Analysis of Coordination between Developers and Users in The Apache Community

Yasutaka Kamei, Shinsuke Matsumoto,
Hirotaka Maeshima, Yoji Onishi,
Masao Ohira, Ken-ichi Matsumoto

Nara Institute of Science and Technology
Japan

International Conference on Open Source Systems@ Milan Italy - 9 Sep. 2008

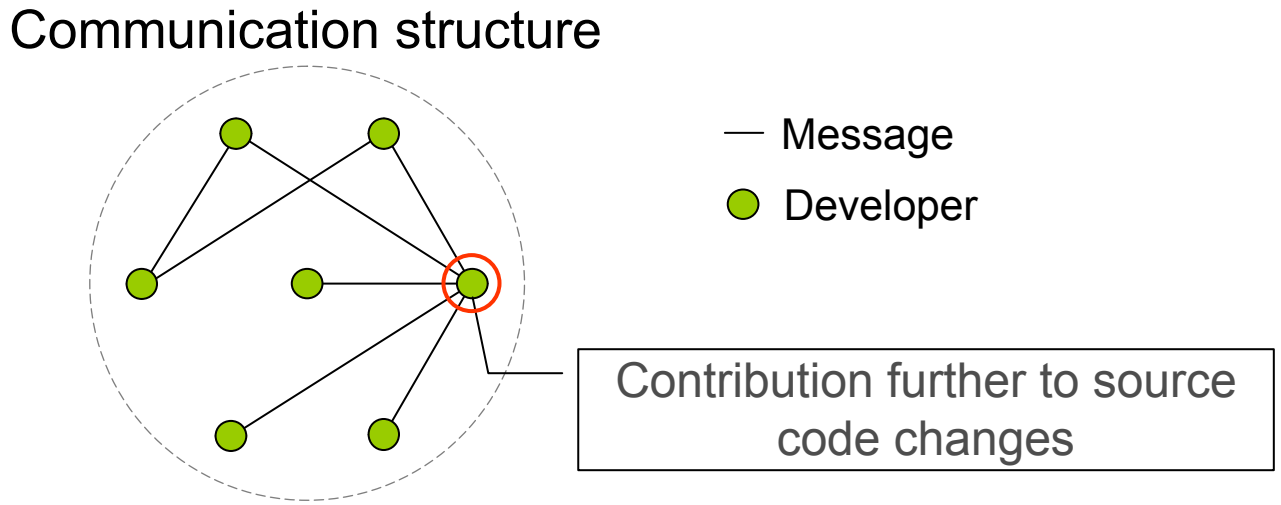
Introduction -background-

- OSS development is geographically distributed software development in which developers around the world communicate with each other

- Analyzing communication among developers in OSS development communities would help us better understand distributed software development
  - *How do developers coordinate with each other ?*
  - *How do developers collaborate with each other ?*

Introduction -analysis of communications among developers-

- Many studies reported analysis of actual communications among developers in OSS communities
  - A developers communicating with many developers much contributed to source code changes*



