Learning and Researching with
Open Source Software

Minghui Zhou

zhmh@pku.edu.cn

Associate Professor

Peking University

IWSO 2012, Oct 15, Nanjing.

Outline

❑ **A snapshot of Open Source Software (OSS)**

❑ **Learning with OSS**

❑ **Research on OSS**

A SNAPSHOT OF OPEN SOURCE SOFTWARE

Is OPEN (Source Software) FREE?

What OSS are out there?

Who are working on OSS?

Is Open Source Software Free?

❑ **Student A: OSS is free**

❑ **Free:** 免费
❑ **Open:** 开放

Open != Free

What OSS are out there?

- ❑ **Microsoft Windows vs. Unix & Linux**
- ❑   **GNOME**
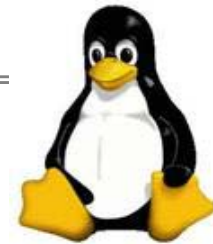- ❑ **Microsoft Office   vs. Open Office**

- ❑ **Microsoft IE   vs. Firefox**

- ❑ **DB2/SQLServer   vs. MySQL**

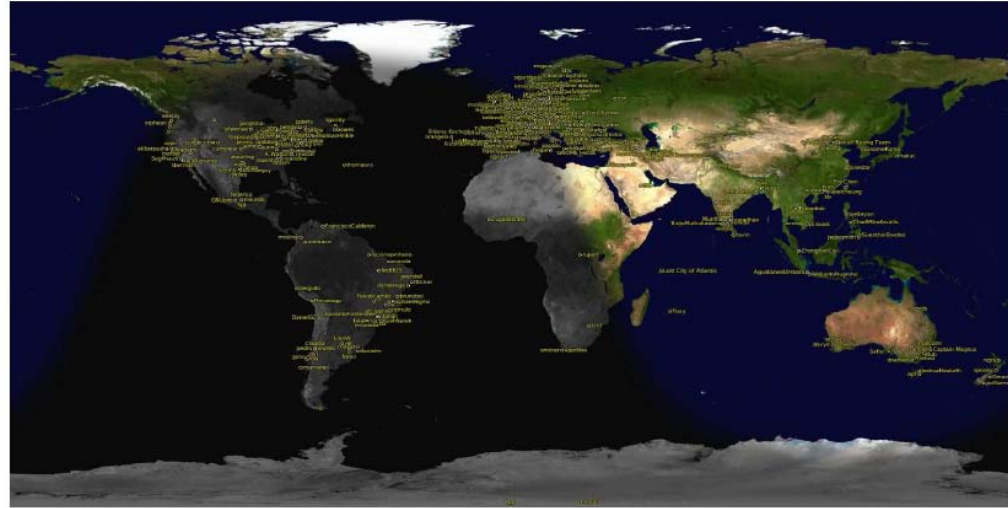- ❑ **IBM Websphere   vs. JOnAS (OW2)**

- ❑   **Apache Tomcat**

Who are working on OSS?
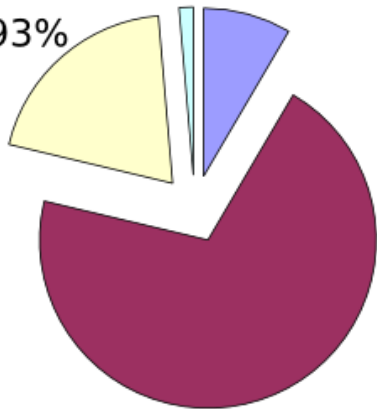
❑ **Volunteers**

