Communication Structure and Performance of Open Source Software Projects

April 29, 2006

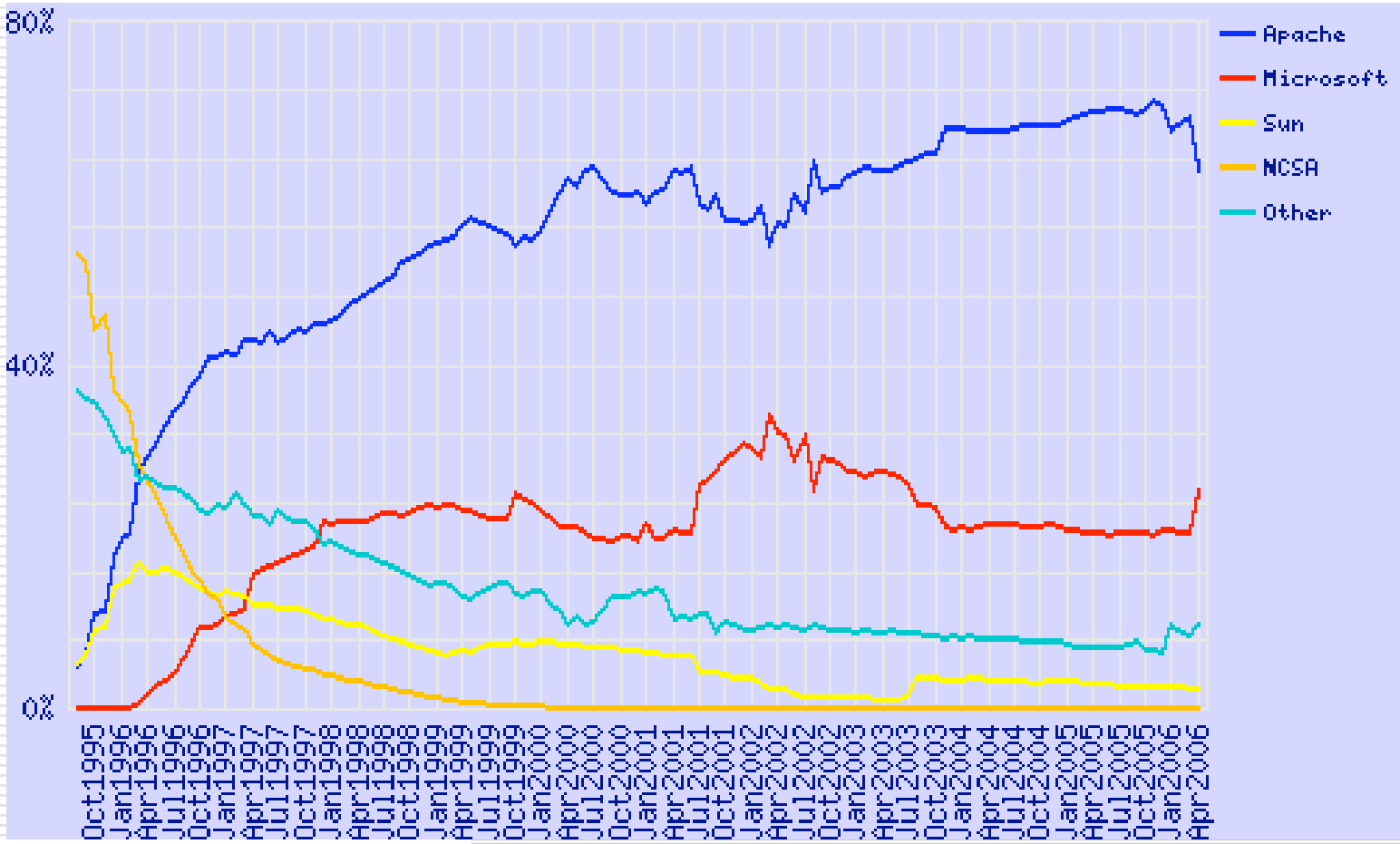
Yutaka Hamaoka
Keio University
Hamaoka@fbc.keio.ac.jp

Agenda

- Motivation
- Previous Research
- Hypotheses
- Data
- Variables
- Analysis
- Result
- Summary and conclusion
- **Limitation and Future Research**

Motivation

- Market share of Server software



Source)netcraft.com

Open Source Software Project as a user-centered innovation System

Manufacturer centered Innovation system

- Manufacturer
  - Develop
  - Marketing
  - User Support
  - Maintain

- User(consumer)
  - Buy
  - Use
  - Post purchase behavior
    - WOM,Complain
    - Repeat purchase

User centered Innovation system

User

developer

- Development
- Marketing
- User Support
- Maintain

user

- Download
- Use
- Post use behavior
  - Feedback
  - Participate

Research on Open source software[-2004]