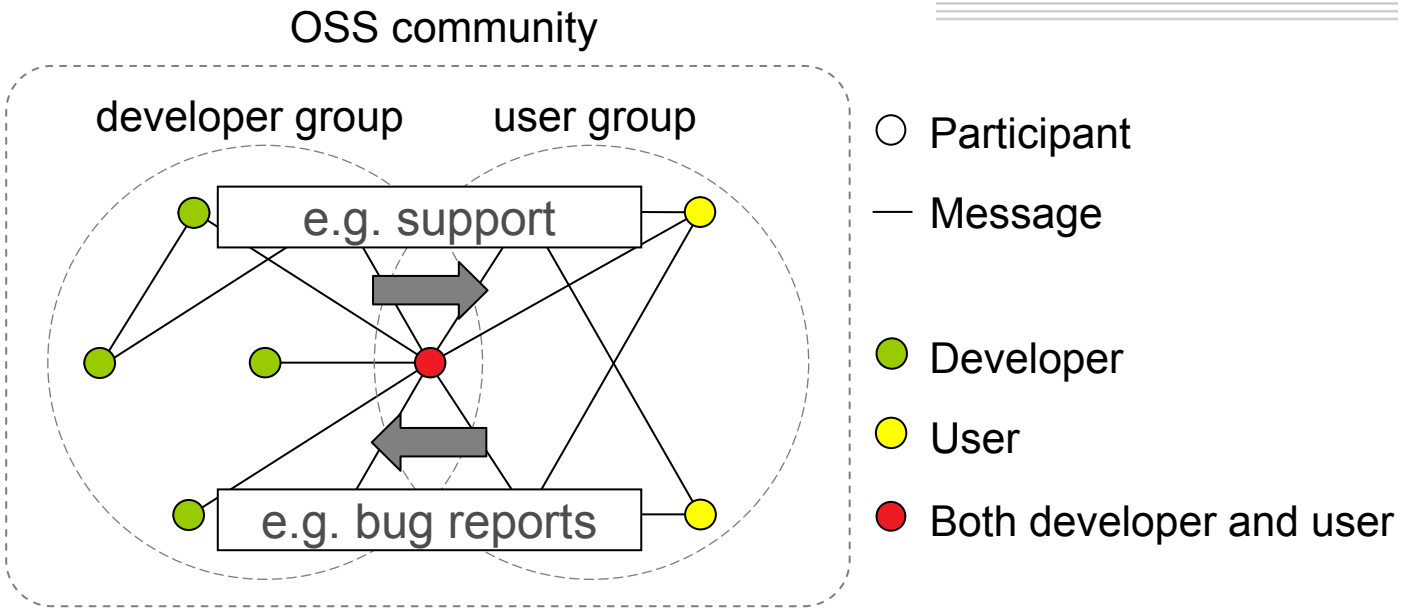
* C. Bird et al. Mining email social networks. In Proc. the 2006 International Workshop on Mining Software Repositories (MSR'06), pages 137–143, 2006.

Introduction -importance of user for OSS community-

- Raymond* pointed out that users in an OSS community played an important role as co-developers
  - A considerable use of software products by many users can lead to bug reports for the products and then the improvement of their quality

* E. S. Raymond. The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary. O'Reilly and Associates, 1999.

Introduction -important role of collaboration-



- We consider that the participants who belong to both developer and user groups have an important role of collaboration among developers and users in an OSS community

Goal and approach

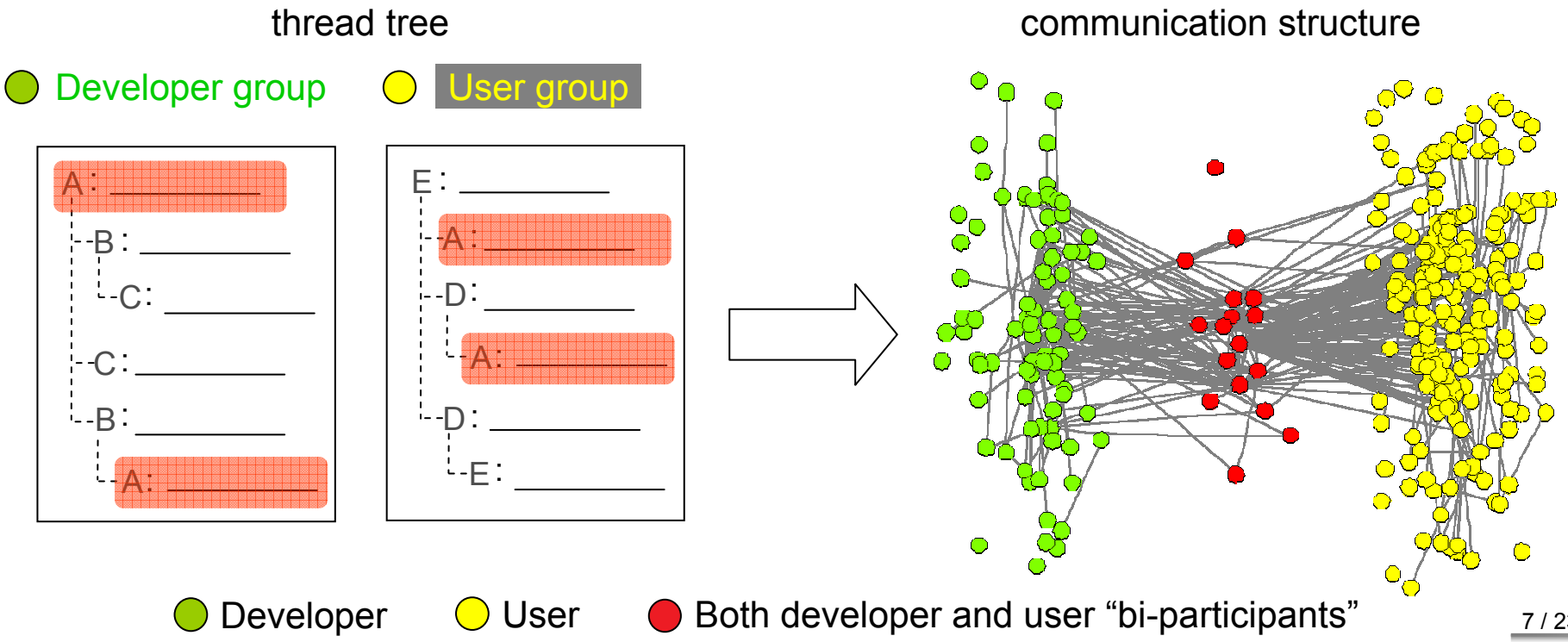
- Goal
  - To obtain a better understanding on collaboration and coordination in OSS communities
- Approach
  - We analyze communication structures in OSS communities with a focus on participants who belong to both developer and user groups “bi-participants” and assist the collaboration between them

