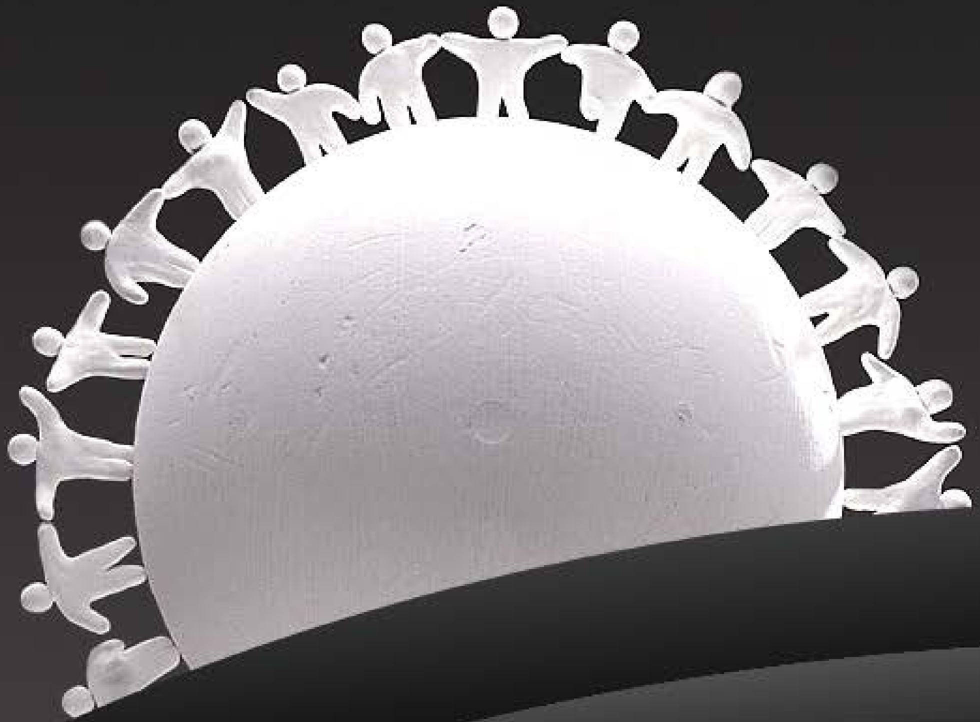
Team Activities Measurement Method for Open Source Software Development Using the Gini Coefficient



Ayako Masuda

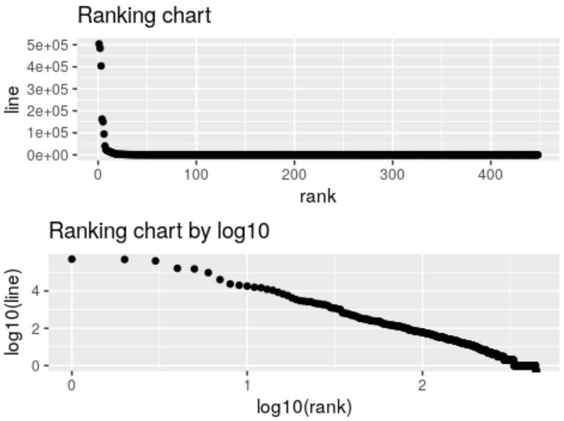
FeliCa Networks, Inc.

Agenda

1. Introduction
2. Summary
3. How to Measure the Team Activities in the OSSD
4. Result of the Measurement
5. Discussion
6. Conclusion

1-1. Background (1/3)

- Open source software (OSS) has made remarkable contributions as social infrastructure.
- The model of the OSS development (OSSD) team is becoming the de facto standard.
- However, there are many areas where OSSD team activities are not clear.
- We are conducting empirical research on OSSD team activities.
- Our previous research revealed the characteristic of the OSSD team, which was that the team consists of a small number of enthusiastic contributors (core members) and a large number of small contributors.