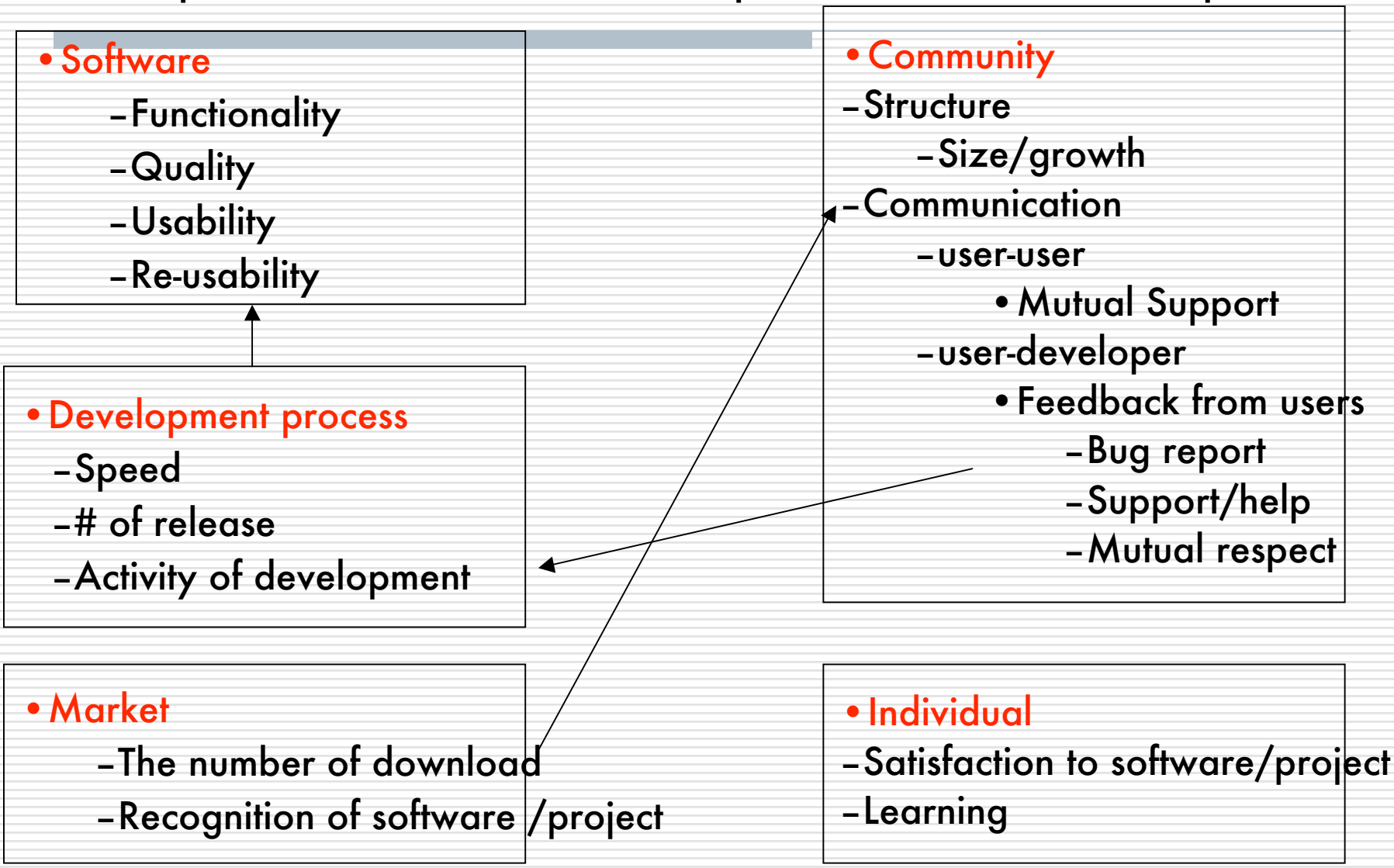
- Case study on Development process/team
  - Apache [Mockus et.al.(2000)]
  - Gnome [Koch and Schneider(2000)]
  - Linux kernel [Tuomi(2000)]
  - Freenet [von Krough et.al.(2003)]
- Survey on Motivation of individuals
  - Linux Developer[Hertel and Herrmann(2003)]
  - Apache help-line[Lakhani and von Hippel (2003)]

- Limitation
  - Focusing *single significantly succeeded* Open source software project
    - No comparison.
    - No quantitative data.
    - *"Why the project succeeded?"* is unexplained.

Hamaoka(2004)

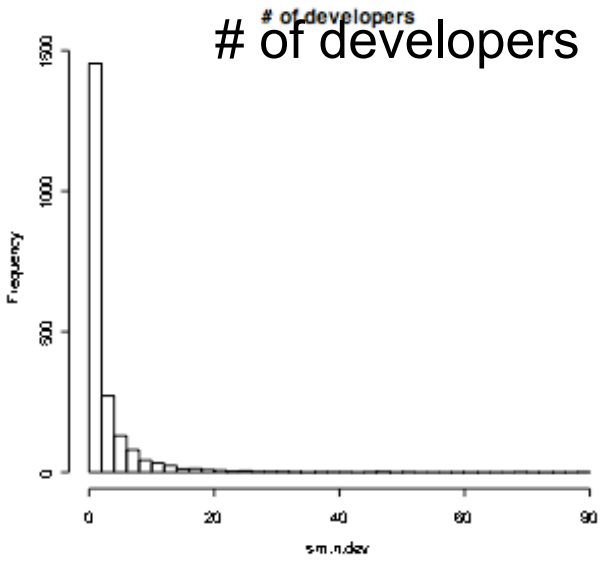
RQ1 What is *success* of Open Source Software Projects?

Proposed "Success metrics" of Open Source Software Projects

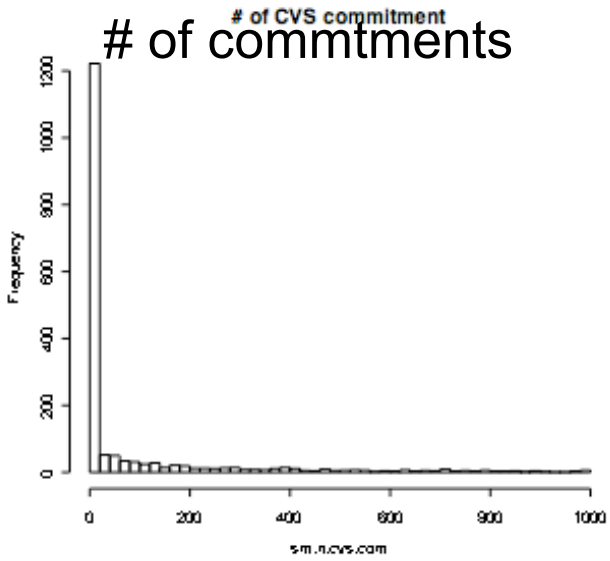


- RQ2 Are Open Source Software Projects really successful?
  - *Data*
    - *Sourceforge.net*
  - Unit of Analysis
    - Project
  - Sampling
    - 2,200 projects were randomly selected
      - 10% of 23,000 projects as of May,2001
    - Top 100 projects in terms of page view, download, activity were added to cover really successful projects.
    - Duplicated projects were removed.
    - 2,101 projects

