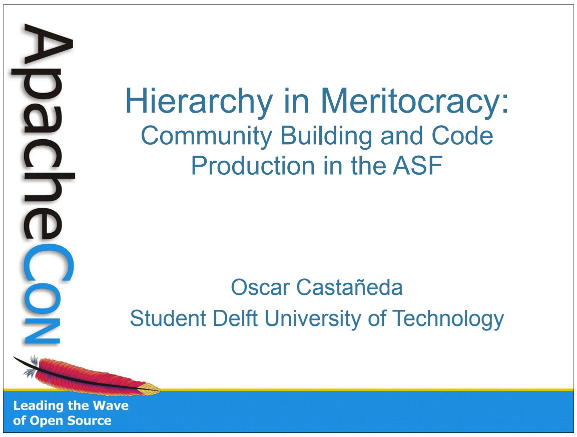
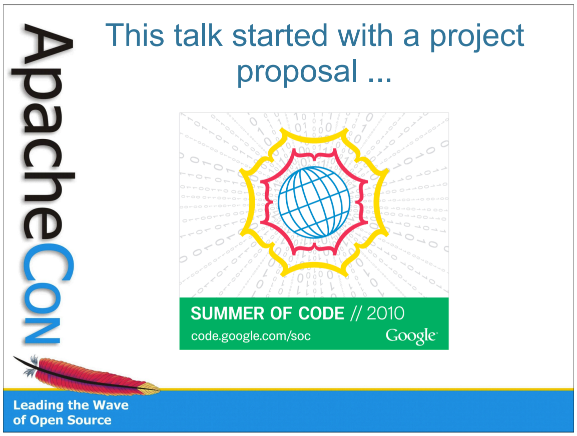


Hierarchy in Meritocracy:

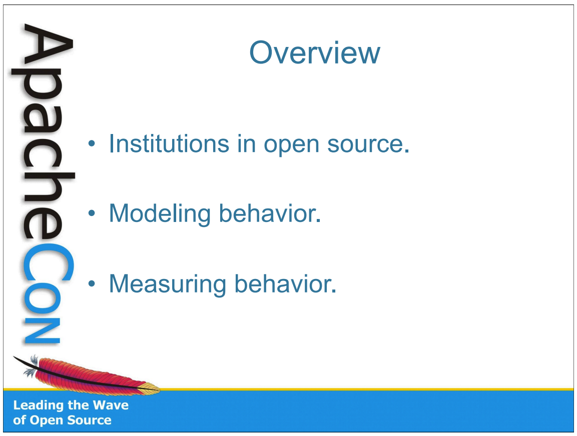
Community Building and Code Production in the ASF

Oscar Castañeda
Student Delft University of Technology



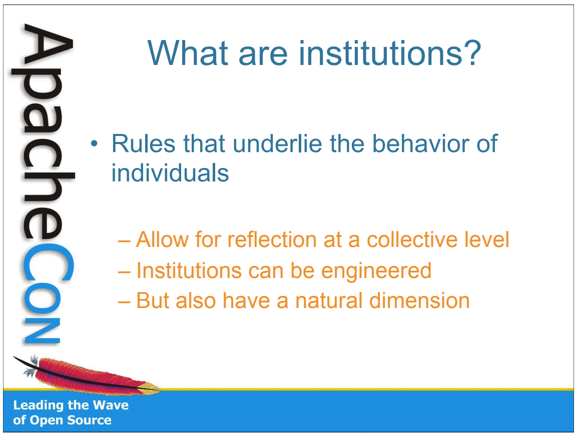
This talk started with a project proposal ...