❑ **Company backing**

How do you participate in the GNOME project?
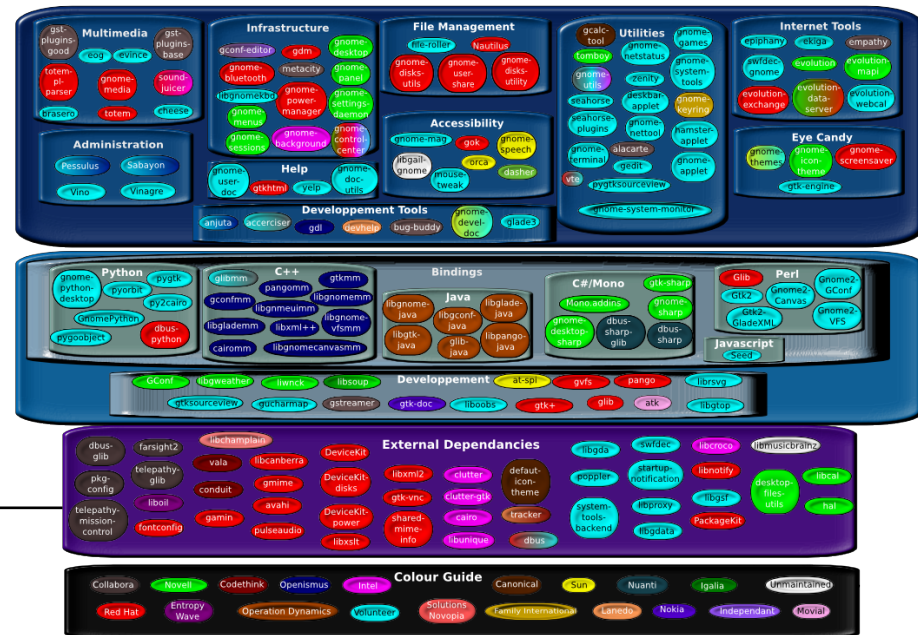
19.93%

- Professionally
- In my spare time
- Both
- No answer

70.27%

Few Chinese

LEARNING WITH OSS

Help with courses.

Learn through participation.

Help with Courses

❑ **There are good code out there,**

➢ **Helping to learn the principles and the implementation of an operating system, a web server, etc.**

❑ **There are free OSS out there,**

➢ **Helping to build all kinds of course infrastructures and applications**

Learning and motivated through participation

❑ " I learned a lot from this leading open source project (Gnome) while working with other contributes."

❑ "I volunteered because it was exciting to think--and as time progressed, know--that my involvement would have a direct effect on the quality and feature set of Mozilla"

❑ "A single ant cannot do much, but many ants can move mountains. That is the strength of Open Source. I want to help move a mountain!"

Trajectories of two volunteers

❑ **Volunteer A**
  ➢ **College student – report bugs from time to time**
  ➢ **Core member**
  ➢ **Now hired by Mozilla**

❑ **Volunteer B**
  ➢ **Director of music at St. Patrick Church, Washington, DC - commit fix from time to time**
  ➢ **Engineering manager at Mozilla**
  ➢ **Now employed by mozdev group, a consulting company specializing in mozilla-based technologies**
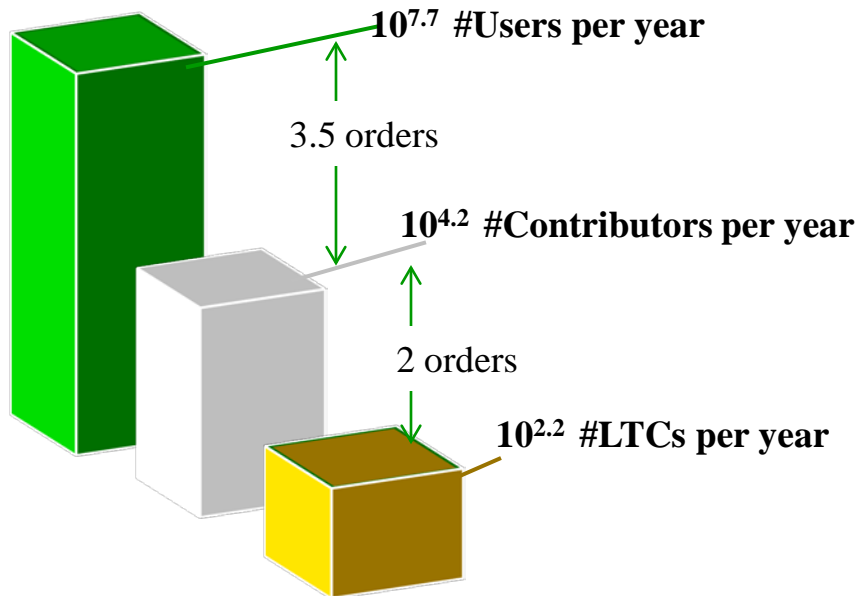
❑ **Volunteer C**

How to attract Long Term Contributors?

How commercial involvement affect OSS?