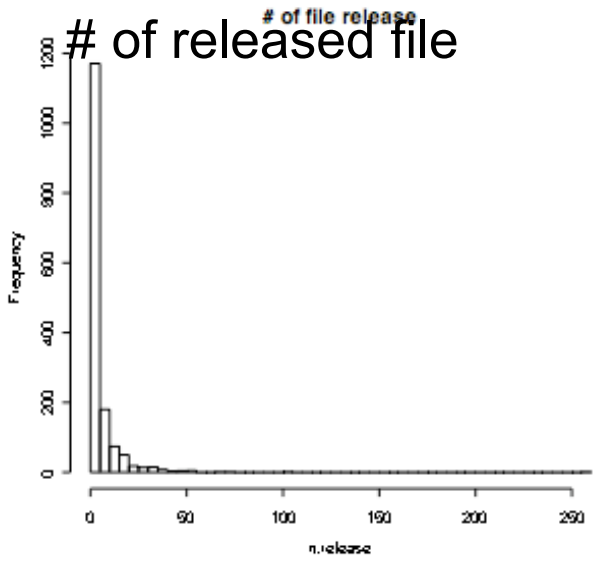
Distribution of success metrics



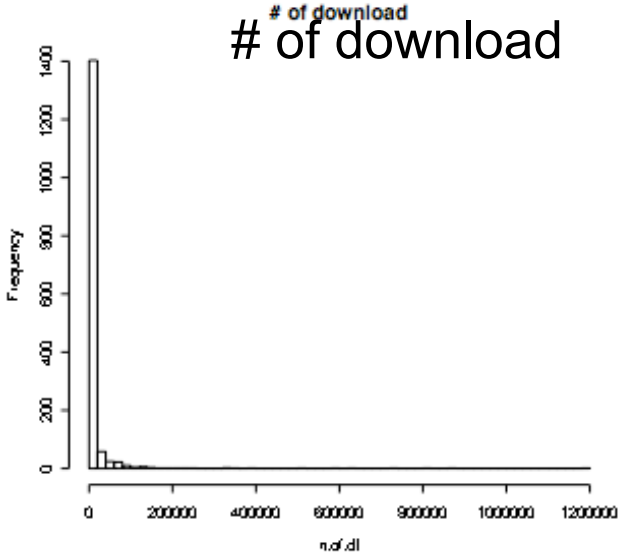
# of developers



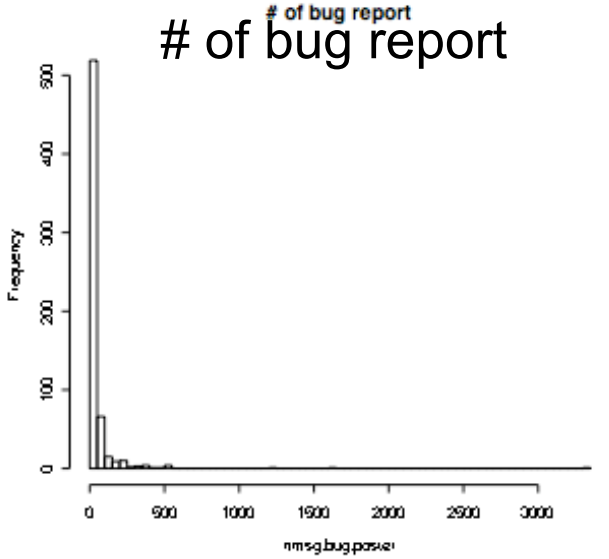
# of commtments



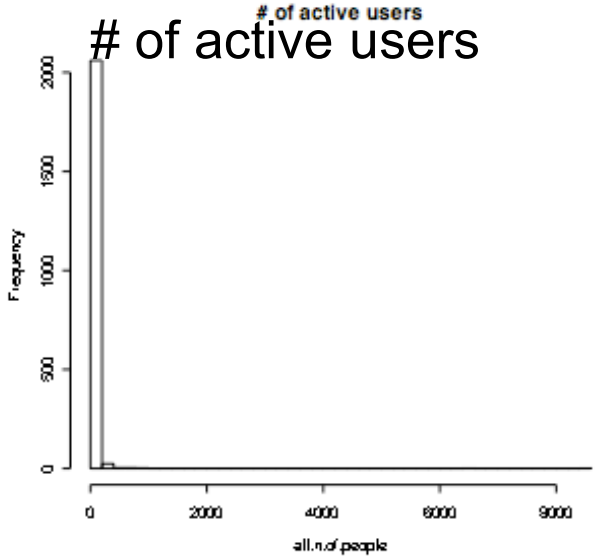
# of released file



# of download

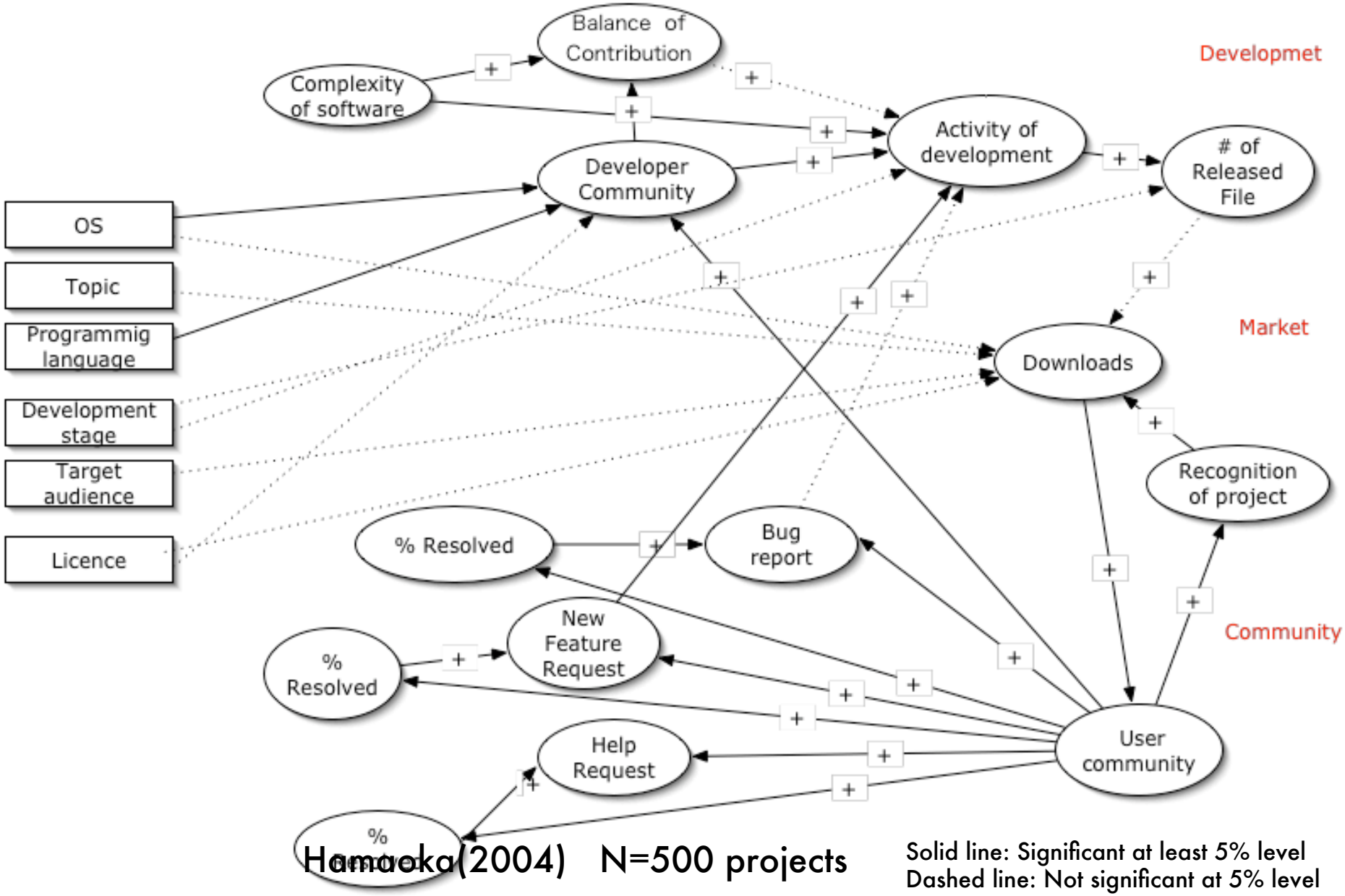


# of bug report



# of active users

RQ3 What makes Open Source Software Project more successful?



Hamaoka(2004) N=500 projects

Solid line: Significant at least 5% level
Dashed line: Not significant at 5% level

Research Question of Present study

- Does communication structure among developers, users, and user-developers affect *development* performance?

Previous Researches on Network structure and Group Performance

	Unit of analysis		Explanatory Variables		Dependent Variables	
	G: Group, I: individual		SNA variables	other	Performance	Creativity
Provan and Sebastian(1998)	G	3 Mental health agencies	Clieaque overlap(+), Service link Multiplexity(+)*		Client satisfaction, QOL	
Ahuja, Manju and Kathleen Carley(1999)	G	3 task groups at SOAR project	Hierarchy, Centrality, Hierarchical levels	Task type	Objective and percieved performance	
Sparrowe, et al.(2001)	I	190 employees in 38 work groups	In-degree centrality [advice(+/+), hindrance(+/-)]		In-role and extra-role performance	
	G		Density[advice(ns), hindrance(-)], Advice network centralization(-)		Assessment by leader	
Cummings and Cross(2003)	G	182 Workgroups at Fortune 500 telecommunications firm	Hierarchy(-), Core-periphery (-), Structural holes of the leader (-)		Manager rated performance	
Kidane and Gloor(2005)	G	33 Open source software projects:Eclipse	Group betweenness centrality(ns/ns), Group density(+/ns)		% of bugs resolved	# of enhancement
Reagans and	I	104 individuals in R&D	Social cohesion(+)	Common	Knowledge transfer	

Hypotheses

- Developer
  - Team size
    - Hd1 Developer team size is positively related to development performance.
  - Distributed work
    - Hd2 Distributed work of development is positively related to development performance.
  - Hierarchy[Ahuja et al.(1999), Cummings and Cross(2003)]
    - Hd3 Hierarchy in developer team is negatively related to development performance.
  - Structural hole[Sparrowe et al.(2001), Cummings & Cross(2003)]
    - Hd4 *Structural hole* within team hinders development performance.