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



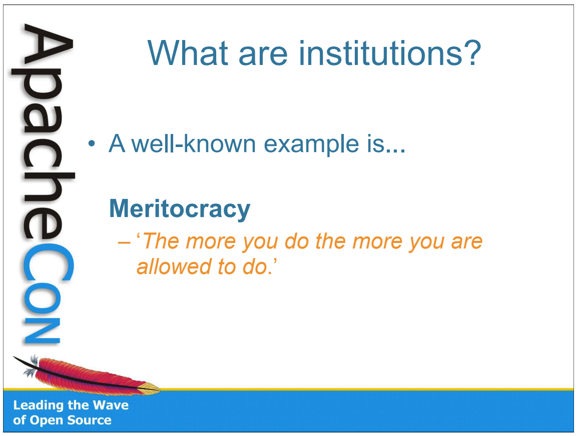
Overview

- 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

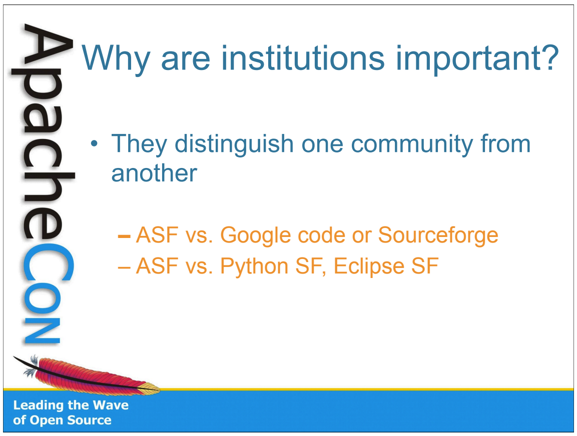


What are institutions?

- A well-known example is...

Meritocracy

- *'The more you do the more you are allowed to do.'*



Why are institutions important?

- They distinguish one community from another
  - 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

Modeling behavior

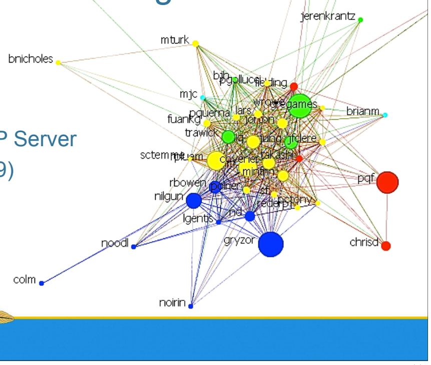
- Needed for deeper understanding of behavior ...
  - How organized?
  - Influence on code production?

Modeling behavior

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

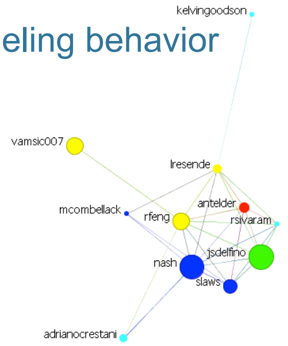
Modeling behavior

HTTP Server
(2009)



Modeling behavior

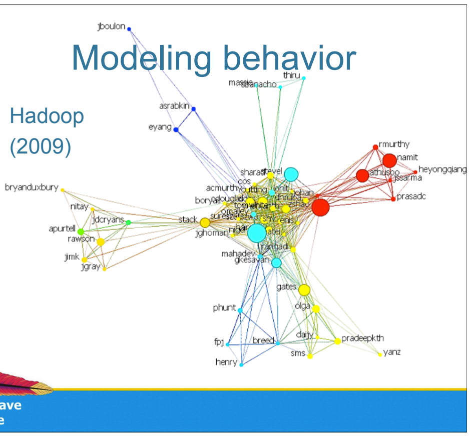
Tuscany
(2009)



Modeling behavior

Hadoop

(2009)



Modeling behavior

- What aspects can be modeled?
  - Connectedness
  - Asymmetry
  - Redundancy

Modeling behavior

- Related institutions
  - Collective choice
  - Conflict resolution
  - Nested enterprise

Modeling behavior

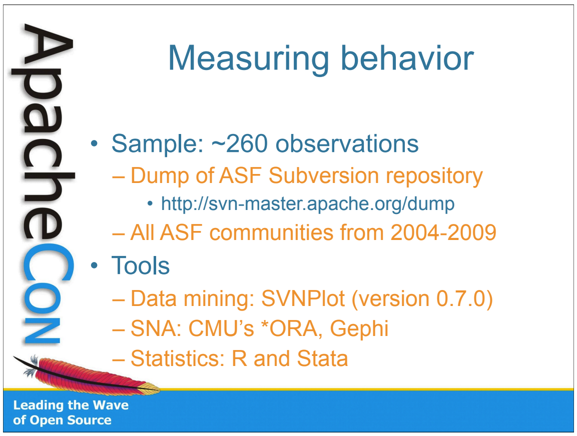
- What other aspects can be modeled?
  - Clustering
  - Average distance

