Inflow and Retention in OSS Communities with Commercial Involvement

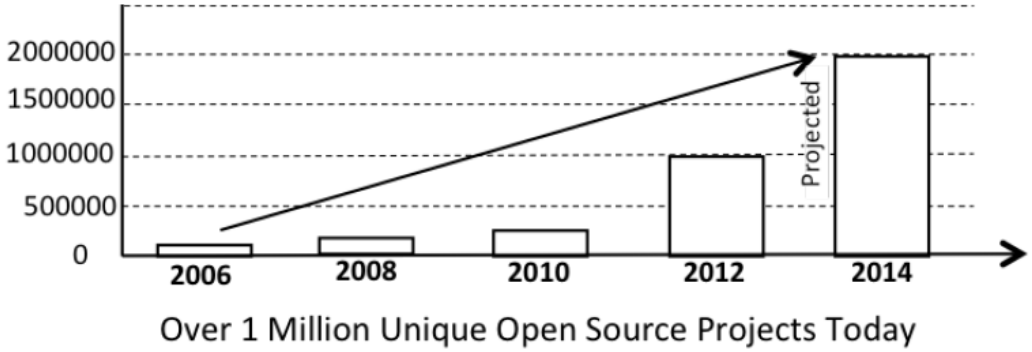
Minghui Zhou
zhmh@pku.edu.cn
http://sei.pku.edu.cn/~zhmh
Peking University

ISHCS, Beijing, 2016-12-18

Open source is eating the world

80% of all Software Development is Open Source

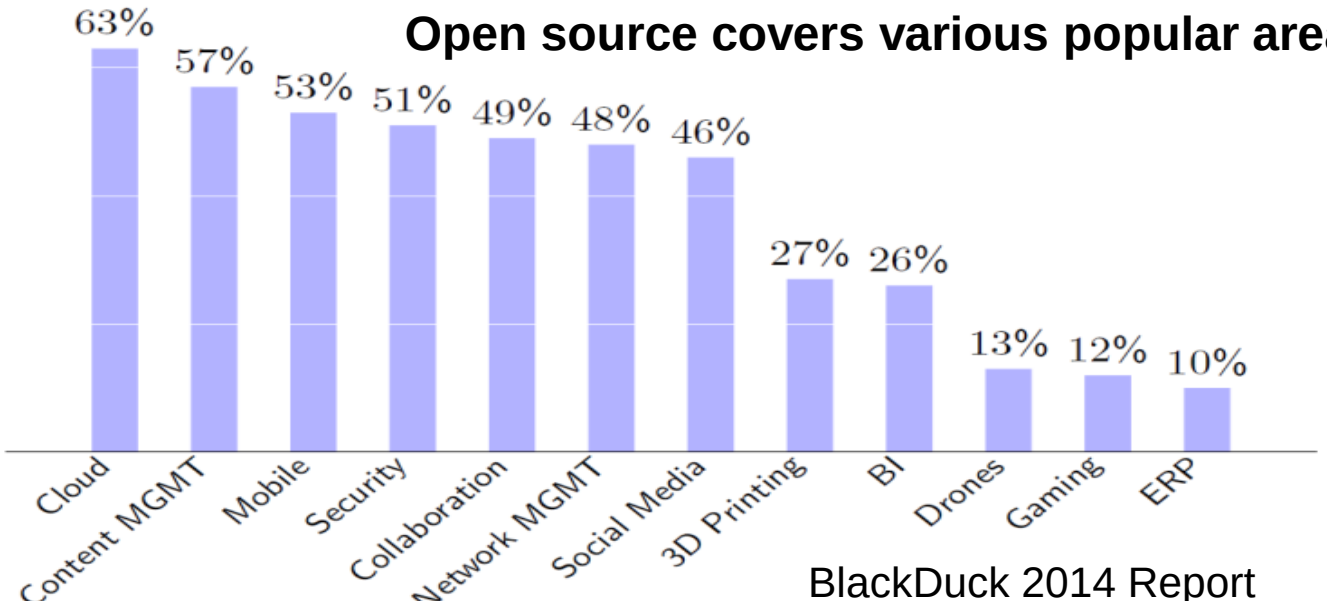
-- Jim Zemlin, Linux Con 2014



Over 1 Million Unique Open Source Projects Today

The growth is exponential

Open source covers various popular areas



BlackDuck 2014 Report

Companies participate in OSS

- Every successful OSS has firms involved.
  - Linux kernel: e.g., Intel contributed 10.5% of changes, RedHat 8.4% (2015)