Analysis of bi-participants

- Visualizing the communication structure
- Calculating the betweenness centrality
- Reading contents of messages

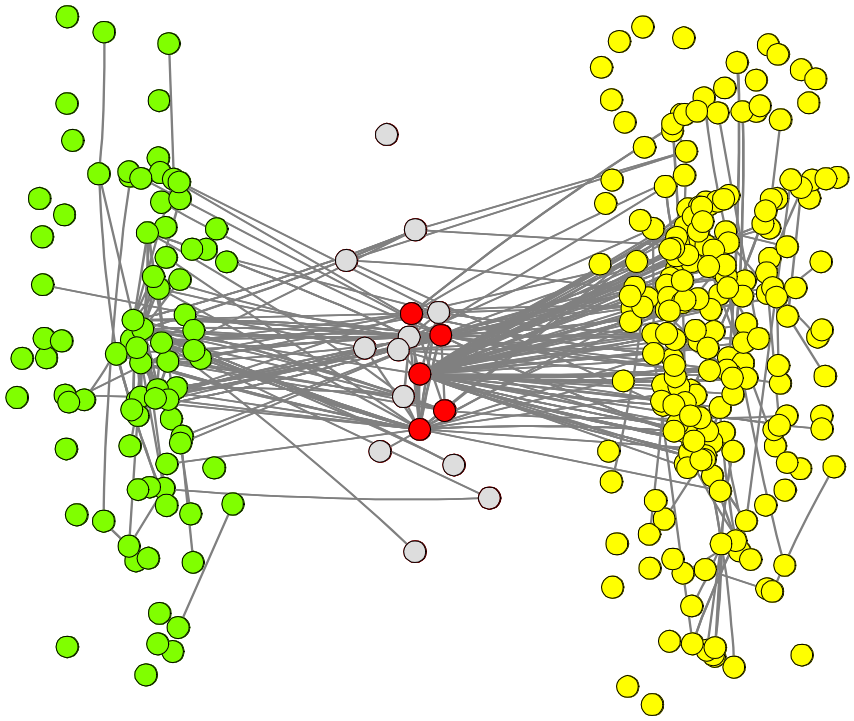
Visualizing the communication structure

- To understand the whole picture of the structure
  - We define a communication structure as sender-receiver relationships in online media