- User as co-developer
  - "Users are wonderful things to have,…. properly cultivated, they can become co-developers."
- *User feedback*
  - *"Given enough eyeballs, all bugs are shallow"*
    - Raymond(1998) *The Cathedral and the Bazaar*

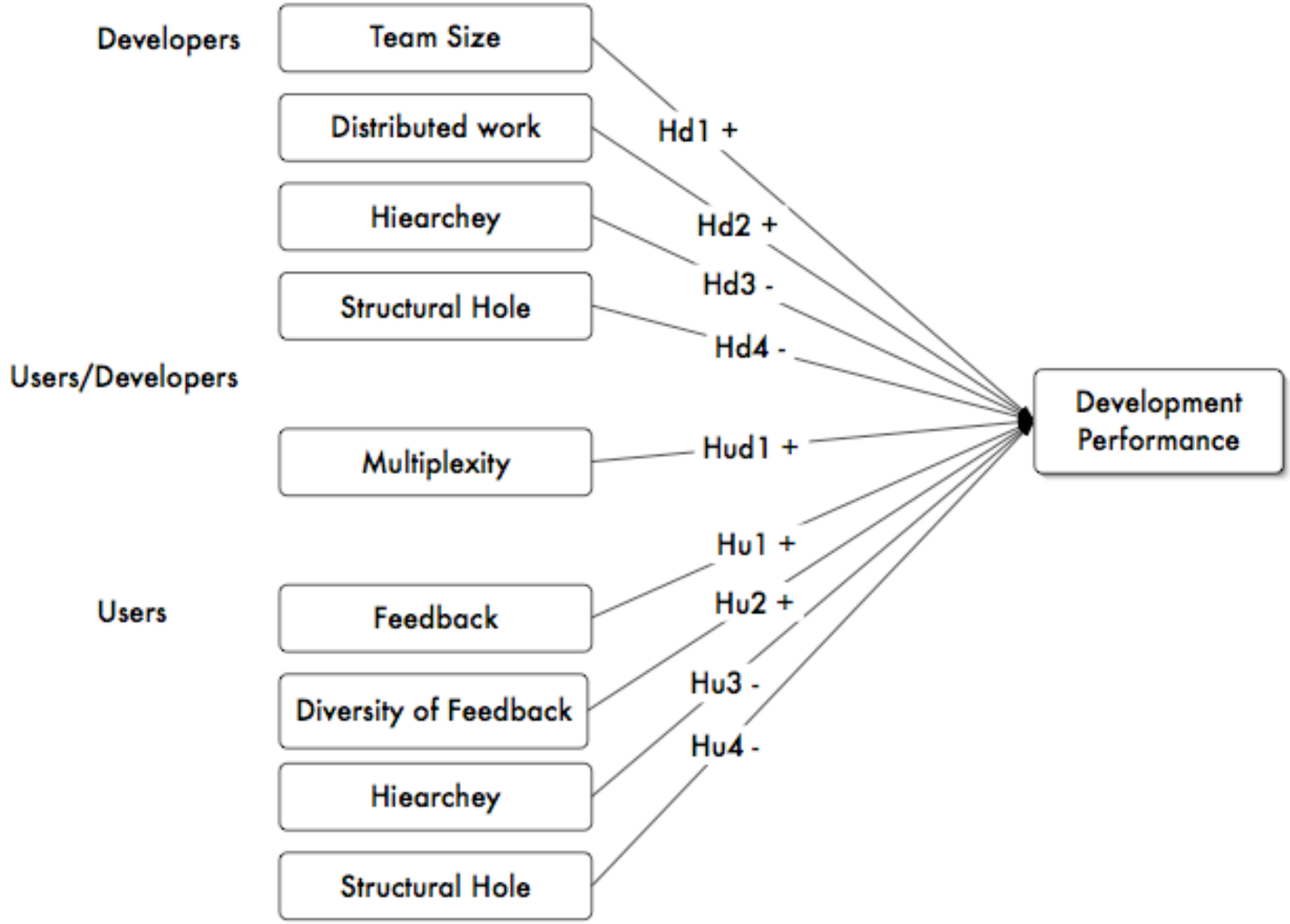
Hypotheses

- User
  - Feedback
    - Hu1 Feedback form users is positively related to development performance.
  - Diversity of feedback
    - Hu2 Diversity of communication among users is positively related to development performance.
  - Hierarchy
    - Hu3 Hierarchy in user community is negatively related to development performance.
  - Structural hole
    - Hu4 *Structural hole* within user community hinders development performance.

Hypotheses

- User/Developer
  - Multiplexity[Provan and Sebastian(1998)]
    - Hud1 Multiplexity of developers is positively related to development performance.

Hypotheses


Developers
Team Size
Distributed work
Hiearchey
Structural Hole
Users/Developers
Multiplexity
Users
Feedback
Diversity of Feedback
Hiearchey
Structural Hole
Hd1 +
Hd2 +
Hd3 -
Hd4 -
Hud1 +
Hu1 +
Hu2 +
Hu3 -
Hu4 -
Development Performance

Data

- Archives form *Sourceforge.net*
- Unit of Analysis
  - Project
  - Pooled data
- Sampling
  - 2 stage sampling
    - 2,200 projects were randomly selected
      - 10% of 22,000 projects as of May,2001
      - Top 100 projects in terms of page view, download, activity were added to cover really successful projects.
      - 2,101 projects
    - Further Screening
      - Released software?
      - Posted at least 100 messages?
      - CVS data is available?
      - 85 projects

Data

- Development performance
  - # of commitments to CVS (Concurrent Versioning System)/day
- Communication
  - Among developers
    - Messages posted to developers forum
  - User feedback
    - Messages posted to Bug Report and Feature Request forum
  - 0/1 matrices were composed to calculate SNA indexes.