Company	Changes	Total
None	11,968	12.4%
Intel	10,108	10.5%
Red Hat	8,078	8.4%
Linaro	5,415	5.6%
Samsung	4,290	4.4%
Unknown	3,842	4.0%
IBM	3,081	3.2%
SUSE	2,890	3.0%
Consultants	2,451	2.5%
Texas Instruments	2,269	2.4%
Vision Engraving Systems	2,089	2.2%
Google	2,048	2.1%
Renesas Electronics	2,004	2.1%
Freescale	1,690	1.8%
Free Electrons	1,463	1.5%
FOSS Outreach Program for Women	1,418	1.5%
Oracle	1,166	1.2%
AMD	1,109	1.1%
NVidia	1,078	1.1%

- Projects benefit from commercial resources, and prosper at the market, sometimes.

Downsides

- Volunteers may feel unappreciated and their work being appropriated.
  - A volunteer from JBossAS (RedHat): “It’s tough to build a real community when you have paid committers and unpaid contributors developing code under (L)GPL with the original copyright assigned to the company that funds the effort.”
- Volunteers may stop contributing.
  - Linux kernel: the number of volunteers contributing to the Linux kernel has been slowly declining for many years, now sitting at 12.4% (13.6% in 2014, and 14.6% in 2013).

Aim

- **Is there a good way for a company to participate in open source community?**
  - Definition: HYBRID
    - OSS project with commercial participation
- **Impact on the community**
  - Newcomers: may bring changes and new values, and may add important features and later join the core team.
  - Retention: projects that are able to retain developers for longer periods increase the pool of expertise; losing an existing developer is detrimental to the project.

Research Questions

- **We select three hybrids to investigate:**
  - **What policies and actions did companies employ to get involved in communities?**
  - **What is the impact of the policies and actions?**
    - **Did they increase the inflow of new contributors?**
    - **Did they improve retention of existing contributors?**

