

ApacheCon

Hierarchy in Meritocracy: Community Building and Code Production in the ASF

Oscar Castañeda
Student Delft University of Technology



Leading the Wave
of Open Source

1

ApacheCon

This talk started with a project
proposal ...



SUMMER OF CODE // 2010
code.google.com/soc Google



Leading the Wave
of Open Source

2

ApacheCon

Overview

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



Leading the Wave
of Open Source

3

ApacheCon

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



Leading the Wave
of Open Source

4

ApacheCon

What are institutions?

- A well-known example is...
 - Meritocracy**
 - *'The more you do the more you are allowed to do.'*



Leading the Wave
of Open Source

5

ApacheCon

Why are institutions important?

- They distinguish one community from another
 - ASF vs. Google code or Sourceforge
 - ASF vs. Python SF, Eclipse SF



Leading the Wave
of Open Source

6

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



Modeling behavior

- Needed for deeper understanding of behavior ...
 - How organized?
 - Influence on code production?



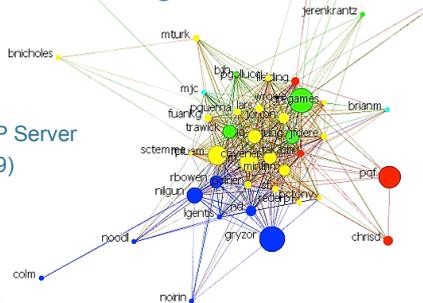
Modeling behavior

- We have a network of file co-authorship
 - Model institutions as dimensions in that social network
 - Network-level measures



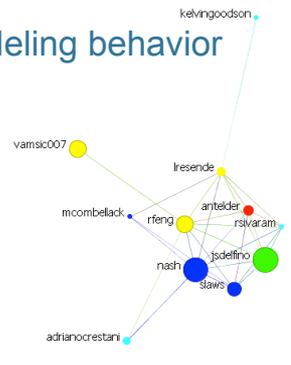
Modeling behavior

HTTP Server (2009)



Modeling behavior

Tuscany (2009)



Measuring behavior

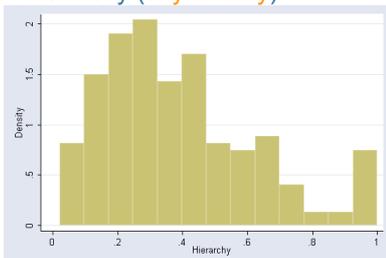
- Sample: ~260 observations
 - Dump of ASF Subversion repository
 - <http://svn-master.apache.org/dump>
 - All ASF communities from 2004-2009
- Tools
 - Data mining: SVNPlot (version 0.7.0)
 - SNA: CMU's *ORA, Gephi
 - Statistics: R and Stata

Measuring behavior

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

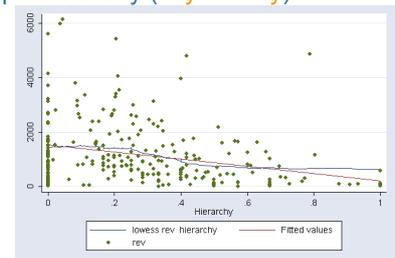
Measuring behavior

- Graph hierarchy (asymmetry)



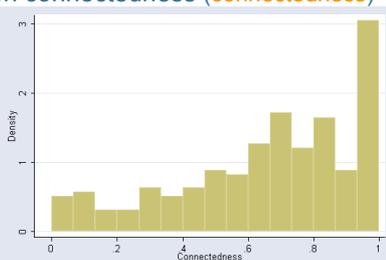
Measuring behavior

- Graph hierarchy (asymmetry)



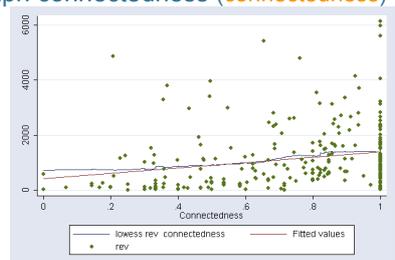
Measuring behavior

- Graph connectedness (connectedness)



Measuring behavior

- Graph connectedness (connectedness)



ApacheCon

Measuring behavior

- Graph efficiency (redundancy)

A histogram showing the density distribution of Graph efficiency (redundancy). The x-axis is labeled 'Efficiency' and ranges from 0 to 1. The y-axis is labeled 'Density' and ranges from 0 to 2. The distribution is unimodal and slightly right-skewed, peaking at an efficiency of approximately 0.7 with a density of about 1.8.

Leading the Wave of Open Source

25

ApacheCon

Measuring behavior

- Graph efficiency (redundancy)

A scatter plot showing the relationship between Graph efficiency (redundancy) and another variable. The x-axis is labeled 'Efficiency' and ranges from 0 to 1. The y-axis ranges from 0 to 6000. The plot shows a dense cloud of green data points with a fitted curve (red line) and a legend indicating 'lowess rev efficiency' and 'Fitted values'.

Leading the Wave of Open Source

26

ApacheCon

Measuring behavior

- Measures of clustering
 - clustering coefficient
 - average distance

Leading the Wave of Open Source

27

ApacheCon

Measuring behavior

- Clustering coefficient

A histogram showing the density distribution of the Clustering coefficient. The x-axis is labeled 'Clustering Coefficient Watts-Strogatz' and ranges from 0 to 1. The y-axis is labeled 'Density' and ranges from 0 to 4. The distribution is unimodal and right-skewed, peaking at a clustering coefficient of approximately 0.7 with a density of about 3.8.

Leading the Wave of Open Source

28

ApacheCon

Measuring behavior

- Clustering coefficient

A scatter plot showing the relationship between Clustering coefficient and another variable. The x-axis is labeled 'Clustering Coefficient Watts-Strogatz' and ranges from 0 to 1. The y-axis ranges from 0 to 6000. The plot shows a dense cloud of green data points with a fitted curve (red line) and a legend indicating 'lowess rev clusteringcoefficientwattsstrogatz' and 'Fitted values'.

Leading the Wave of Open Source

29

ApacheCon

Measuring behavior

- Average distance

A histogram showing the density distribution of Average distance. The x-axis is labeled 'AverageDistance' and ranges from 0 to 2.5. The y-axis is labeled 'Density' and ranges from 0 to 1.5. The distribution is unimodal and right-skewed, peaking at an average distance of approximately 1.5 with a density of about 1.4.

Leading the Wave of Open Source

30

Measuring behavior

- Average distance

The scatter plot displays 'AverageDistance' on the x-axis (ranging from 0 to 2.5) and 'rev' on the y-axis (ranging from 0 to 6000). A dense cluster of green data points is visible, with a red line representing 'Fitted values' showing a positive correlation. A legend at the bottom identifies the plot elements: 'lowess rev' (blue line), 'averagedistance' (green dots), and 'Fitted values' (red line).

Leading the Wave of Open Source

31

Conclusions

- Modeling and measuring behavior gives insights into code production
- Some institutions have a negative impact on code production
- Other institutions have a positive influence
- Self-organization also plays a role

Leading the Wave of Open Source

32

Future Directions

- Propose an Apache Lab
- Develop an Apache Agora script extension for SVNPlot
- Recommendation mining using Apache Mahout: recommend files to developers based on behavior

Leading the Wave of Open Source

33

QA / Discussion

Leading the Wave of Open Source

34

Acknowledgements

- Charel Morris, Stone Circle Productions.
- Nitin Bhide, Founder of SVNPlot and GSoC mentor.
- Google's Open Source Programs Office.

Leading the Wave of Open Source

35

Thanks.

Leading the Wave of Open Source

36