What are best practices in OSS?

RESEARCH ON OSS
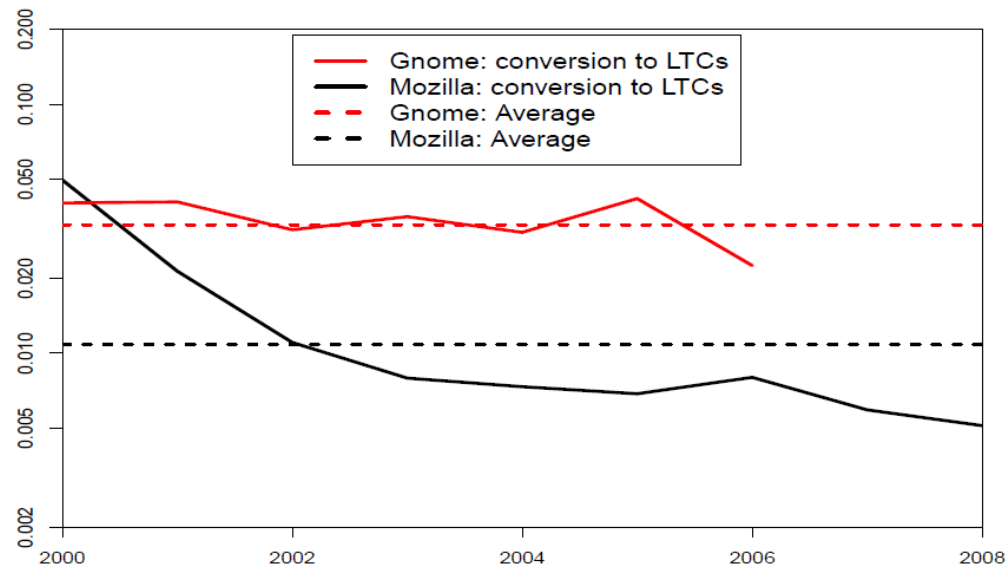
How to attract Long Term Contributors?

❑ **Different types of participators vary in orders of magnitude**

Mozilla (Average over 2000-2008)

$10^{7.7}$ **#Users per year**

3.5 orders

$10^{4.2}$ **#Contributors per year**

2 orders

$10^{2.2}$ **#LTCs per year**

每年用户：千万级别
每年新的参与人数：万
每年成为长期贡献者的人数: 百

Ratio of Contributors to LTCs in Gnome and Mozilla over years

Gnome: conversion to LTCs
Mozilla: conversion to LTCs
Gnome: Average
Mozilla: Average