JBossAS : since 1999

Geronimo : since 1999

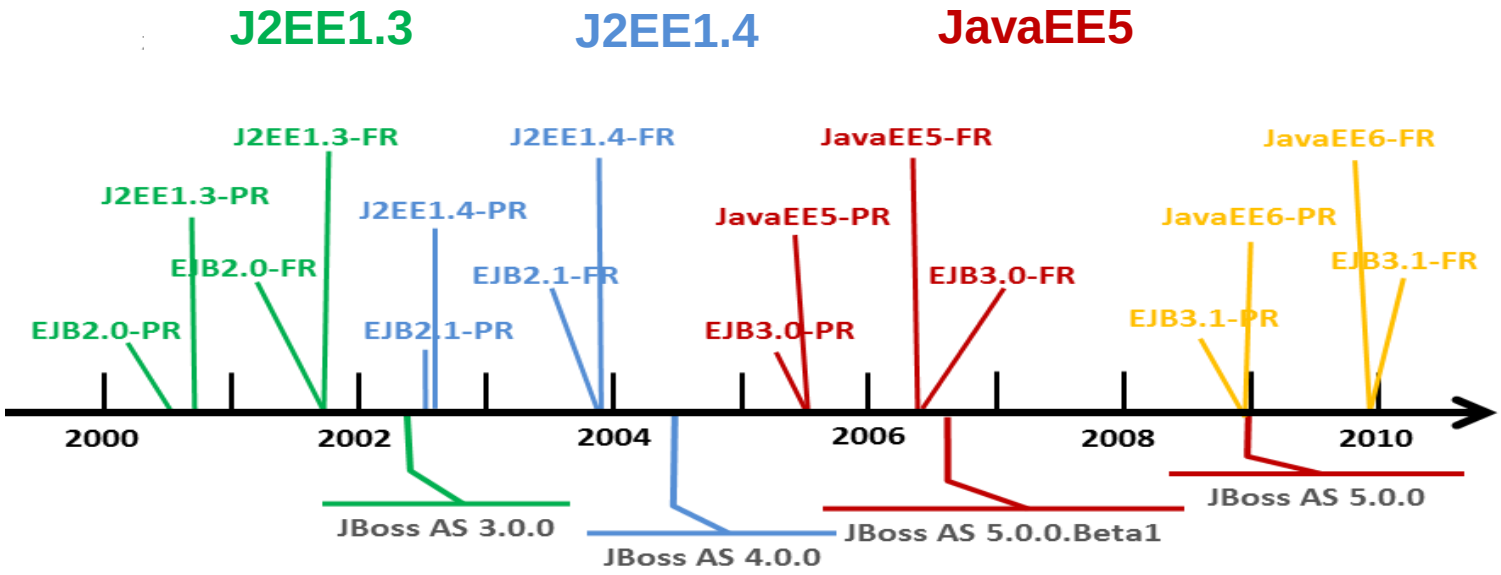
JOnAS : since 2003

The Apache Software Foundation

The open source community for infrastructure software

Why these three JavaEE projects?

- **To control for variation inherent in software, we select three projects that develop the same specification in the similar time period.**
  - Control for external factors (economy, technology)
  - Control for context
  - We were a major contributor to one of the projects



Mixed-Method Study

- Main STEP1:
  - Use literature to derive dimensions of commercial involvement.
- Main STEP2:
  - Use online documents to obtain commercial involvement models.
- Main STEP3:
  - Analyze project repository data, and quantify the impact of commercial involvement.
- Validation:
  - Conduct interviews and surveys to validate the findings.

Results

- **STEP1: Use literature to derive dimensions of commercial involvement.**
- STEP2: Use online documents to obtain commercial involvement models.
- STEP3: Quantify the impact of different commercial involvement models.

Two Dimensions of Commercial Involvement

Motivation:

Q1: commercial objectives?

Actions:

Q2: manage IP?

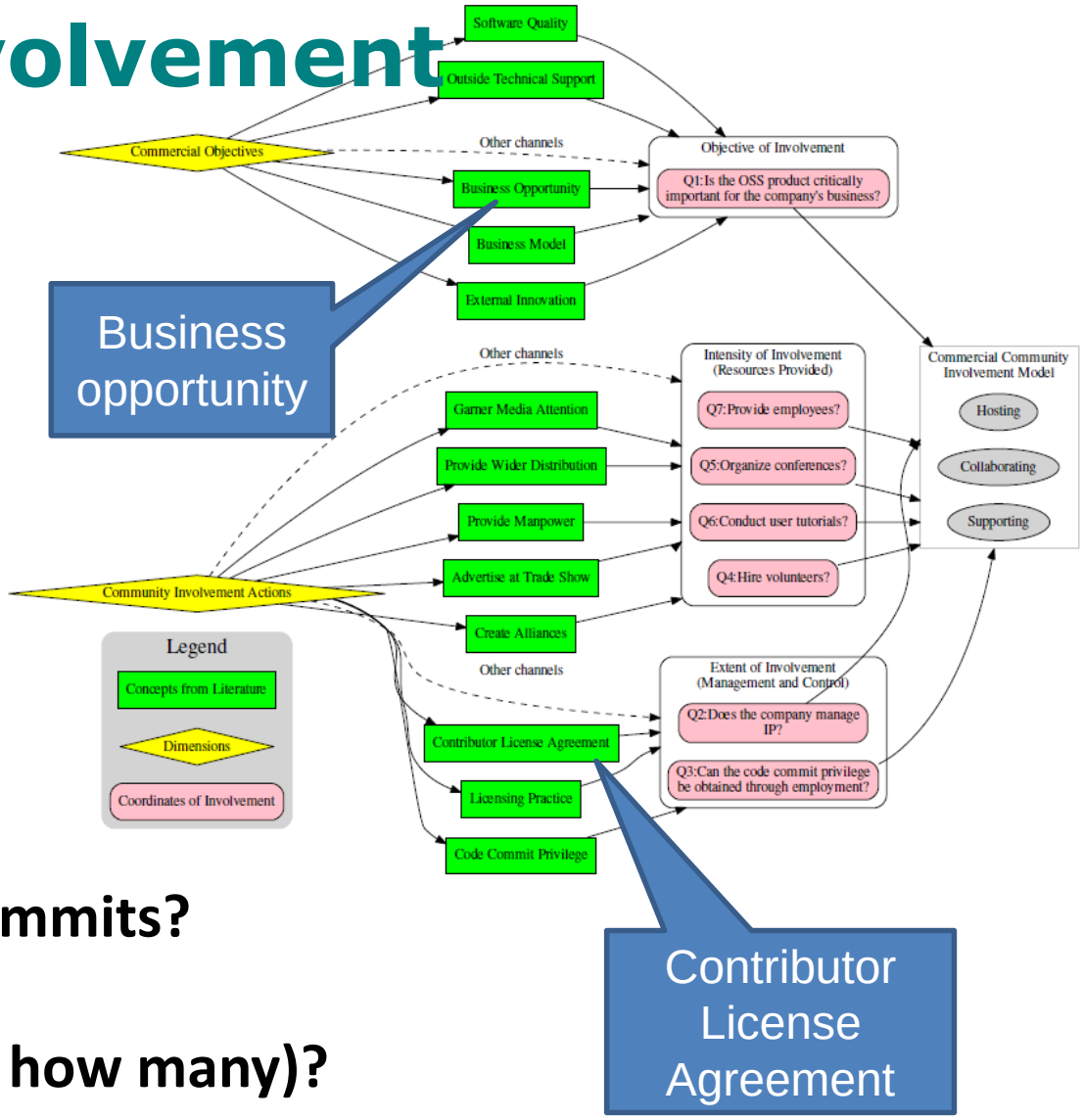
Q3: code commit privilege?