Calculating the betweenness centrality

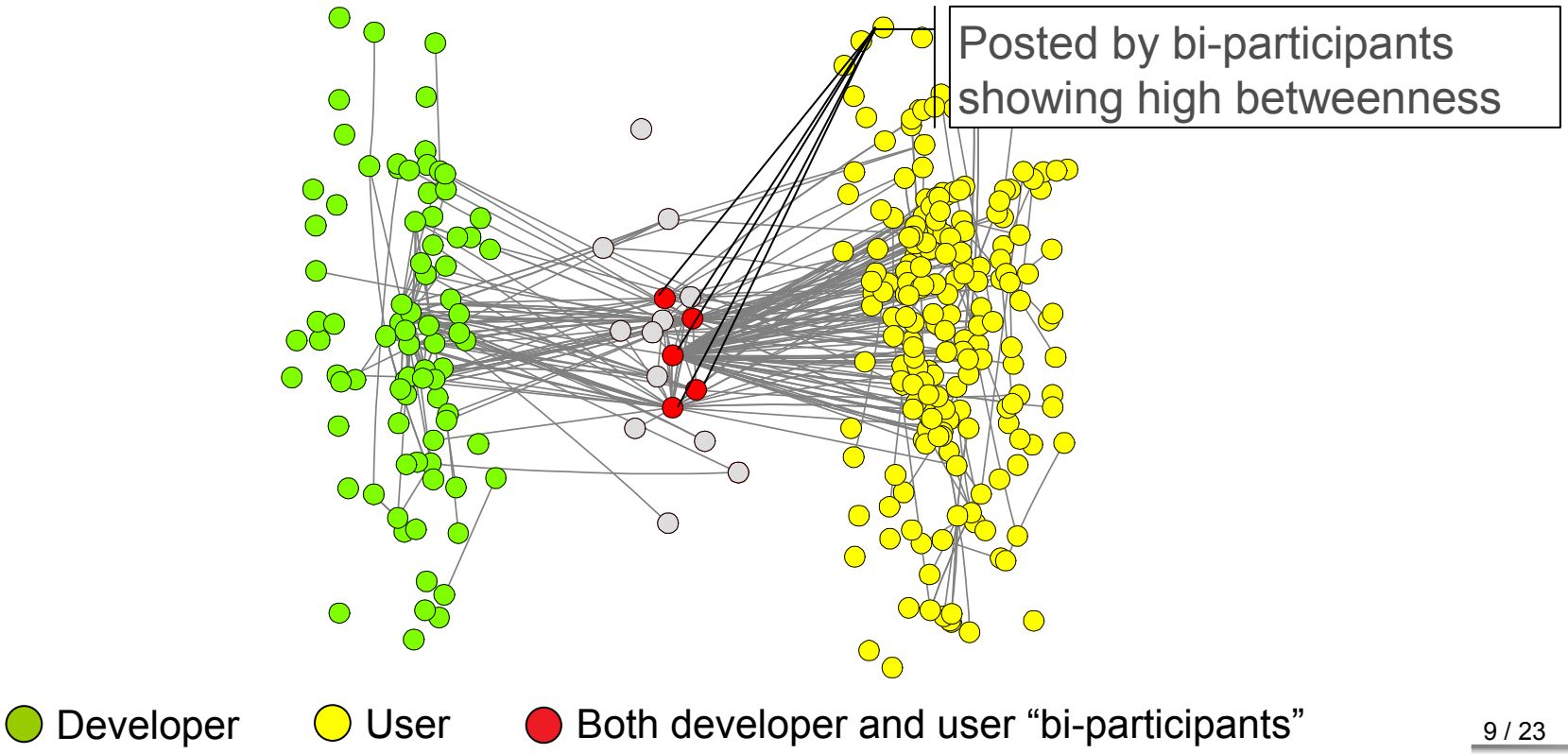
- The betweenness centrality
  - It takes a value from 0 to 1, and a node with higher betweenness indicates the node which plays a role of a intermediator in the network.



Developer User Both developer and user “bi-participants”

Reading contents of messages

- To confirm whether bi-participants coordinates activities between the both groups, or not.



Case Study

- Target OSS project
  - Apache HTTP Server

    It has been developing a web server software product with the biggest market share
- Target Communication data
  - Developers mailing list
  - Users mailing list
- Target Period

Data Cleaning

- We clean the data to identify the participants who have several email addresses before our analysis
  1. A sender of messages with the same email address and different ”Name”:
  2. A sender of messages with the same name at “From” and partially the same address before at mark(@):
  3. A sender of messages with the same name at ”From” and different email addresses:

Data Cleaning -Step 1.-

- A sender of messages with the same email address and different ”Name”:

Message A

```
From: Yasutaka Kamei <yasuta-k@is.naist.jp>
To: dev@httpd.apache.org
Subject: test

 --- Email Body ---
```

Message B

```
From: Yasu <yasuta-k@is.naist.jp>
To: dev@httpd.apache.org
Subject: test

 --- Email Body ---
```

Data Cleaning -Step 2.-

- A sender of messages with the same name at “From” and the same address before at mark(@):

Message A

From: Yasutaka Kamei <yasuta-k@is.naist.jp>
To: dev@httpd.apache.org
Subject: test

--- Email Body ---

Message B

From: Yasutaka Kamei <yasuta-k@gmail.com>
To: dev@httpd.apache.org
--- Email Body ---

We found 678 unique senders by applying the Step 1 and 2 to the data.