Ranking chart

Ranking chart by log10

Ranking of additional lines in atom.

1-1. Background (2/3)

- We tried to analyze the test activities of OSSD, but the 48 OSSD teams analyzed in this study did not confirm the existence of the members specializing in testing.

- In OSSD, in order to use a lot of other software, it is necessary to test including dependencies in the development stage.
- Moreover, developers fix bugs and respond to new requirements through frequent releases.
- One of the factors is that products are changing from software products themselves to services.
- In OSSD, the speed of response time supports product quality.

- In other words, it is considered that the difference in quality requirements is the difference in testing activities.

1-1. Background (3/3)

- Key factors for evaluating GitHub projects are its community structure and testing activity.
- In particular, evidence of testing in a GitHub project implies that the developers have spent considerable time and effort to ensure that the product adheres to its intended behavior.

- For example, 47.8% of the openlayers project, which is one of our target project codes, are test codes.
- Thus, half of the OSSD team activity is related to testing.

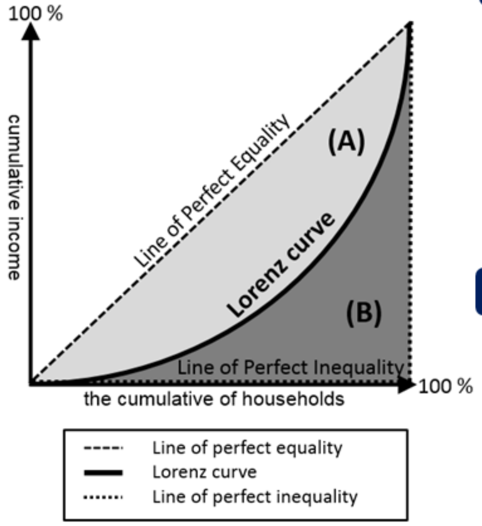
- The OSSD team develops source codes for realizing requirements and, further, test codes for testing the source codes.

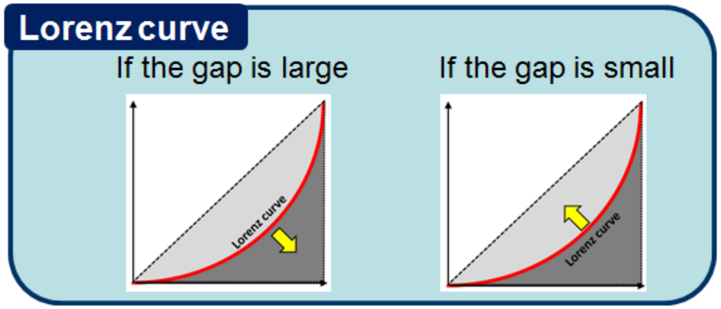
- This study measures the team activity in OSSD, and it includes aspects of software testing activity.

1-2. The Purpose of This Study

- This study focused on the load distribution of software development activities and measures the variance of contributors' activity using the Gini coefficient.

- The Gini coefficient and Lorenz curve are common indexes used to analyze the distribution of household income*.

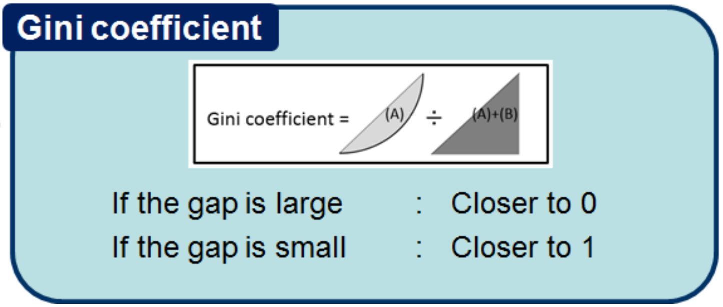

100 %
cumulative income
Line of Perfect Equality
(A)
Lorenz curve
(B)
Line of Perfect Inequality
100 %
the cumulative of households
----- Line of perfect equality
▬ Lorenz curve
······· Line of perfect inequality