Modeling behavior

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

Measuring behavior

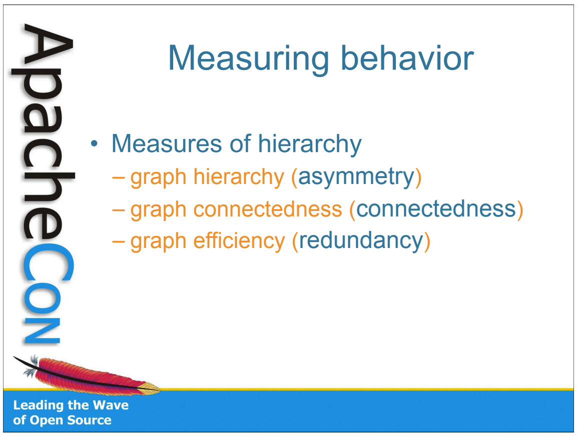
- Institutionalized behavior
  - Follows rules or norms
- Self-organized behavior
  - Emergent



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

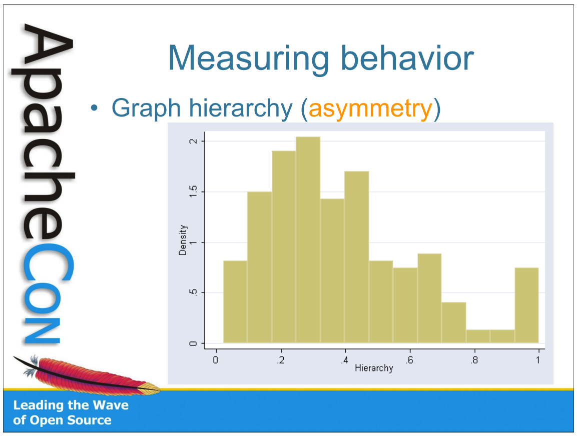
Leading the Wave
of Open Source



Measuring behavior

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

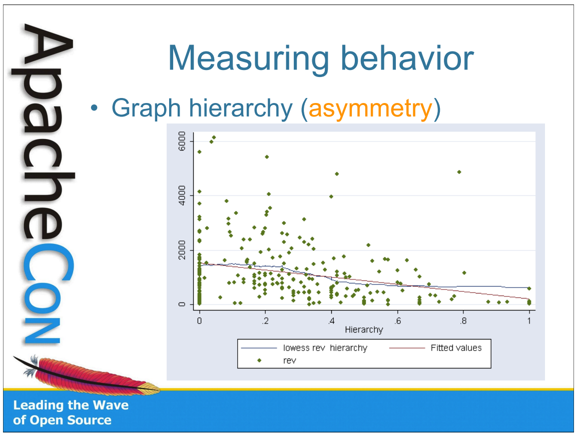
Leading the Wave
of Open Source



Measuring behavior

- Graph hierarchy (asymmetry)

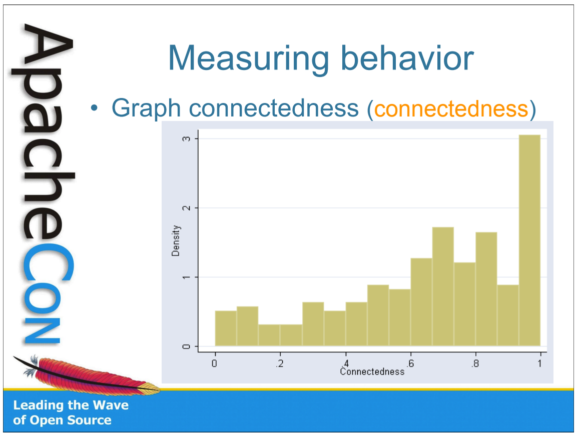
Leading the Wave
of Open Source



Measuring behavior

- Graph hierarchy (asymmetry)