Data Cleaning -Step 3.-

- A sender of messages with the same name at "From" and different email addresses:
  - We judged if they were used by the same person, by confirming the body of messages (e.g. messages with the same signature)

Message A

From: Yasutaka Kamei <yasuta-k@is.naist.jp>
To: dev@httpd.apache.org
Subject: test

Yasu

Message B

From: Yasutaka Kamei <yasu@gmail.com>
To: dev@httpd.apache.org
[illegible]

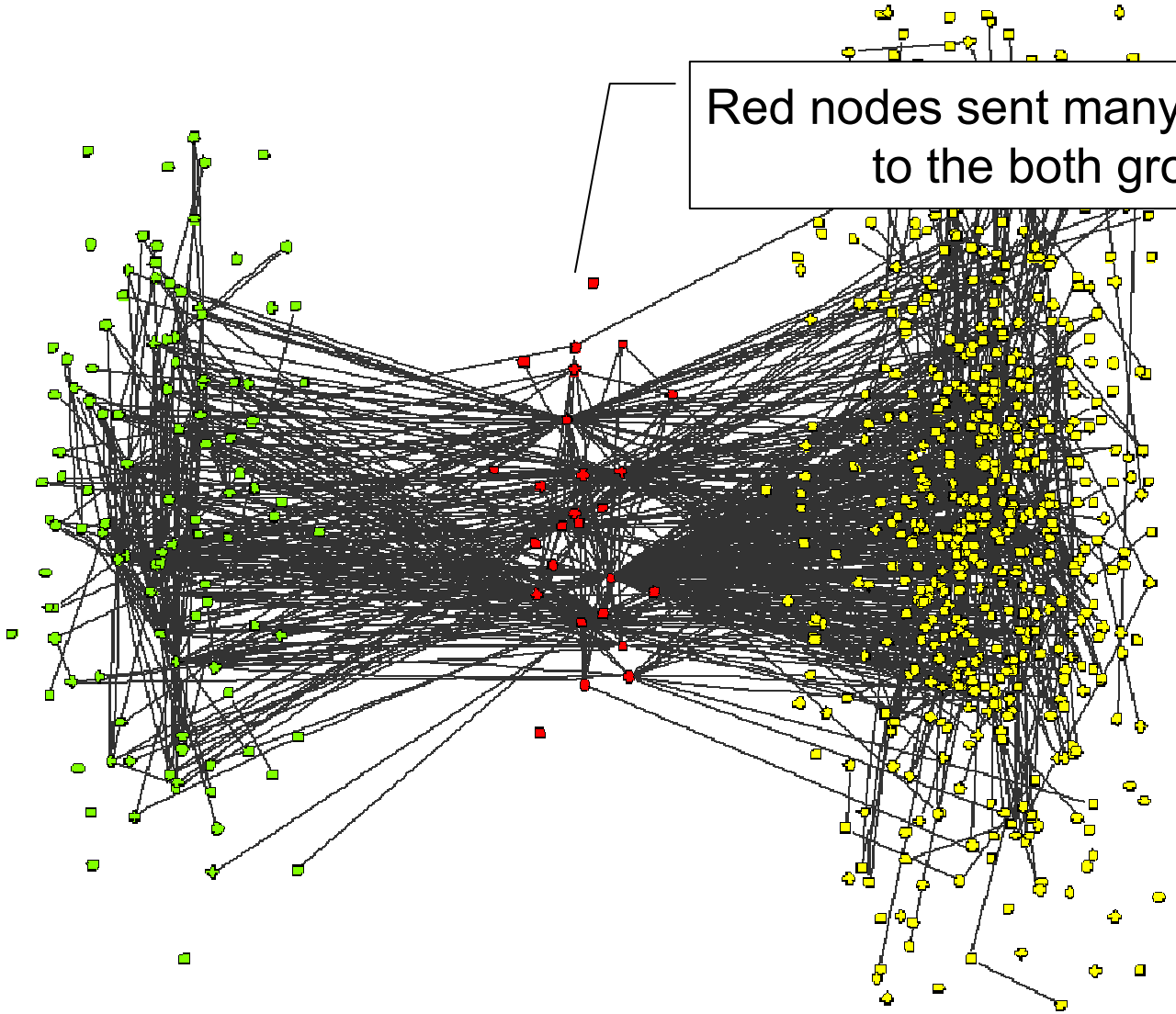
Yasu

Only 5 of 678 senders match the Step 3.

