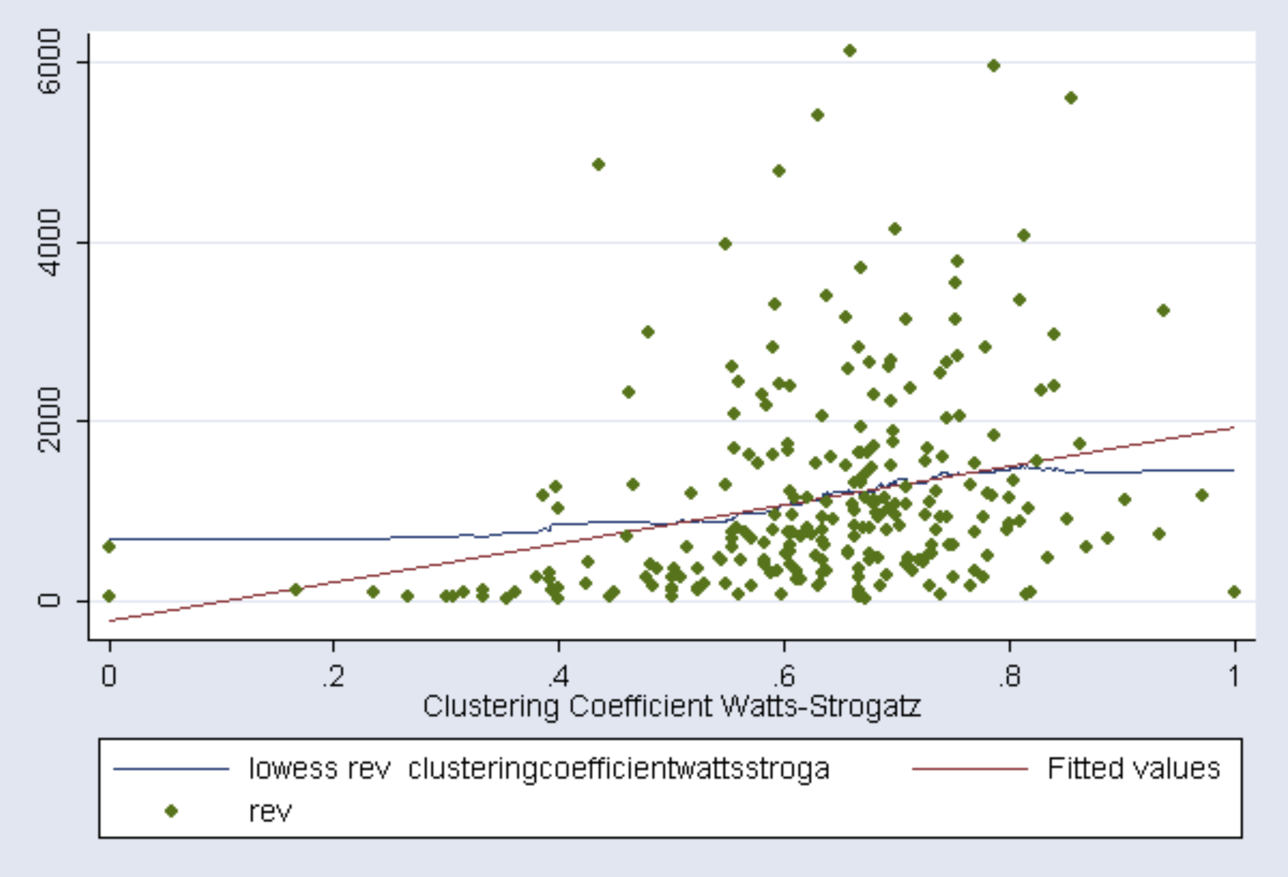
Measuring behavior

- Clustering coefficient



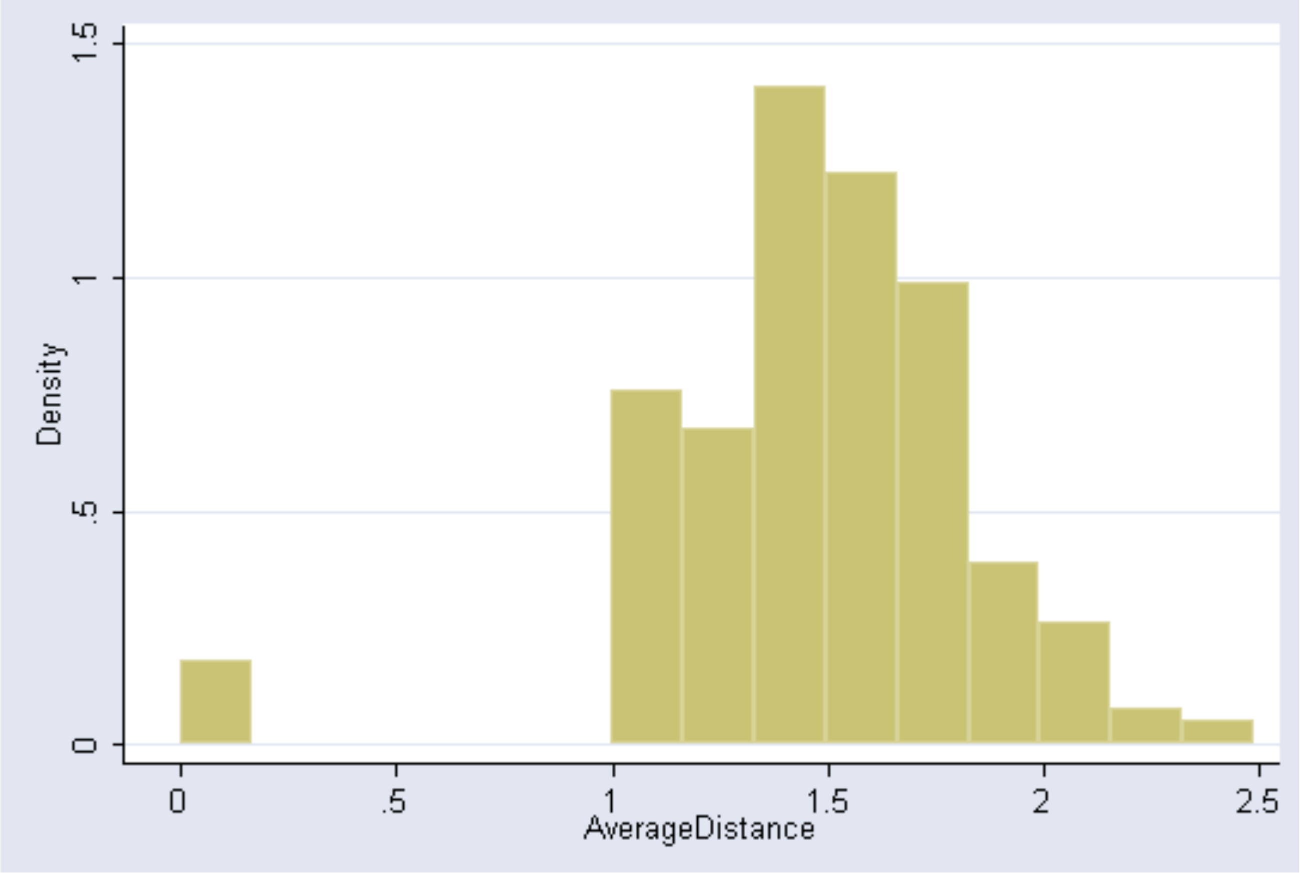
Measuring behavior

- Clustering coefficient



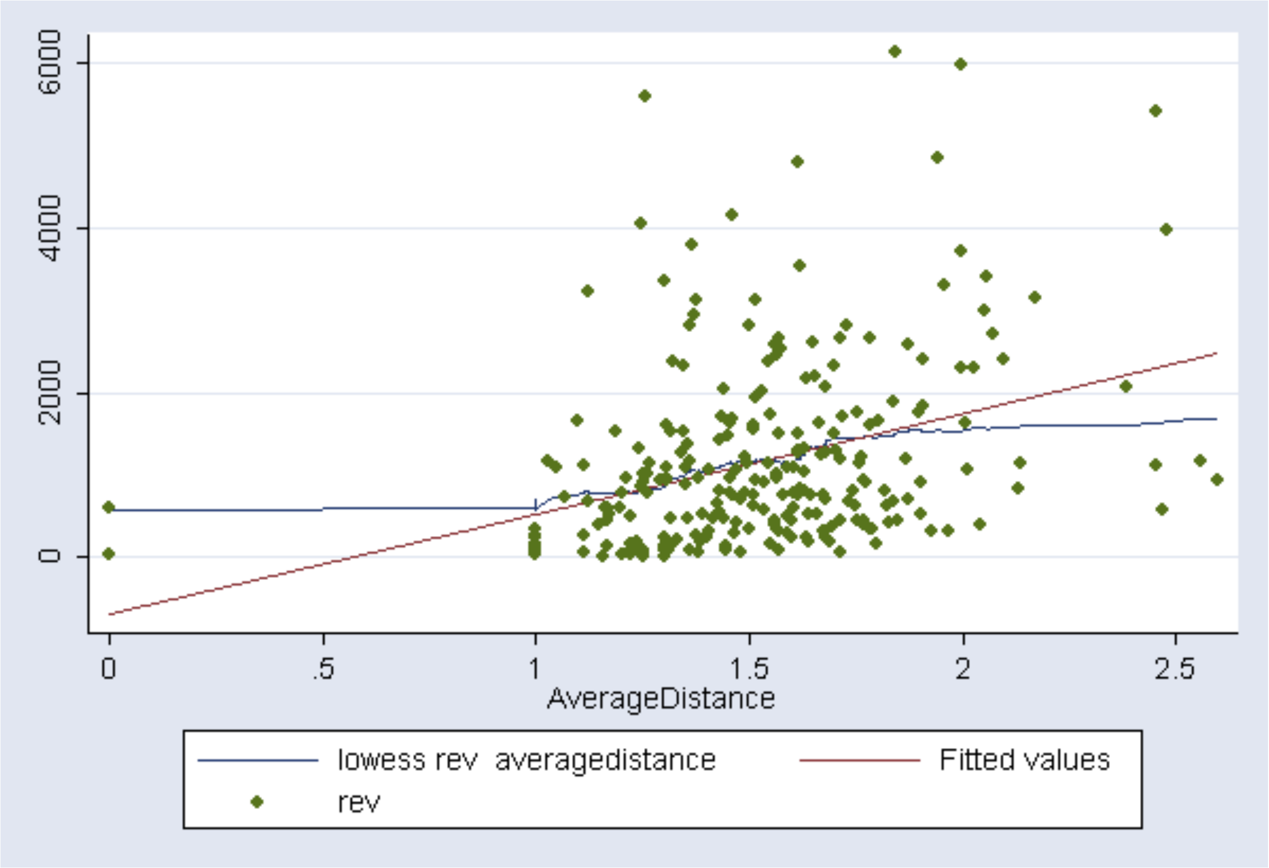
Measuring behavior

- Average distance



Measuring behavior

- Average distance



Conclusions

- Modeling and measuring behavior gives insights on code production
- Modeled aspects: varied impact on code production
- Self-organization also plays a role
  - Apache communities:
    - Highly clustered
    - 1-2 degrees of separation (low average distance)
    - Appear to be small-world networks

Future Directions

- Compare with PSF, ESF, SourceForge, Google Code
- Develop an Apache Agora script extension for SVNPlot
- Recommend files to developers based on behavior
- All data up on my Apache page:
  - http://people.apache.org/~ocastaneda/
  - Collected SVN db's data available offline.

Acknowledgements

- Charel Morris, Stone Circle Productions
- The ASF, Apache TAC
- Karl Fogel
- Tony and Daniel ASF Infrastructure team
- Nitin Bhide, Founder of SVNPlot and GSoC mentor
- Google's Open Source Programs Office

QA / Discussion

- Motivation: understand, not focus on metrics.
- Q. Does sustained code production indicate health?
  - Is a healthy community one that produces lots of code?

Thanks.