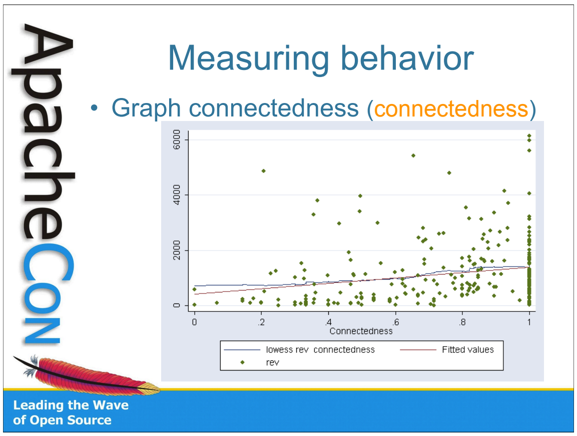
Leading the Wave
of Open Source



Measuring behavior

- Graph connectedness (connectedness)

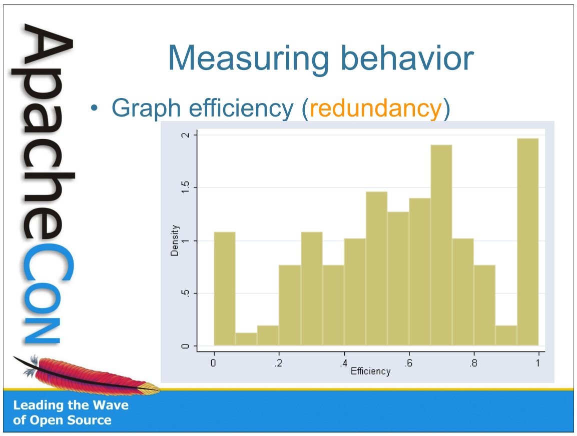
Leading the Wave
of Open Source



Measuring behavior

- Graph connectedness (connectedness)

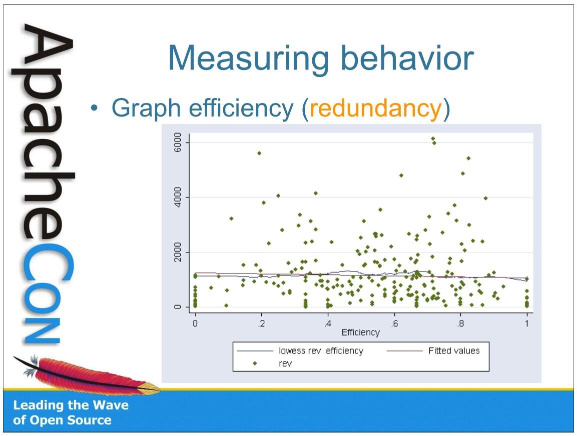
Leading the Wave
of Open Source



Measuring behavior

- Graph efficiency (redundancy)

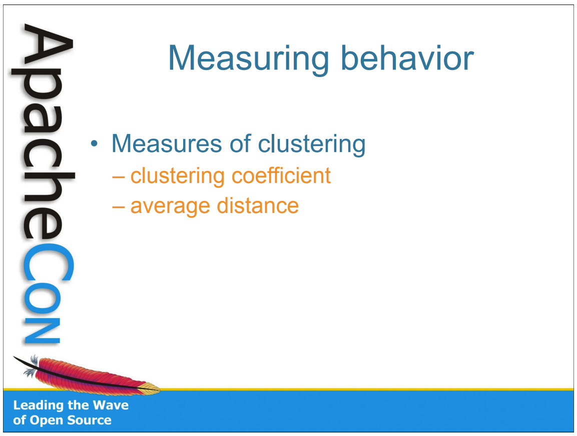
Leading the Wave of Open Source



Measuring behavior

- Graph efficiency (redundancy)