Result of analysis

- Visualizing the communication structure
- Calculating the betweenness centrality
- Reading contents of messages

Result of analysis -communication structure-



Developer User Both developer and user “bi-participants”

Result of analysis -betweenness centrality-

- The betweenness centrality of them is 10 times higher than the median, that means they intermediate between many developers and users

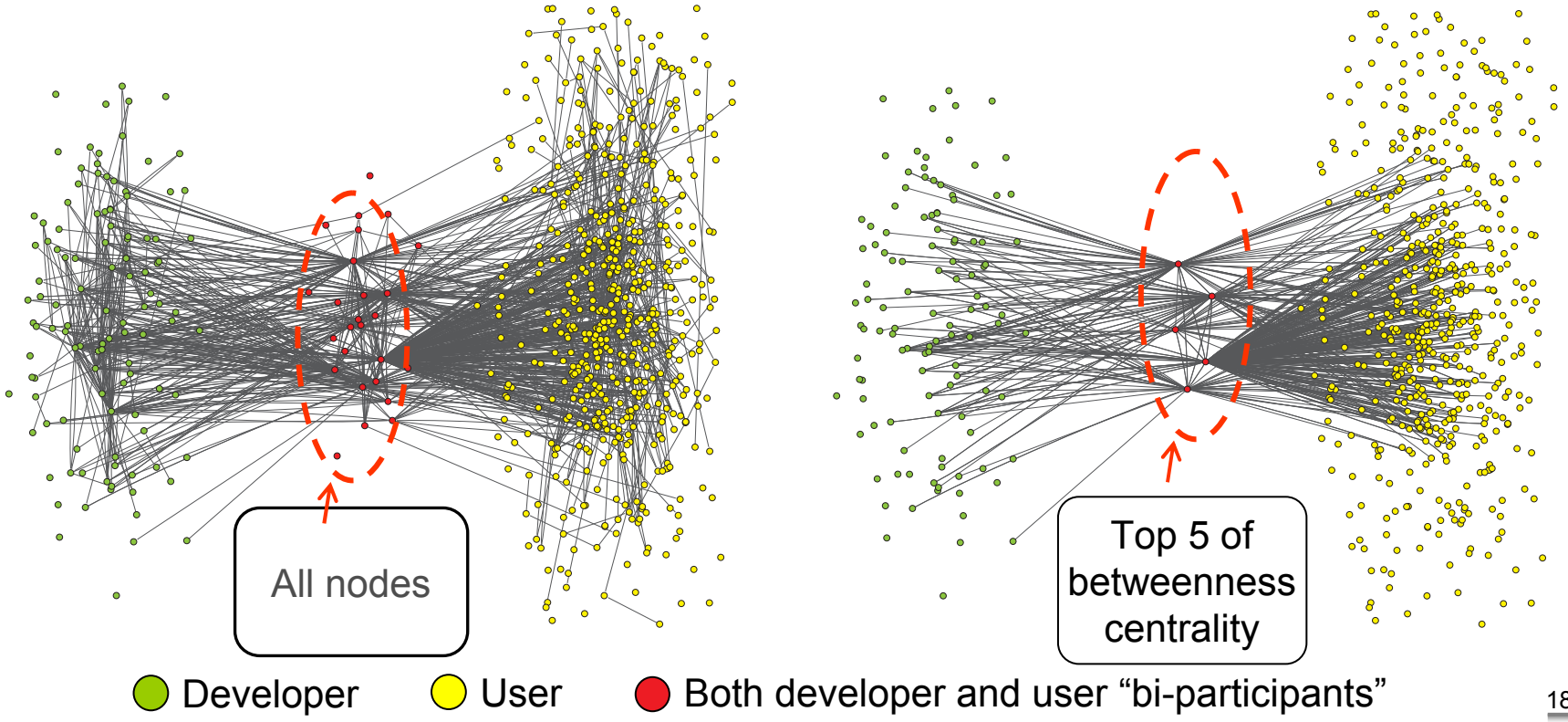
- P1 with the highest betweenness has extremely a larger number of degrees with users

The statistics value of the top 5 bi-participants

	P1	P2	P3	P4	P5	median
Betweenness	0.179	0.044	0.043	0.022	0.019	0.001
Num. of emails	592	193	261	127	62	17
Num. of degrees with dev.	15	29	33	18	11	2
Num. of degrees with users	189	31	28	19	21	1