Lorenz curve

If the gap is large

If the gap is small



Gini coefficient

Gini coefficient = (A) ÷ (A)+(B)

If the gap is large : Closer to 0
If the gap is small : Closer to 1

* Gastwirth, 1972; Nakamura, 2005

Agenda

1. Introduction
2. Summary
3. How to Measure the Team Activities in the OSSD
4. Result of the Measurement
5. Discussion
6. Conclusion

Summary (1/3)

- Measurement of load distribution using Gini coefficient
  - This study focused on the load distribution of software development activities and measures the variance of contributors' activity using the Gini coefficient.
  - Effort Person-Months in the months with commit record in the measurement period was used as the unit of measurement to measure amount of activity, and the Gini coefficient, commonly used in field of economics, was used as an index corresponding to the occupancy rate.
  - It was confirmed that this method can be used to calculate the Gini coefficient of Effort Person-Months in real OSSD projects.

Summary (2/3)

- The influence of activity variance
  - The influence of activity variance was investigated by confirming whether the Gini coefficient of Effort Person-Months correlated with the development period, contributor number, total Effort Person-Months number, total commits numbered, repository size, star number, fork number, and issue number.
  - However, it was found that there was almost no correlation with items other than the development period.
  - Given that there is almost no impact of large variance in activity amount on commit, which is directly linked to the deliverable, it is assumed that there is almost no lowering of occupancy rate of personnel due to free participation.

Summary (3/3)

- The characteristics of OSSD from the viewpoint of Gini coefficient
  - The OSSD projects were classified with the Gini coefficient in order to investigate the characteristics of the OSSD project from the Gini coefficient, and the distribution of each item among the projects were compared, as well.
  - However, OSSD projects with a long development period did not have any characteristics other than the trend that the Gini coefficient was large.

Agenda

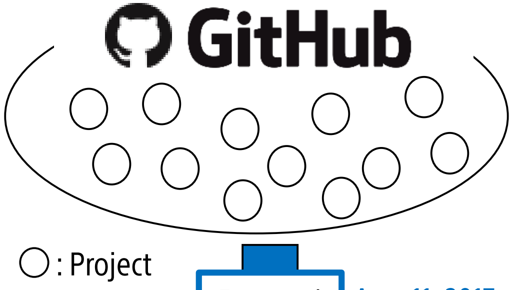
1. Introduction
2. Summary
3. How to Measure the Team Activities in the OSSD
4. Result of the Measurement
5. Discussion
6. Conclusion

Analysis Target

The require conditions of the project :

- Continued development for a certain period of time
- Participation by a certain number of contributors
- Continuous entry of new contributors

Previous Study



Extracted

Aug. 11, 2017

50 Projects

1. Using GitHub Advanced Search function, extract projects that satisfy the following conditions:
   - Registered with GitHub in 2012 and continued until 2017
   - Repository size is 15 MB or larger
   - More than 200 forks
   - More than 1,000 stars showing evaluation by OSS users

2. In the project extracted by 1, both values of the number of contributors and commits belong to the second quartile or more

Updated

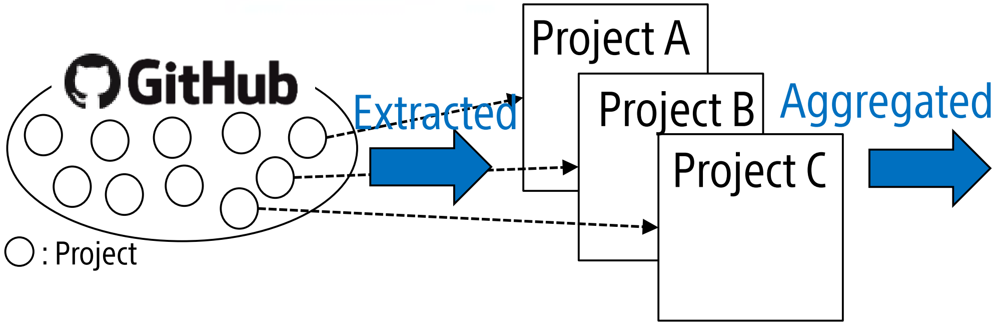
Nov. 30, 2018

This Study

48 Projects

In updated process, because there were two projects in which the repository was moved, 48 OSSD project data were used.