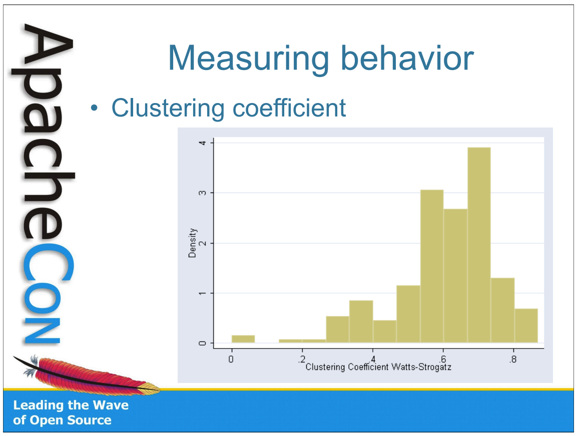
Leading the Wave of Open Source



Measuring behavior

- Measures of clustering
  - – clustering coefficient
  - – average distance

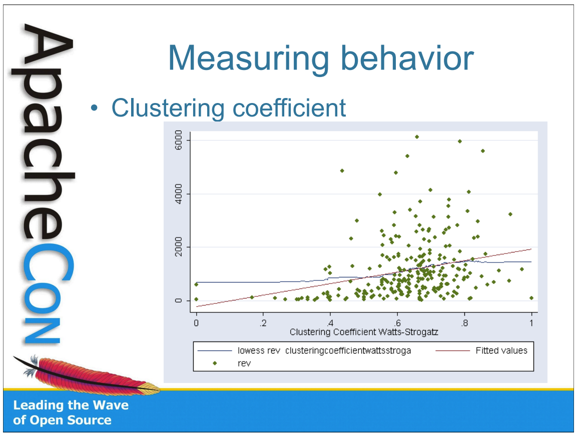
Leading the Wave of Open Source



Measuring behavior

- Clustering coefficient

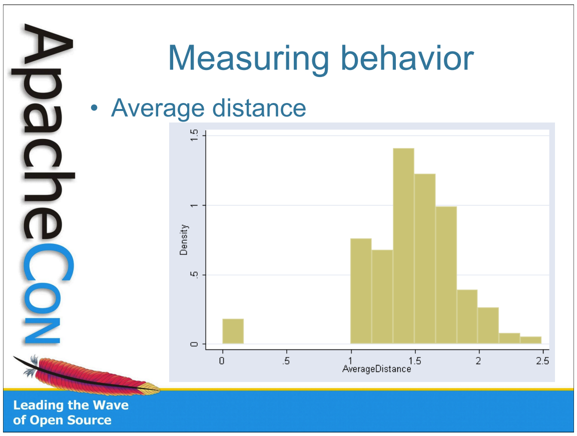
Leading the Wave of Open Source



Measuring behavior

- Clustering coefficient

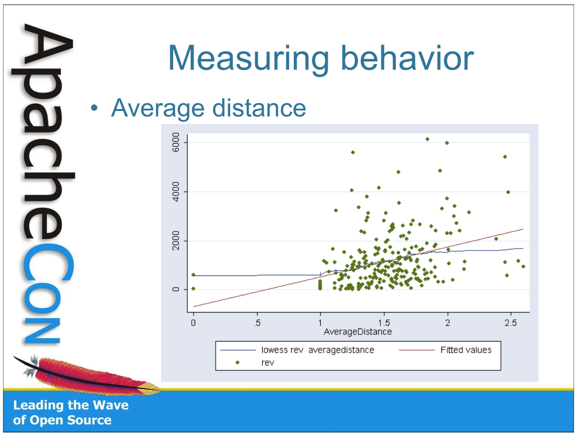
Leading the Wave of Open Source



Measuring behavior

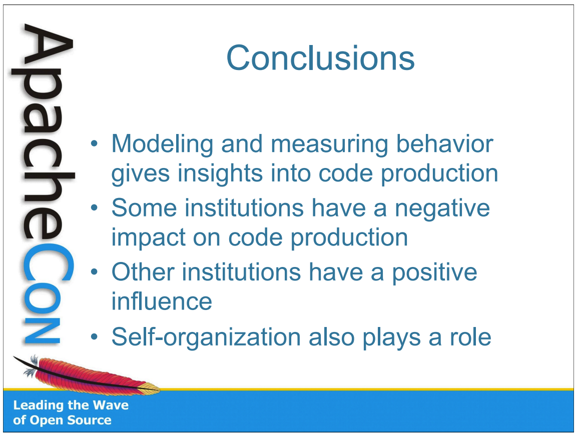
- Average distance

Leading the Wave of Open Source