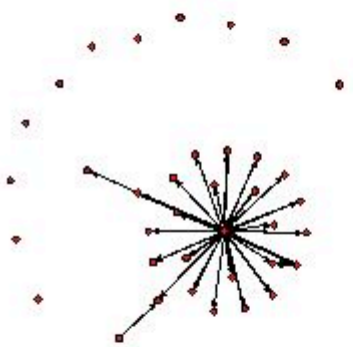
Variables

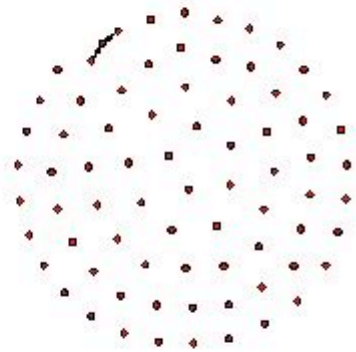
- Performance
  - Log(1+# of commitment /day)
- Developer
  - Team Size
    - Log(1+# of committers to CVS/days)
  - Distributed work
    - Freeman degree based Entropy at Development forum
  - Hiearchey
  - Structural hole
    - Density of development forum message matrix
- User
  - Feedback
    - Log(1+# of bug reports & feature request)
  - Variety of feedback
    - Freeman degree based Entropy at bug/feature request report forum
- User/Developer
  - Multiplexity
    - % of developers who posted messages to Bug & feature Request forum