Gnome和Mozilla中
贡献者与长期贡献者的比率逐年降低

Data Sources

❑ **Project Repositories**

Version Control
Version
Time
Login
File
MR

Issue Tracking
MR
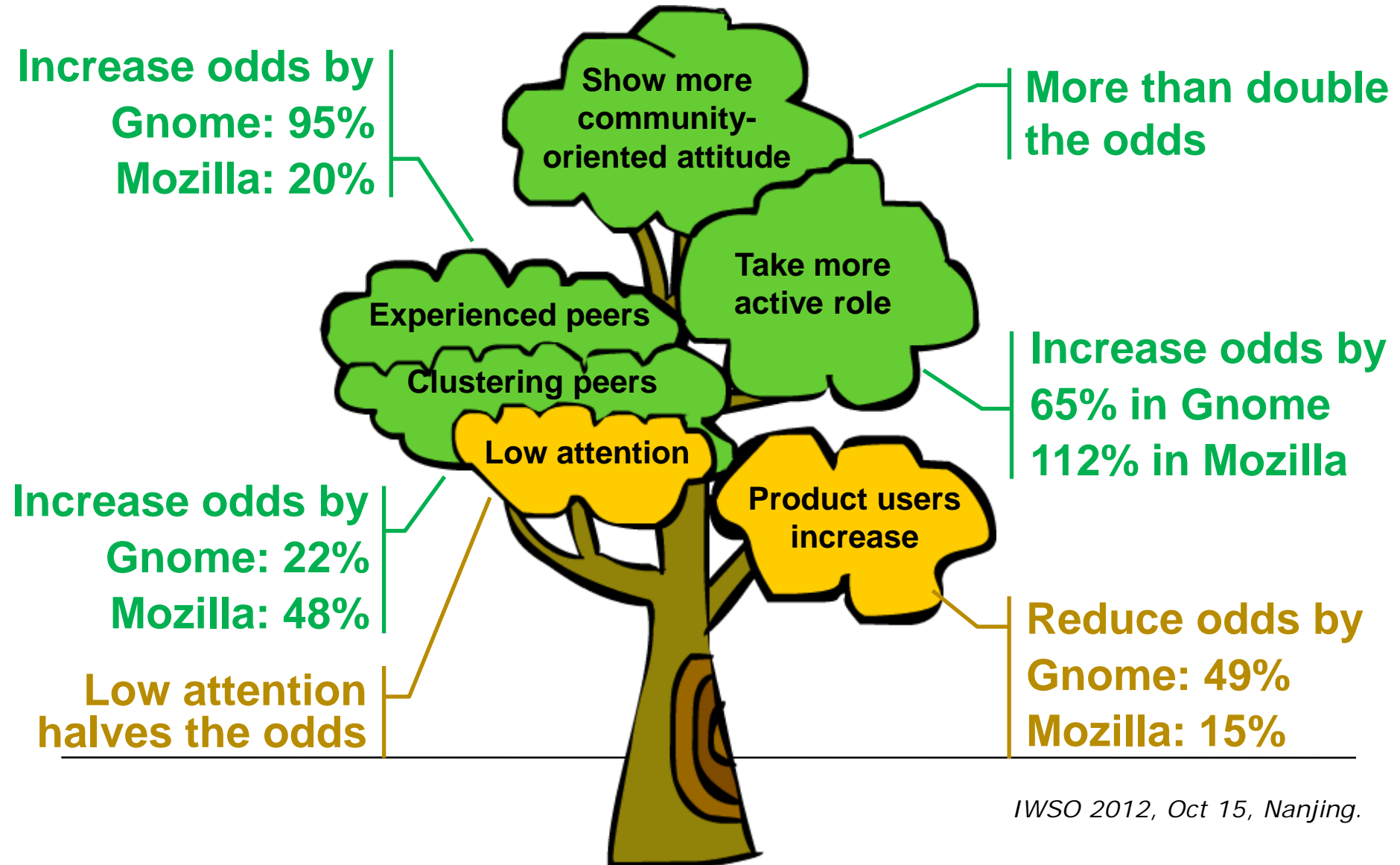Activity
-When
- Who
- What
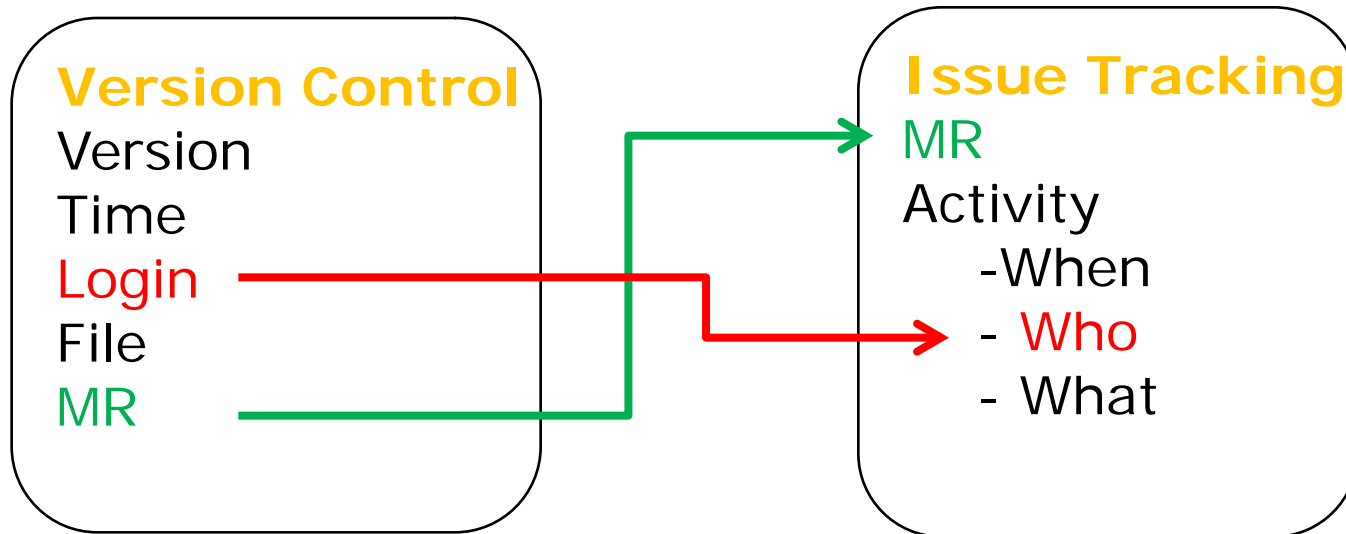
Gnome: 156,332 id,
517,801 bugs,
6,398,475 activities