Result of analysis -betweenness centrality-

- The top 5 bi-participants intermediate between more than half of users and developers
  - 55 of 112 nodes in the developers mailing-list
  - 249 of 540 nodes in the users mailing-list



Result of analysis -contents of messages-

- 5 experimenters checked the contents of messages which were posted by the top 5
  - If more than 3 experimenters judge that a message implied the coordination actions, we decided it as a message relates to the coordination actions

The coordination actions of the top 5 bi-participants

	P1	P2	P3	P4	P5
# of all e-mails	592	193	261	127	62
# of candidate coordination actions	15	5	23	4	22
# of coordination actions	8	1	10	3	14

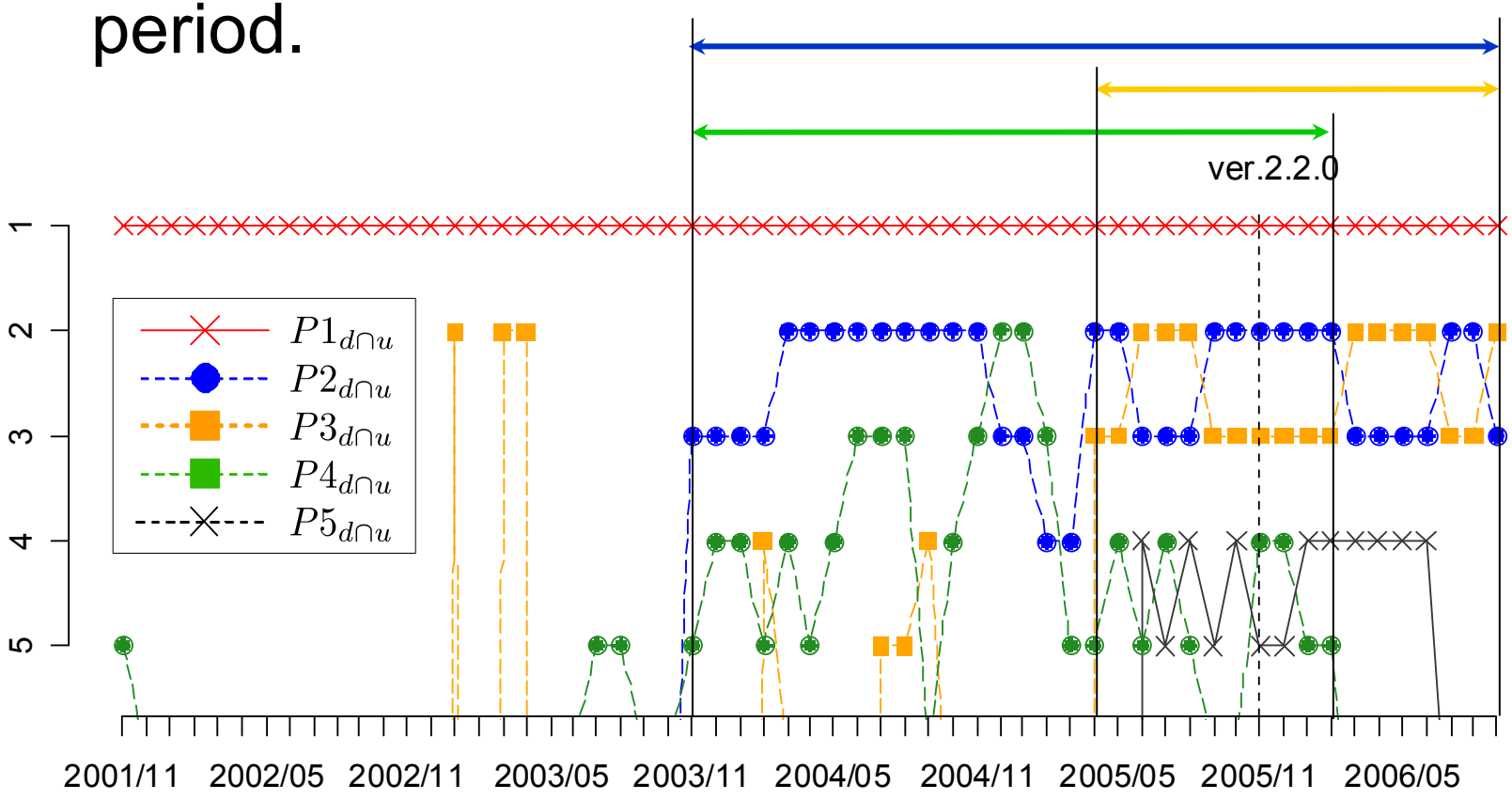
Did they coordinate activities for only a limited period of time?