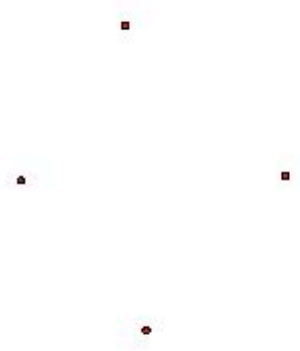
Look at Some Network

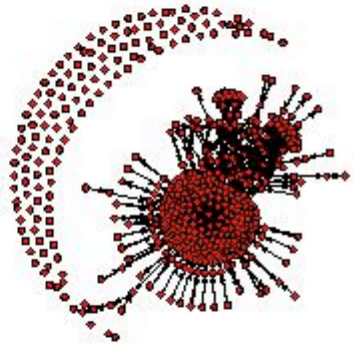
Pcmcia-cs

a) Open Discussion b)Feature request/bug report c)Support d)Development



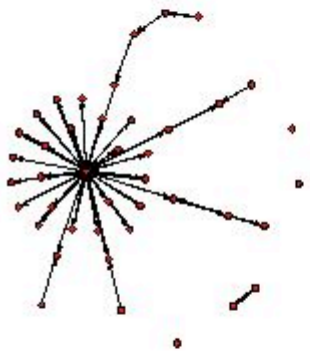


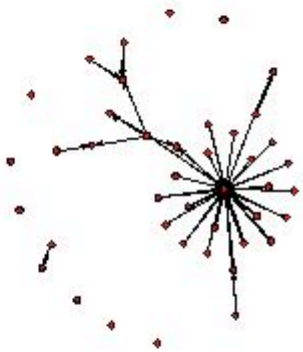


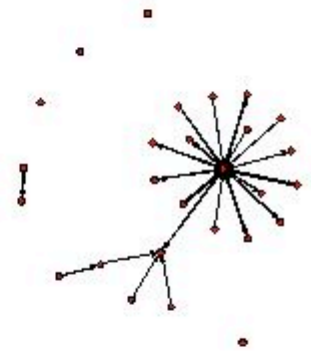


Cplus-plus

a) Open Discussion b)Feature request/bug report c)Support d)Development

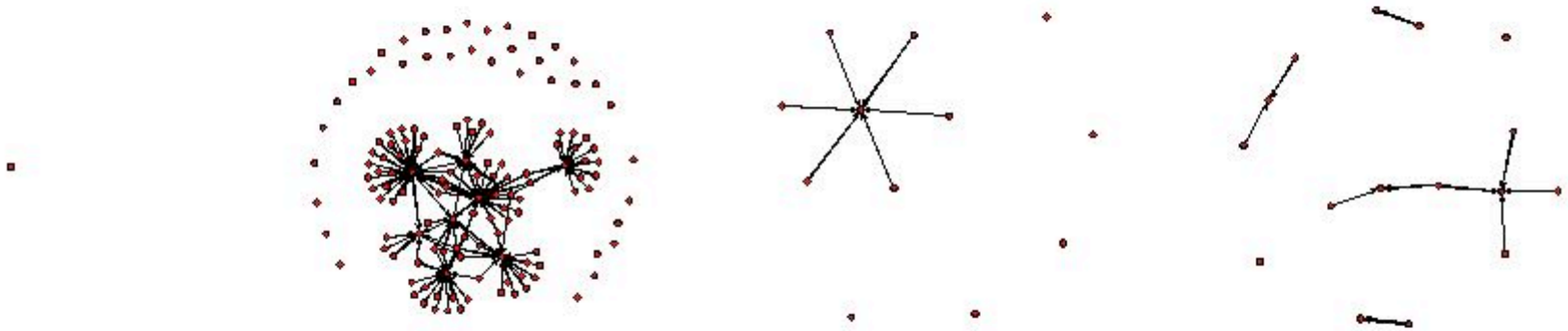






- dri

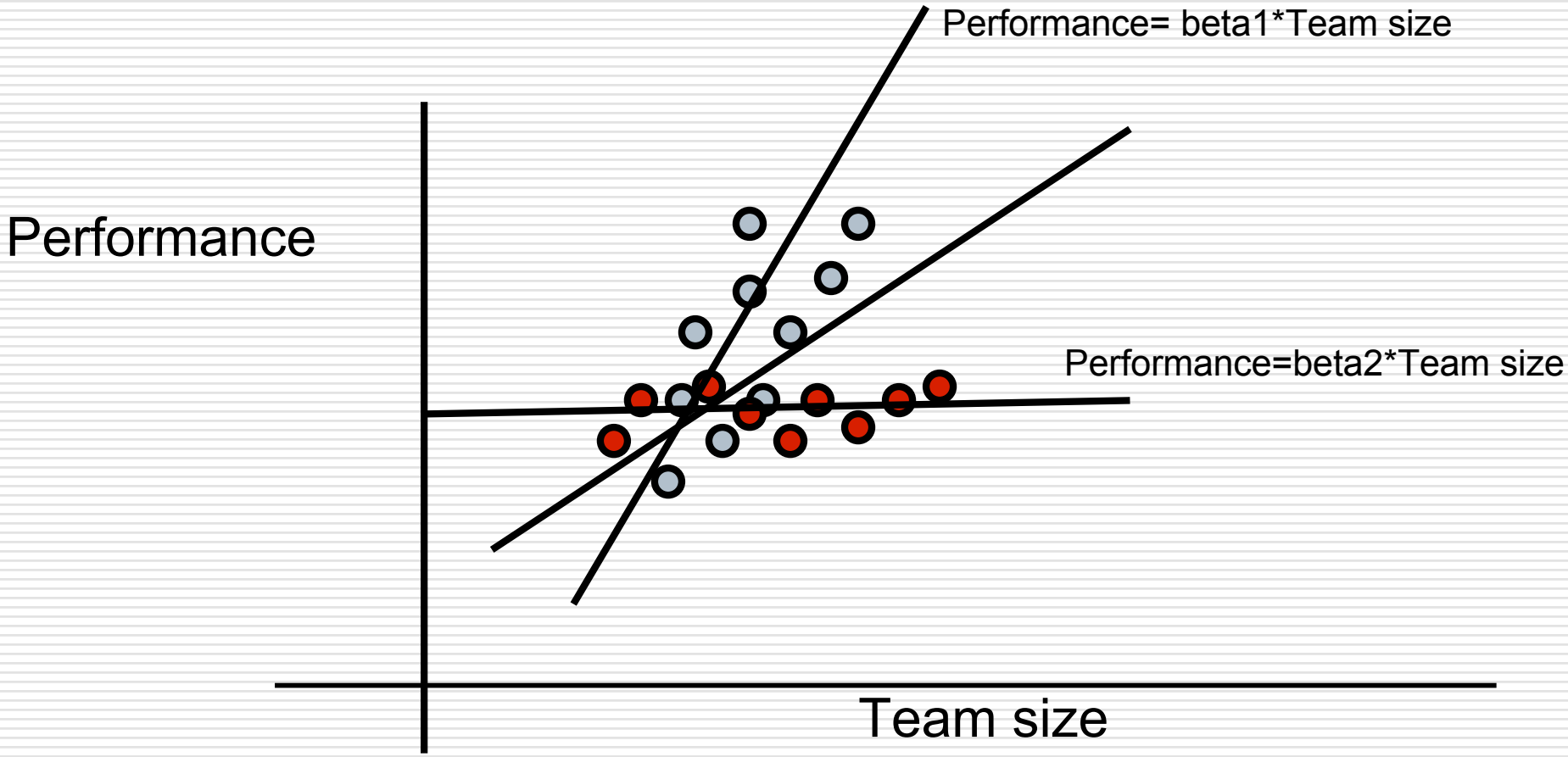
a) Open Discussion b)Feature request/bug report c)Support d)Development



Analysis

- Netrworks are heterogeneous
  - OLS,NLS,…., any method that assume homogeneous population will be misleading.
- Latent class regression model
  - A kind of *finite mixture model*[McLachlan and Peel 2000]
    - *flexmix library on R*
    - Typical application in marketing
      - Consumers are different in terms of price sensitivity. But we don't know who is price sensitive.
      - We don't have enough data to estimate parameters at individual level.

Example of Mixture of heterogeneous population

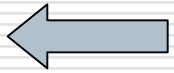


Result

- How many segments?