Q4: hire project volunteers?

Q5: organize conferences/summits?

Q6: conduct user tutorials?

Q7: provide employees (and how many)?



Results

- STEP1: Use literature to derive dimensions of commercial involvement.
- **STEP2: Use online documents to obtain commercial involvement models.**
- STEP3: Quantify the impact of different commercial involvement models.

Q1: commercial objectives?

Bull on JOnAS: to gain profit indirectly through other commercial products.

IBM on Geronimo: to gain profit indirectly through the commercial products.

RedHat on JBossAS: to gain profit through subscription-based services.

Bull (JO)	General interest: To provide open source enterprise …eware solutions and support for the customers, and to adver…e the open source strategy for OW2. To gain profit indirectly through other commercial products.
Bull (JO.E1)	Specific Interest(SI) in E1: To develop its early version, through cooperation with several French organizations
Bull (JO.E2)	SI in E2: To impleme… and to be effective at… velopers start building… own products.
Bull (JO.E3)	SI in E3: To implement JavaEE 5 and to develop …tive version conforming to new specifications based on OS… to cooperate… …den its …ce. …ption to …f services and tools.
RHT (JB.E4)	To ensure the continu… an independent op… ource application server and integra… as a division of RedH… To gain profit through subscription-based services.
IBM (GE.E2)	To support Geronimo and distribute a commercial …ommunity version based on it, and to obtain and test innovations transferable to the commercial products, e.g WebSphere. To gain profit indirectly through the commercial products.

Q2-7: community involvement actions?

Extent of Community involvement | **Intensity of Community involvement**

Epoch	Management and Control		Resources Provided by Companies				
	IP Management Entity	Way to Be Committer	R1-R4*	Employees	Hire	Conferences	Tutorials
JB.E1	JBoss Org	Contribute	—	—	—	—	—
JB.E2	JBoss Group LLC.	Be employed+	Y	Y(15)	Y	Y	N
JB.E3	JBoss Inc.	Be employed+	Y	Y(30)	Y	Y	Y
JB.E4	RedHat	Be employed+	Y	Y(150)	Y	Y	Y
GE.E1	ASF	Contribute	—	—	—	—	—
GE.E2	ASF	Contribute	Y	Y(35)	Y	—	—
JO.E1	Bull, France Telecom, Lifl, INRIA	Contribute**	Y	Y(12)	—	—	—
JO.E2	the same as above	Contribute**	Y	Y(13)	—	—	—
JO.E3	Adding: SerLi, U of Fortaleza, Peking U.	Contribute**	Y	Y(15)	—	—	Y