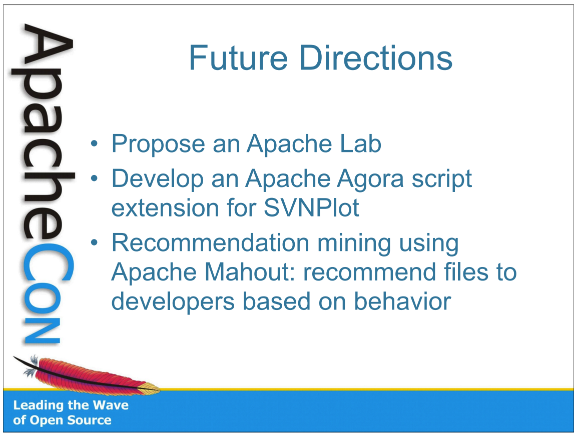
Measuring behavior

- Average distance


ApacheCon
Leading the Wave of Open Source

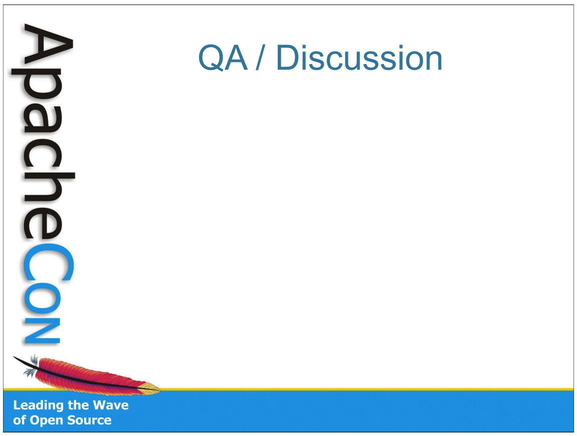
Conclusions

- 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

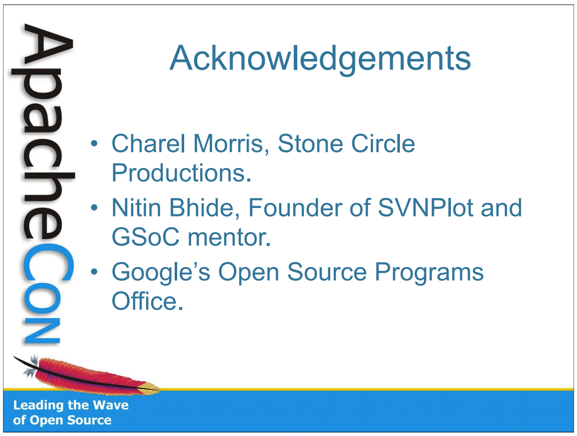

ApacheCon
Leading the Wave of Open Source

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

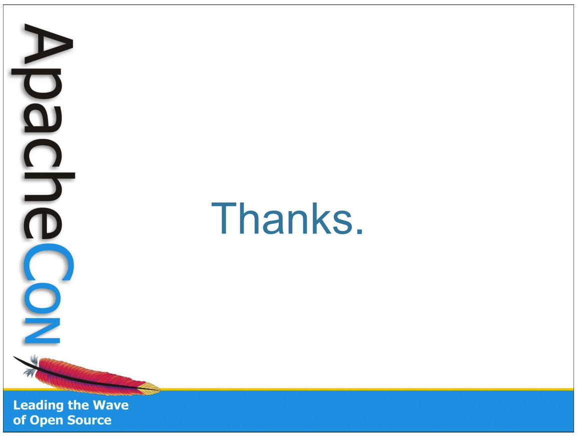

ApacheCon
Leading the Wave of Open Source

QA / Discussion


ApacheCon
Leading the Wave of Open Source

Acknowledgements

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


ApacheCon
Leading the Wave of Open Source

Thanks.