Mozilla: 187,333id,
620,511bugs,
15,662,706activities

❑ **Questionnaires: Sent emails to developers**

What happens during the first month?

Increase odds by Gnome: 95% Mozilla: 20%

Show more community-oriented attitude

More than double the odds

Experienced peers

Take more active role

Clustering peers

Increase odds by 65% in Gnome 112% in Mozilla

Low attention

Increase odds by Gnome: 22% Mozilla: 48%

Product users increase

Low attention halves the odds

Reduce odds by Gnome: 49% Mozilla: 15%

IWSO 2012, Oct 15, Nanjing.

How commercial involvement affect OSS?

❑ **Different types of commercial involvement**

  ➢ **Hosting,**
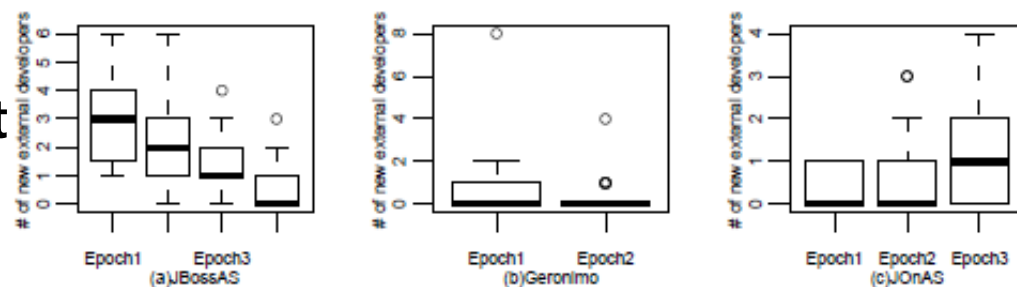 ➢ **Supporting,**
  ➢ **Collaborating**

Fig. 1.   Inflow of External Developers in JBossAS, Geronimo, and JOnAS

❑ **For example, Hosting mechanism**

  ➢ **Decrease the number of newcomers, but,**
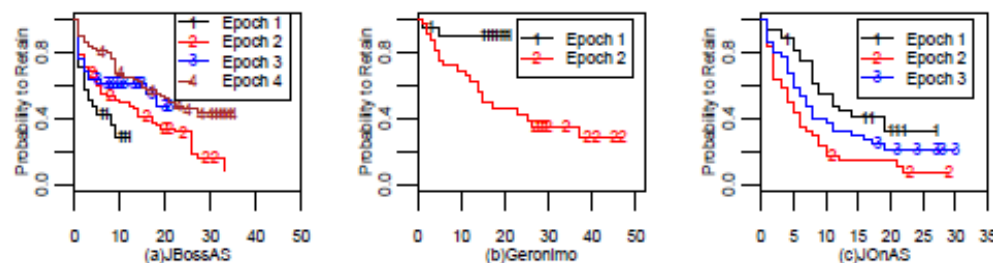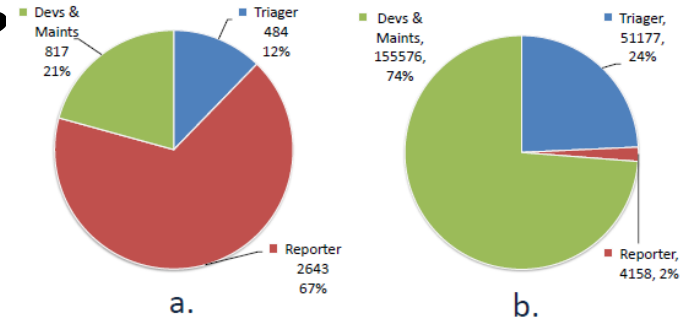 ➢ **Increase their retention**

Fig. 2.   Survival Curves of New Joiners in JBossAS, Geronimo, JOnAS each epoch

What are the best practices in OSS?