Analysis Data



GitHub

○ : Project

Extracted

Project A

Project B

Project C

Aggregated

48 PJ's Commit Records

Project A

author	month	Commit	files	add	delete	EM
aaa	2012-12	668	866	3465	5519	1
aaa	2013-01	450	720	2370	13197	1
aaa	2013-02	366	683	3128	3569	1
			:			
bbb	2013-11	1	2	38	3	1
bbb	2013-12	1	1	1	1	1

Aggregated

Project A

author	Start	commit	files	add	delete	EM
aaa	2012-12	702	3112	217356	88372	49
bbb	2013-11	1332	5442	127000	91494	137
ccc	2014-12	5487	20775	573063	478714	286
			:			
yyy	2015-12	6914	33529	553354	444034	541
zzz	2017-12	1648	6552	279058	134294	422

Effort Person-Months (EM)

- Collects commit records for each contributor for the entire period of the project from Git log
- EM for each contributor was calculated on a monthly basis
- EM was counted in contributor units

Agenda

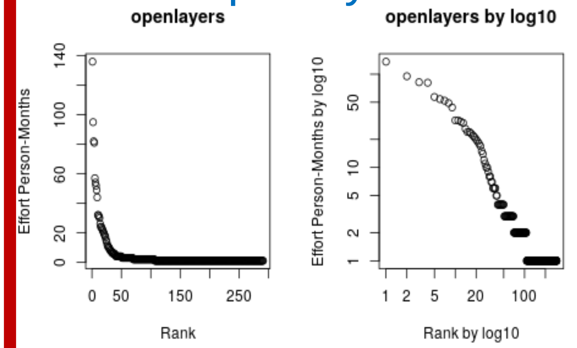
1. Introduction
2. Summary
3. How to Measure the Team Activities in the OSSD
4. Result of the Measurement
5. Discussion
6. Conclusion

Selected 3 OSSD Projects Data

	openlayers	meteor	llvm
Product	A JavaScript library that displays dynamic maps on web pages.	A real-time web application framework that enables reactive web programming.	A compiler base that can be used for any programming language.
Started in	May 2006	May 2006	June 2001
Period (year)	12	7	17
Contributors (number)	267	476	919
Repository size (KB)	77,614	77,585	940,703
Forks (number)	1.778	5,028	1,720
Stars (number)	4,453	40,706	3,540
Issues (number)	620	267	0
Effort-Months (EM)	152	144	2,181
Gini Coefficient (value)	0.7217878	0.5706287	0.6091536