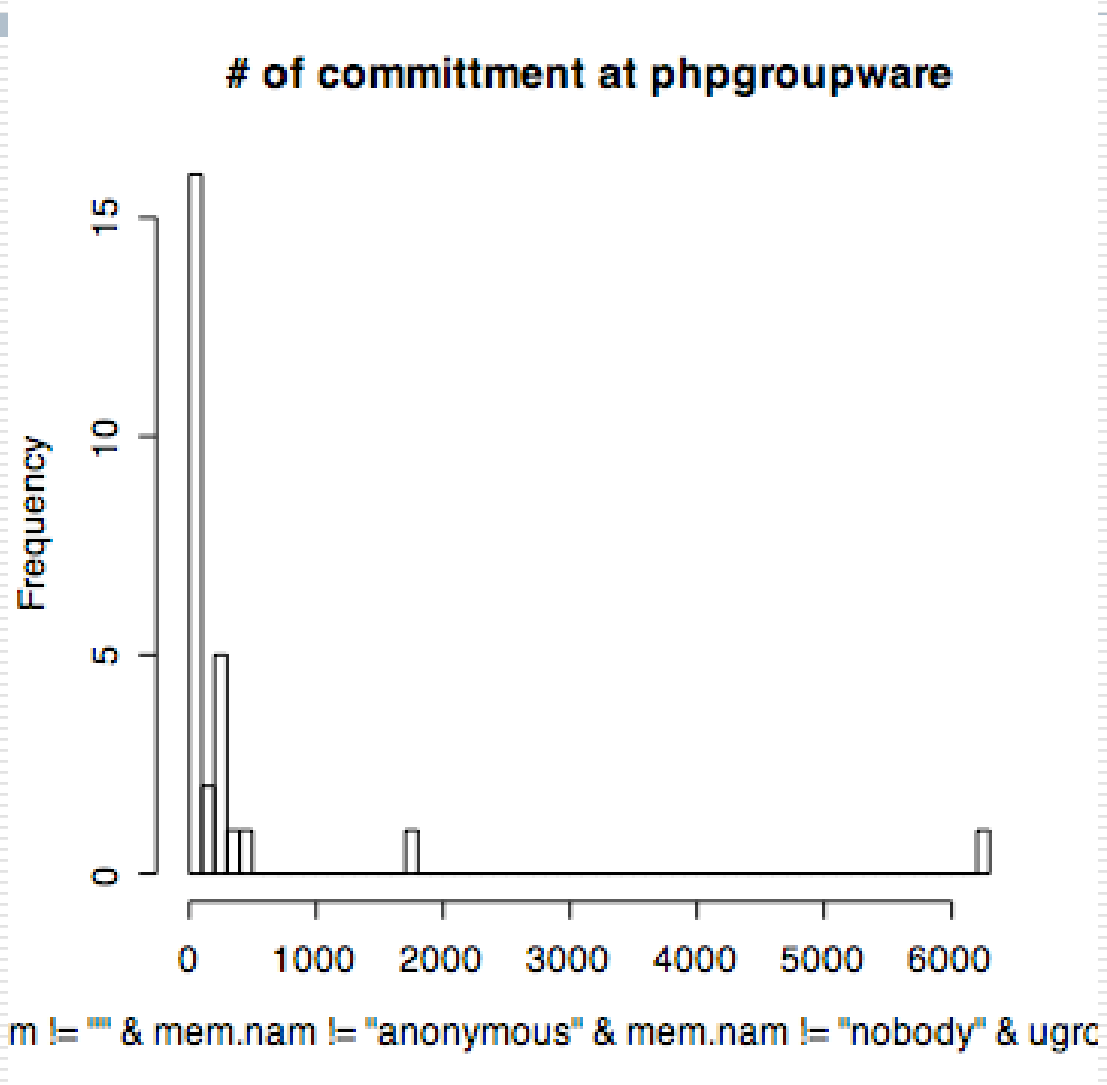
# of segment and model fit

# of segment	AIC	BIC	Segment size
1	312.3	408.2	85
2	218.0	387.3	25 60
3	68.9	366.8	25 34 26
4	not converged	not converged	



Estimated Parameters

			Expected sign	One Segment			Two Segments					
							Segment 1 (N=25)			Segment 2 (N=60)		
	Variables			b	t-value		b	t-value		b	t-value	
	Intercept			1.663	1.59		2.515	2.14	*	7.483	9.02	***
Developer	Team size	Hd1	+	0.602	2.14	**	0.686	3.18	**	0.064	0.35	
	Distributed work	Hd2	+	-1.245	-2.03	**	-4.282	-3.36	**	-1.240	-3.90	***
	Team size x distributed			0.522	2.28	**	1.082	2.67	**	0.648	5.24	***
	Hierarchy	Hd3	-	-0.525	-0.21		-19.348	-3.30	**	-0.103	-0.08	
	Structural Hole(-Density	Hd4	+	3.743	0.23		96.961	3.05	**	5.057	0.58	
Users	Feedback	Hu1	+	0.491	1.33		0.427	0.66		0.786	4.03	***
	Diversitiy of feedback	Hu2	+	-0.495	-0.67		-3.443	-2.20	*	-0.591	-1.43	
	Feedback x diversity			0.039	0.26		0.833	2.30	*	-0.091	-1.07	
	Hierarchy	Hu3	-	-1.326	-0.55		-9.129	-2.31	*	-1.378	-1.02	
	Structural Hole(-Density	Hu4	+	27.757	1.62		81.583	3.89	***	21.118	2.11	**
Dev/Users	Multiplexity	Hud1	+	0.207	0.72		0.370	0.82		0.435	2.95	***
Control Va	Log(1+# of Open Discussion)		+	0.656	1.80	*	-0.344	-0.40		0.425	2.25	**
	Diversity			0.305	0.58		1.157	1.91		-0.020	-0.07	
	OD x Diversity			-0.204	-1.93	*	-0.175	-1.14		-0.094	-1.63	
	Structural Hole(-Density)			-5.032	-0.93		28.733	2.67	**	-3.791	-1.34	
	Have developers forum?			-0.182	-0.38		5.510	6.06	***	-1.030	-4.36	***
	Have users forum?			-1.644	-1.20		-7.471	-4.60	***	-2.326	-3.00	***
	Have open discussion forum?			-2.372	-1.98	*	-2.750	-1.55		-5.751	-7.11	***
		AdjR2		0.353			0.965			0.754		


# of committment at phpgroupware
Frequency
0
5
10
15
0
1000
2000
3000
4000
5000
6000
m != "" & mem.nam != "anonymous" & mem.nam != "nobody" & ugrc

Descriptive statistics of each segments

		Segment 1	Segment 2	ANOVA
Developers	# of committers to CVS	12.720	6.683	ns
	Entropy	0.628	0.788	ns
	Hierarchy	0.135	0.115	ns
	Density	0.027	0.023	ns
Users	# of feedback	107.240	110.583	ns
	Entropy	1.768	1.772	ns
	Hierarchy	0.069	0.073	ns
	Density	0.020	0.019	ns
Developers/Users	Multiplexity	0.680	0.633	ns

Summary and Conclusion

- Research Question
  - Does communication structure among developers, users, and user-developers affect *development* performance?
- *Yes!*
  - Pattern of effect depends on type of OSSP projects.
    - We identified two types of projects.
- Methodology
  - Latent class regression

Limitation & Future Research

- Pooled data/cross sectional analysis
  - Correlation?
- Development Performance
- Group level analysis

- Panel data analysis
  - To test causality
- Innovativeness of software
- # of good idea(feedback) form users
- Individual level analysis

Communication Structure and Performance of Open Source Software Projects

April 29, 2006

Yutaka Hamaoka
Keio University
Hamaoka@fbc.keio.ac.jp