❑ **What is the best practice for triaging?**

  ➢ **Official website is not reliable**

  ➢ **Tradeoff between developer convenience (efficiency), user experience, and volunteer efficiency**

a. Number of roles of triage actors in Gnome
b. Number of issues each roles triaged in Gnome

❑ **How people reuse code?**

  ➢ **Which code are reused most often? What attributes do they have?**

  ➢ **Who wrote them? Their characters?**
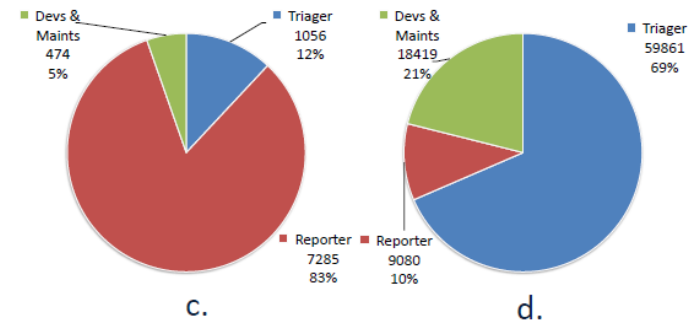
c. Number of roles of triage actors in Mozilla
d. Number of issues each roles triaged in Mozilla

A cloud of software repositories

❑ We keep tracking various commercial and open source projects.

❑ This "universal" repository records data from:
  ➢ Version control
  ➢ Issue tracking
  ➢ Email archives
  ➢ ......

| Forge	Type	Files	File/Ver.	Unique File/Ver.	Branching	From
Large cmpny.	Var.	3,272K	12,585K	4,293K	2.9	1988
SourceForge	CVS	**26,095K**	81,239K	**39,550K**	2.1	1998
code.google	SVN	5,675K	14,368K	8,584K	1.7	1996
repo.or.cz	Git	2,519K	11,068K	5,115K	2.2	1986
Savannah	CVS	852K	3,623K	2,345K	1.5	**1985**
git.kernel.org	Git	12,974K	**97,585K**	856K	**114**	1988
OpenSolaris	Hg	77K	1,108K	91K	12.2	2003
FreeBSD	CVS	196K	360K	75K	4.8	1993
Kde	SVN	2,645K	10,162K	527K	19.3	1997
gnome.org	SVN	1,284K	3,981K	1,412K	2.8	1997
Gcc	SVN	3,758K	4,803K	395K	12.2	1989
Eclipse	CVS	729K	2,127K	575K	3.7	2001
OpenJDK	Hg	32K	747K	60K	12.4	2008

Open Data -- Internetware

❑ Machines
- ➤ DELL R910(4U), 64GbRAM, 16-cores X7550
- ➤ DELL MD3200, 12*2TB SAS
- ➤ DELL R710 * 4, 64GbRAM

❑ Data Levels
- ➤ Level0: raw data
- ➤ Level1: filtered data
- ➤ Level2-n: standardized data