Q2 Q3 Q4 Q5 Q6 Q7

Three commercial models

- Hosting: RedHat on JBossAS

  - Clear and strong commercial objectives
  - extensive resources, full control
- Supporting: IBM on Geronimo

  - Not directly related to revenue
  - Partial Resources, support project (control is on 3rd party)
- Collaborating: BULL on JOnAS
  - Not directly related to revenue
  - Partial resources, shared control

Results

- STEP1: Use literature to derive dimensions of commercial involvement.
- STEP2: Use online documents to obtain commercial involvement models.
- **STEP3: Quantify the impact of different commercial involvement models.**

Epochs and Contrasts

- Each project underwent changes in commercial involvement, epoch represents constant
  - Compare the **similar epochs** in **different projects**
  - Compare **different epochs** within the **same project**

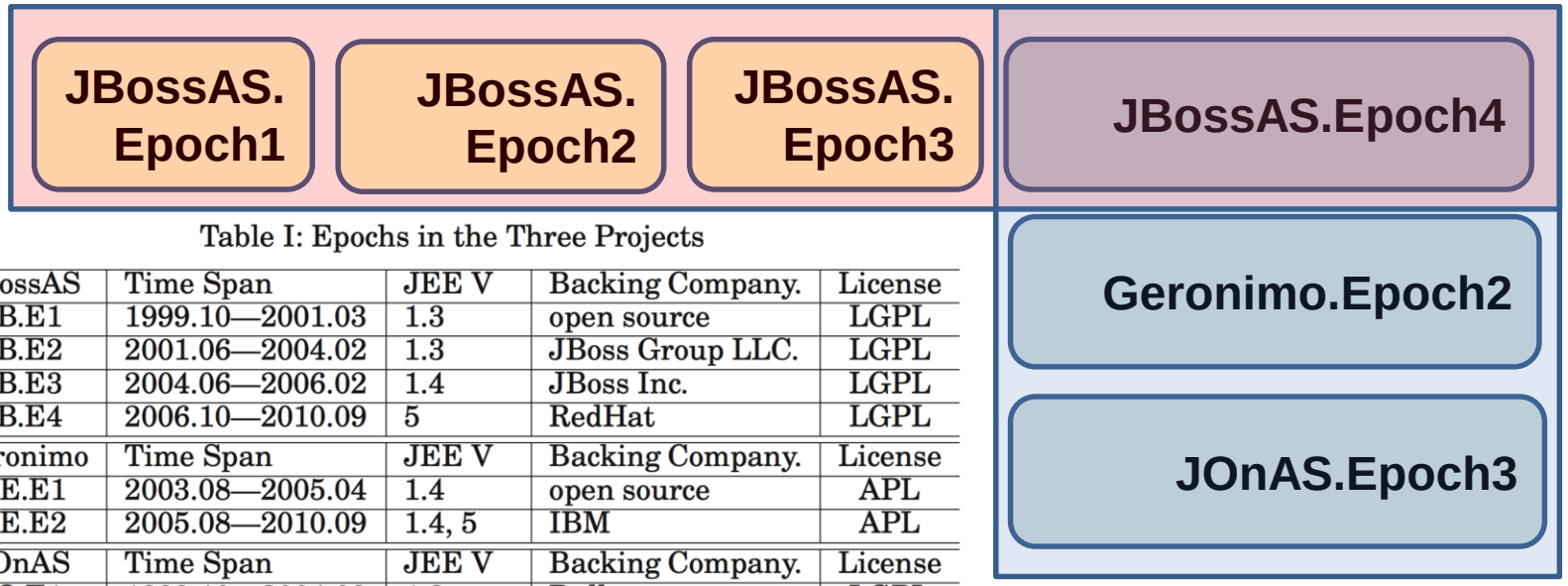


Table I: Epochs in the Three Projects

JBossAS	Time Span	JEE V	Backing Company.	License
JB.E1	1999.10—2001.03	1.3	open source	LGPL
JB.E2	2001.06—2004.02	1.3	JBoss Group LLC.	LGPL
JB.E3	2004.06—2006.02	1.4	JBoss Inc.	LGPL
JB.E4	2006.10—2010.09	5	RedHat	LGPL
Geronimo	Time Span	JEE V	Backing Company.	License
GE.E1	2003.08—2005.04	1.4	open source	APL
GE.E2	2005.08—2010.09	1.4, 5	IBM	APL
JOnAS	Time Span	JEE V	Backing Company.	License
JO.E1	1999.10—2004.03	1.3	Bull	LGPL
JO.E2	2004.06—2006.10	1.4	Bull	LGPL
JO.E3	2007.02—2010.09	5	Bull	LGPL

How different models affect number of newcomers?

JBossAS	Gernonimo	JOnAS
decrease	decrease	increase

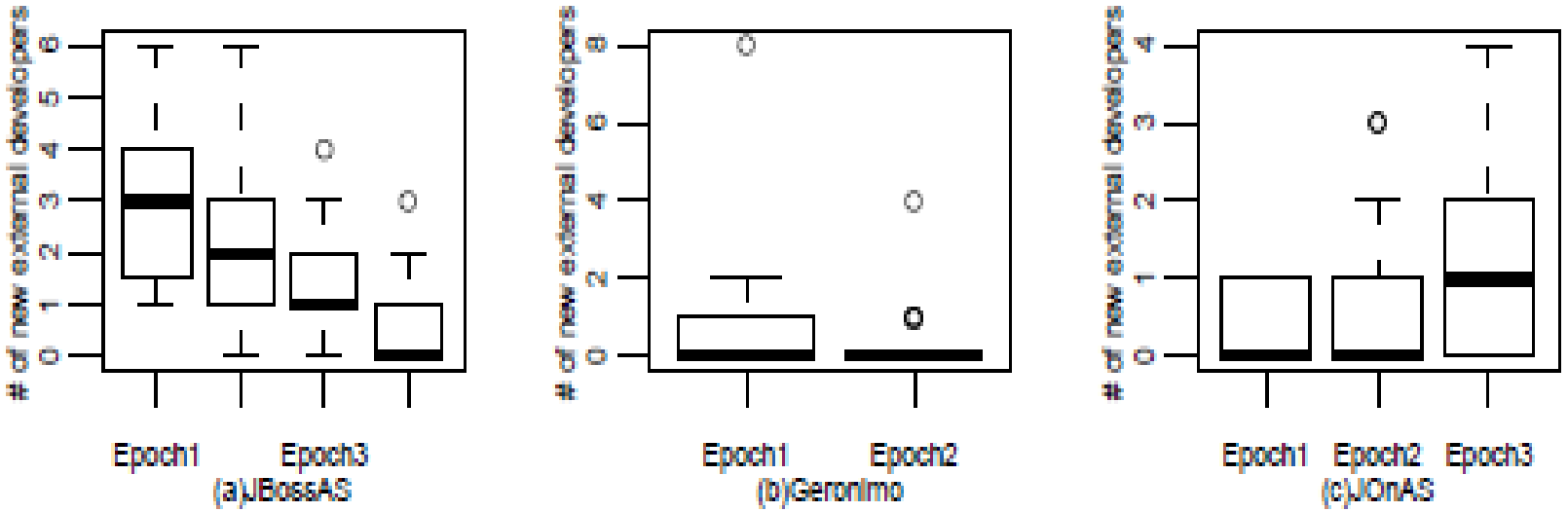


Fig. 1. Inflow of External Developers in JBossAS, Geronimo, and JOnAS

How different models affect contributor retention?