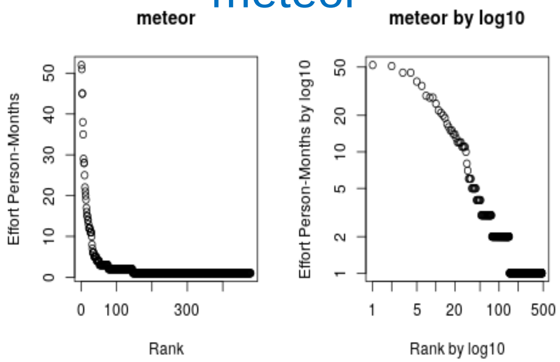
Selected 3 OSSD Projects Data

openlayers



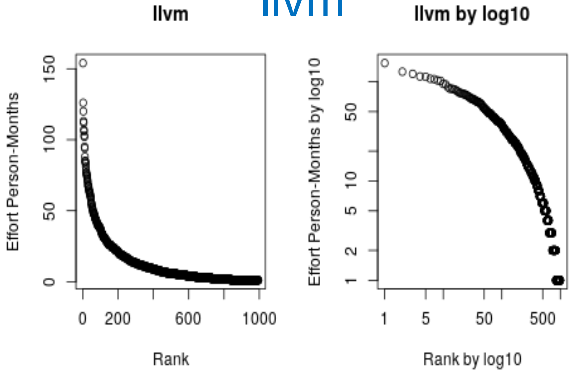
Distribution of Effort Person-Months

meteor



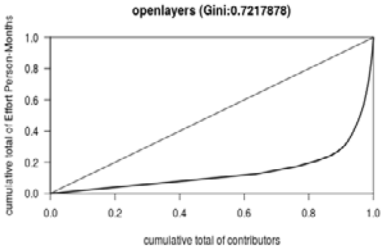
Distribution of Effort Person-Months

llvm



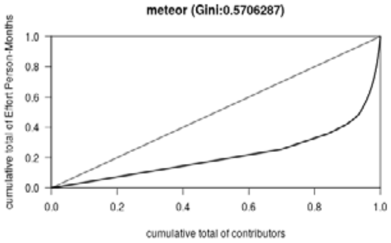
Distribution of Effort Person-Months

Effort Person-Months for each contributor was found to be a power-law distribution



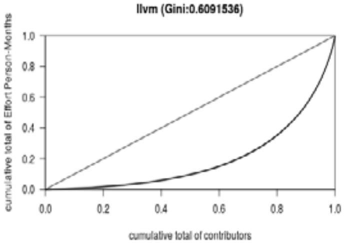
Lorenz curve

Gini : 0.7217878



Lorenz curve

Gini : 0.5706287



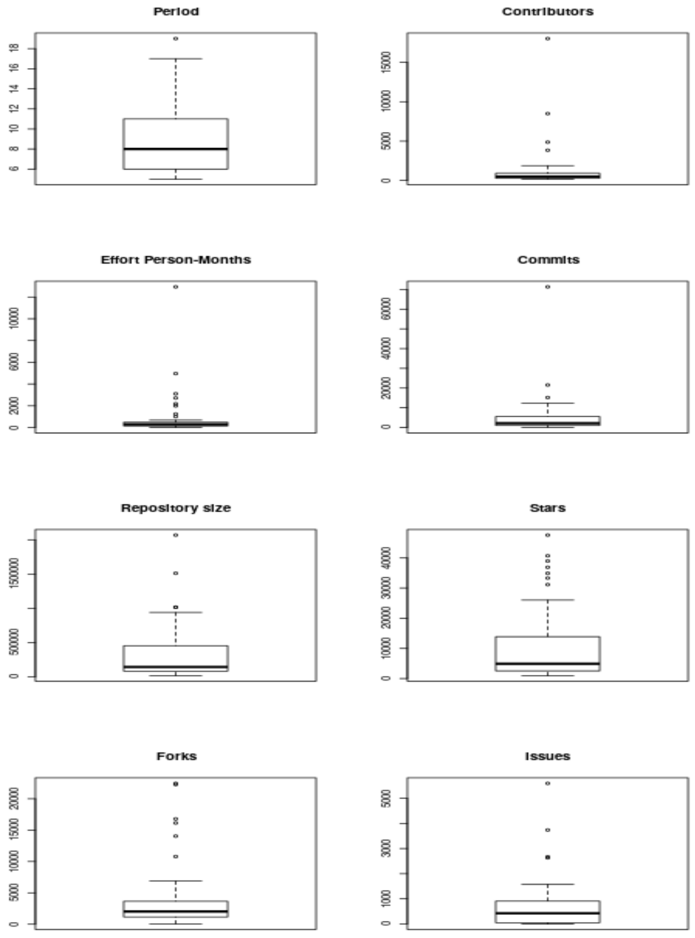
Lorenz curve

Gini : 0.6091536

The variance of contributor's activity amount is large.