Passion-Lab
Software Enginnering Institute
Peking University

Data Center
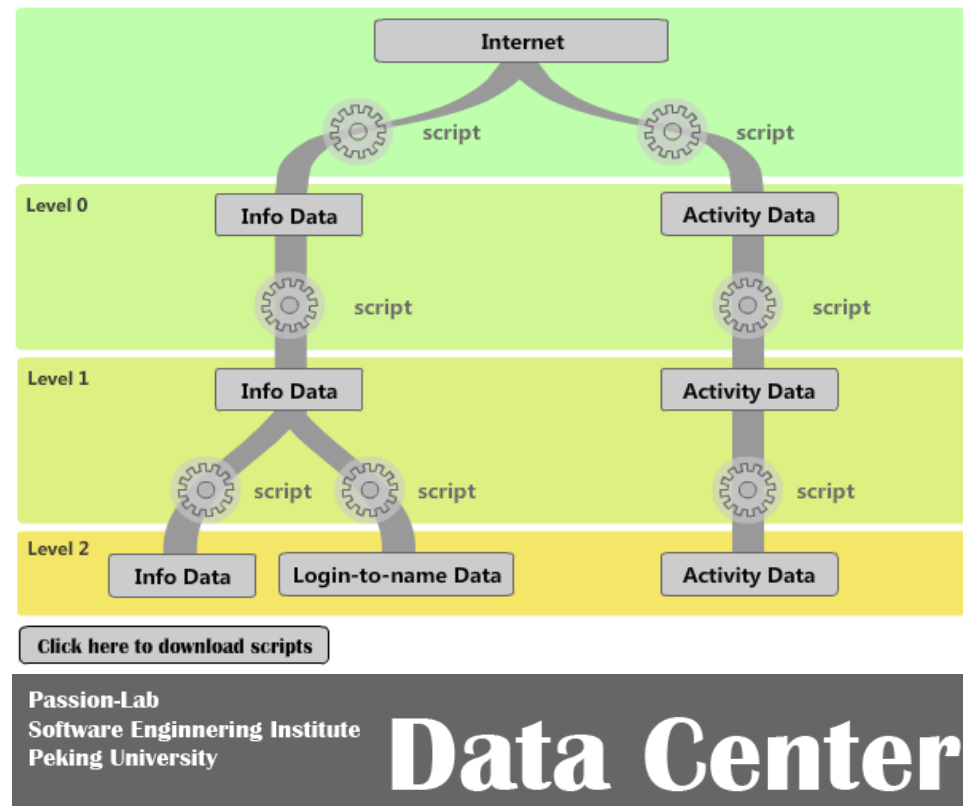
HTTP://passion-lab.org

What we could do with this data?

❑ To enable better user experience, and software engineering practices in a large-scale

➢ Understand the past, predict the future

Passion-Lab
Software Enginnering Institute
Peking University
Data Center

HTTP://passion-lab.org

References

❑ B. Boehm. Software Engineering Economics. Prentice-Hall, 1981.

❑ B. Curtis. Fifteen years of psychology in software engineering: Individual differences & cognitive science. In ICSE'84, pp 97–106, 1984.

❑ A. Mockus, R. T. Fielding, and J. Herbsleb. A Case Study of Open Source Development: The Apache Server, 22nd ICSE, pp. 263-272, Limerick, Ireland, June 4-1, 2000.

❑ A. Mockus. Software support tools and experimental work. In V. Basili and et al, editors, Empirical Software Engineering Issues: Critical Assessments and Future Directions, volume LNCS 4336, pages 91–99. Springer, 2007.

❑ A. Mockus. Amassing and indexing a large sample of version control systems: towards the census of public source code history. In 6th IEEE Working Conference on Mining Software Repositories, May 16–17 2009.

❑ http://blogs.gnome.org/bolsh/2010/07/28/gnome-census/

❑ M. Zhou and A. Mockus. Growth of Newcomer Competence: Challenges of Globalization. In FoSER(future of software engineering) on ACMSIGSOFT / FSE, Santa Fe, New Mexico, Nov 7-8, 2010, pp443-448.

References

- M. Zhou and A. Mockus. Does the initial environment impact the future of developers? ICSE 2011, Honolulu, Havaii, May 21-28, 2011, pp71-80.
- M. Zhou and A. Mockus. What Make Long Term Contributors: Willingness and Opportunity in OSS Community. ICSE 2012, Zurich, Switzerland, 2-9 June 2012, pp.518-528.
- Xiujuan Ma, Minghui Zhou, Hong Mei. A case study of internetware development. Proceedings of the Second Asia-Pacific Symposium on Internetware, 2010.
- X Ma, M Zhou, H Mei. How developers participate in open source projects: a replicate case study on JBossAS, JOnAS and Apache Geronimo. Proceedings of the 1st International Workshop on Replication in Empirical Software Engineering Research, May, 2010
- J. Xie, M. Zhou and A. Mockus. Visualizing the evolution of software issue-tracking practices. 5th International Symposium on Empirical Software Engineering and Measurement (ESEM 2011).
- M. Zhou. 2012. Looking for micro-process in large-scale data. In Proceedings of the 2nd international workshop on Evidential assessment of software technologies (EAST '12). ACM, New York, NY, USA, 39-42.

THANK YOU