JBossAS	Gernonimo	JOnAS
increase	decrease	decrease

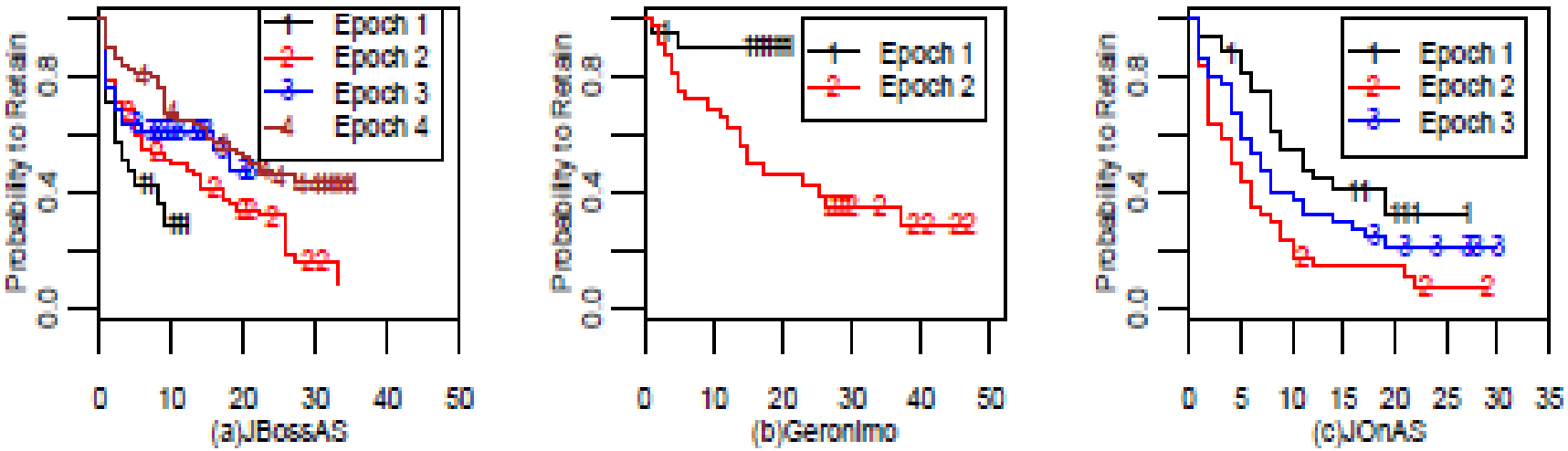


Fig. 2. Survival Curves of New Joiners in JBossAS, Geronimo, JOnAS each epoch

Validation

- Validation study
  - Round 1: used mailing-lists and face-to-face interviews => obtained 17 email responses and conducted 11 interviews
  - Round 2: conducted a survey through private emails, selected two developers from each project =>
    four developers from Geronimo and JOnAS responded
- Collected alternative data to verify that
  - We identified the external dvprs correctly,
  - The homogeneous periods and the related commercial practices were defined appropriately
  - The involvement actions affected contributor inflow and retention.

Recommendations and Implications

Recommendations for practice

- Full control mechanisms and high intensity of commercial involvement were associated with a decrease of external inflow and with improved retention.
- A shared control mechanism was associated with increased external inflow contemporaneously with the increase of commercial involvement.