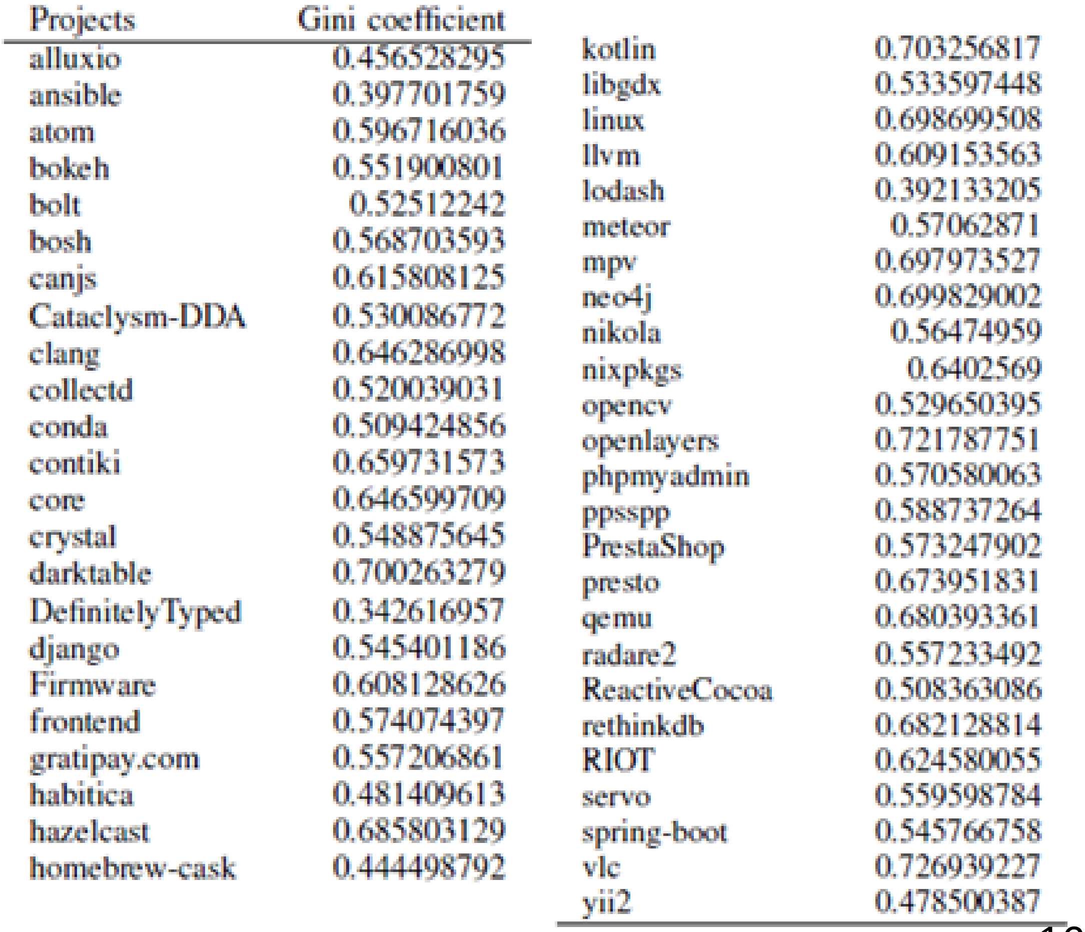
The Gini coefficient of Effort Person-Months can be calculated in three extracted OSSD projects.

Result of the Measurement



Distribution of each item in all PJs

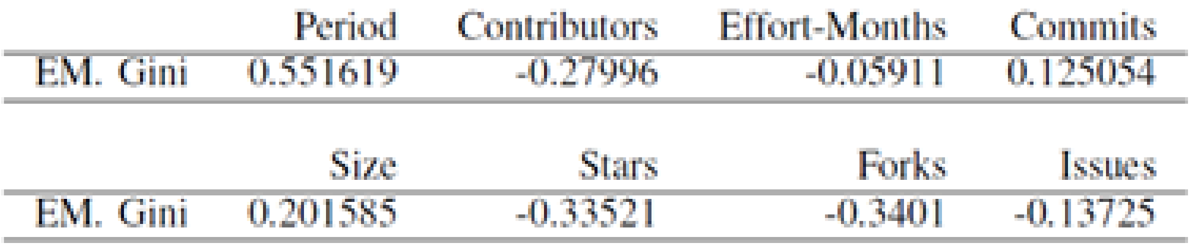
The Gini Coefficient of Effort Person-Months in All PJs.



Projects	Gini coefficient
alluxio	0.456528295
ansible	0.397701759
atom	0.596716036
bokeh	0.551900801
bolt	0.52512242
bosh	0.568703593
canjs	0.615808125
Cataclysm-DDA	0.530086772
clang	0.646286998
collectd	0.520039031
conda	0.509424856
contiki	0.659731573
core	0.646599709
crystal	0.548875645
darktable	0.700263279
DefinitelyTyped	0.342616957
django	0.545401186
Firmware	0.608128626
frontend	0.574074397
gratipay.com	0.557206861
habitica	0.481409613
hazelcast	0.685803129
homebrew-cask	0.444498792
kotlin	0.703256817
libgdx	0.533597448
linux	0.698699508
llvm	0.609153563
lodash	0.392133205
meteor	0.57062871
mpv	0.697973527
neo4j	0.699829002
nikola	0.56474959
nixpkgs	0.6402569
opencv	0.529650395
openlayers	0.721787751
phpmyadmin	0.570580063
ppsspp	0.588737264
PrestaShop	0.573247902
presto	0.673951831
qemu	0.680393361
radare2	0.557233492
ReactiveCocoa	0.508363086
rethinkdb	0.682128814
RIOT	0.624580055
servo	0.559598784
spring-boot	0.545766758
vlc	0.726939227
yii2	0.478500387

Influence of variance of activity amount

- To investigate the impact of activity variance, it was found if the Gini coefficient of Effort Person-Months correlated with the development period, contributor number, total Effort Person-Months number, total commits numbered, repository size, star number, fork number, and issue number.



	Period	Contributors	Effort-Months	Commits
EM. Gini	0.551619	-0.27996	-0.05911	0.125054

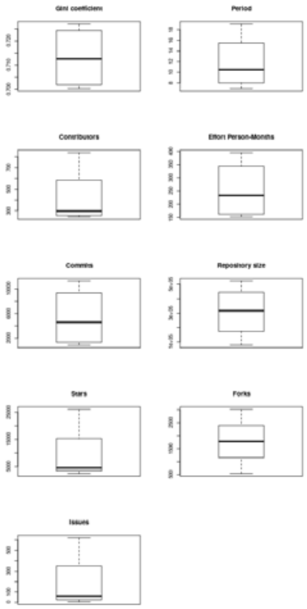
	Size	Stars	Forks	Issues
EM. Gini	0.201585	-0.33521	-0.3401	-0.13725

- According to Table , although the development period had a correlation, the results showed weak or almost nil correlation with other items.

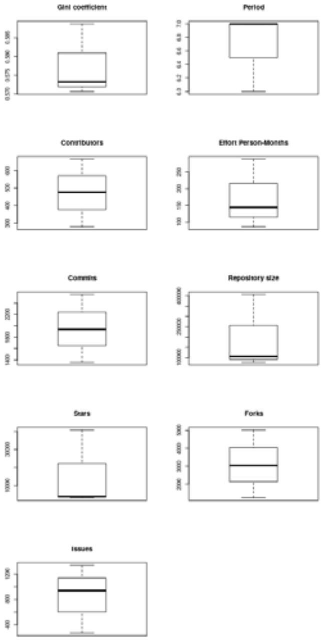
This implies that the influence of the variance of the activity amount on the development activity is weak.