Implications for research

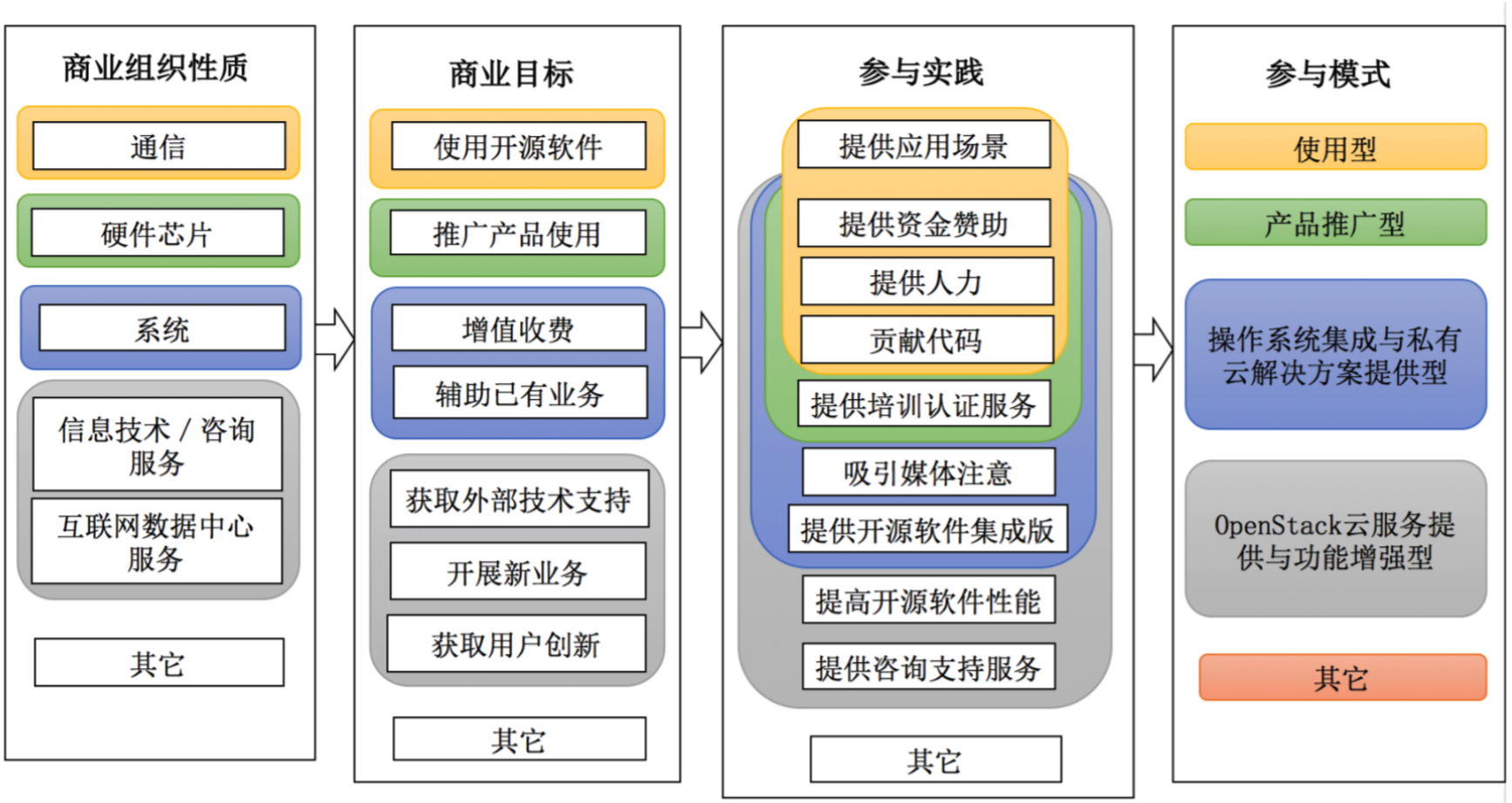
- Natural Experiment could help to control for nuisance factors.
- Two-dimension (seven-question) framework of hybrid space can be used for classifying and structuring commercial involvement.

TOSEM May 2016.
download: dl.acm.org/citation.cfm?id=2876443

Evolution

- Commercial involvement is evolving
  - Commercial objectives drive involvement
  - New technology is an important motivator, e.g.:
    - JaveEE is currently an established market, commercial interest has waned, and the OSS project may lose support and fail.
    - A new generation of technology and contributors may arise and produce new projects and new communities from old ones, e.g., TomEE is taking dvprs from Geronimo.
- We can categorize hybrids: Linux kernel, OpenStack, Android, Docker, ……

OpenStack商业参与模式



- 模式 1:以 AT&T 为代表的使用型
- 模式 2: 以 Intel 为代表的产品推广型
- 模式 3:以 RedHat、Canonical 和 SUSE 为代表的OS集成和私有云解决方案 供型
- 模式 4:以 IBM、HPE、和 Rackspace 为代表的云服务提供与功能增强型

Next: Linux Kernel

- The Linux kernel is the result of one of largest cooperative software projects ever attempted
  - 1,300 companies, 14,000 individuals ever contributed code to Linux kernel
  - Each release contains work of > 1600 dvprs representing > 200 companies, now > 20 million LOC
- How does that work out?
  - How do different stakeholders work together to form a community that develop and maintain such a huge software,
  - and how does an ecosystem with millions of repositories and developers operate given the lack of centralized planning?

THANKS