Classification by Gini coefficient difference

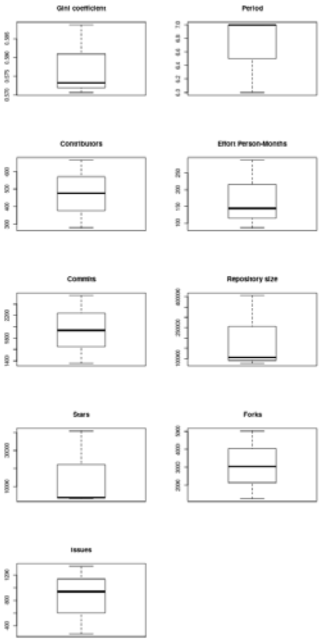
- We classified the Gini coefficients and tried to found the characteristics of the project on the basis of the differences among the groups.
- In the group A with a large Gini coefficient, the development period was observed to be long.
- However, the variance among the individual OSSD projects was large, and the characteristics of the project could not be investigated by comparing the groups.



Group A



Group B



Group C

Agenda

1. Introduction
2. Summary
3. How to Measure the Team Activities in the OSSD
4. Result of the Measurement
5. Discussion
6. Conclusion

Discussion

- The distribution of contributors' activity amount in an OSSD project obeys a power-law distribution with a small number of enthusiastic core members corresponding to the head and a large number of members engaged in a small amount of activity, such as beginners corresponding to the tail.

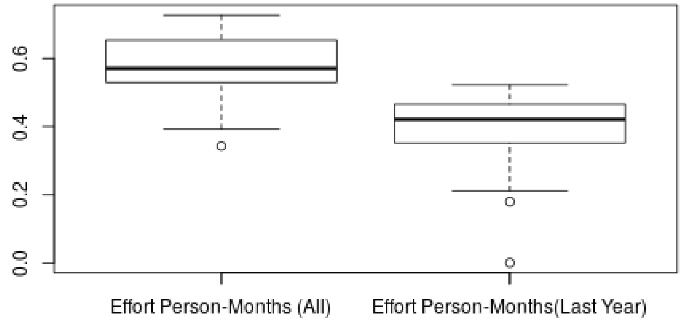
- Given that there is almost no impact of large variance in activity amount on commits, which is directly linked to the deliverable, it is considered that there is almost no lowering of occupancy rate of personnel due to free participation.

- OSSD projects are diverse in terms of the target and difficulty level of development, project operation, and so on.

- Among various OSSD projects, the Gini coefficient tended to be large for those with a long development period.

Discussion

- The Gini coefficient of Effort Person-Months calculated over the entire measurement period was compared with the Gini coefficient calculated over the last year in the measurement period.



Comparison of the Gini coefficient for the whole period and the last year

- The Gini coefficient for the last year of the measurement period was lower than the coefficient of the whole period.
- This result is because the cumulative number of small activity contributors who have committed only once is small in relation to the whole period.
- This difference in the Gini coefficient indicates the occupancy rate of contributors whose period of participation in the OSSD project was short.
- The time series analysis of the OSSD project will help clarifying the activity of the contributors.

Agenda

1. Introduction
2. Summary
3. How to Measure the Team Activities in the OSSD
4. Result of the Measurement
5. Discussion
6. Conclusion

Conclusion

Findings

- In OSSD, development activities and testing activities are not separated, and testing activities are conducted in the process of development.
- Development activities are promoted mainly by a few core members.
- Some projects have been promoted by the same core members, while others have new core members appearing one after another.

- The measurement of load distribution of the activity of OSSD can use Gini coefficient as an index.
- Given that there is almost no impact of large variance in activity amount on commit, which is directly linked to the deliverable, it is assumed that there is almost no lowering of occupancy rate of personnel due to free participation.

Future Task

- Future research should focus on a time series analysis in the OSSD projects and clarify the activity situation of the contributors.

Thank you for your attention.