- It remains unclear whether the top 5 bi-participants supported the activities of developers and users for only the analysis period or not
  - Only the data for 45 days before and after the latest major version of Apache 2.2.0 released

- We analyze the transition of the betweenness centrality of the top 5 bi-participants
  - November 2001 - September 2006

Change in the ranking of the top 5 bi-participants

- The betweenness centrality of P1 was the highest for the whole period.
- That of P2, P3 and P4 was high for the long period.



Discussion -change in the ranking of the top 5-

- Coordinators with consistently higher betweenness such as P1, P2… would continue coordinating activities between developers and users

- One of the success factors is the existence of bi-participants who facilitate and coordinate activities among the members
  - In the Linux community, Linus Torvalds, coordinator of the Linux Kernel community, has been contributing to its development since its start in 1991

Summary of this study

- The communication structure among developers and users, using two mailing-lists of developers and users in the Apache community
  - Participants with high betweenness coordinated activities between developers and users
  - Some of the participants would have been playing a role of a coordinator in the community for a long term

- Thank you for listening

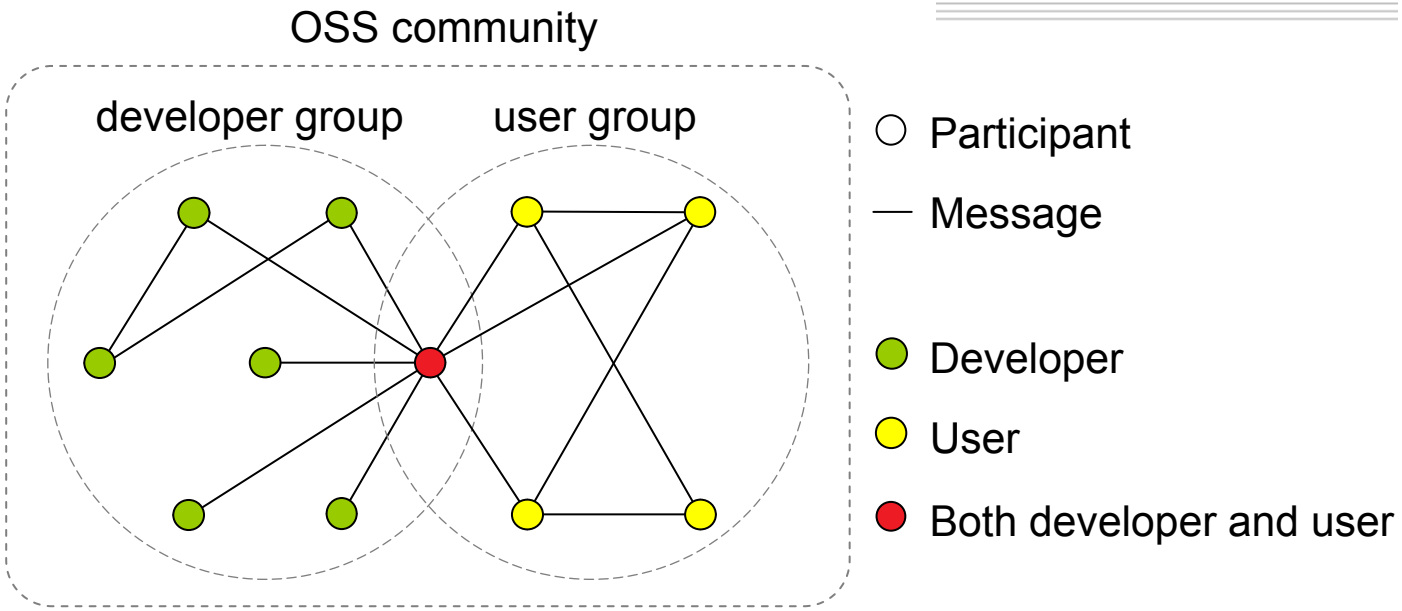
想定質問

Betweenness centrality

- It takes a value from 0 to 1, and a node with higher betweenness indicates the node which plays a role of a intermediator in the network.
  - The betweenness centrality measure is proposed by Freeman.

$$C_{betweenness}(v_i) = \frac{\sum^{n}\sum_{j<k}^{n} p_{jk}(v_i)}{\sum^{n}\sum_{j<k}^{n} p_{jk}}$$

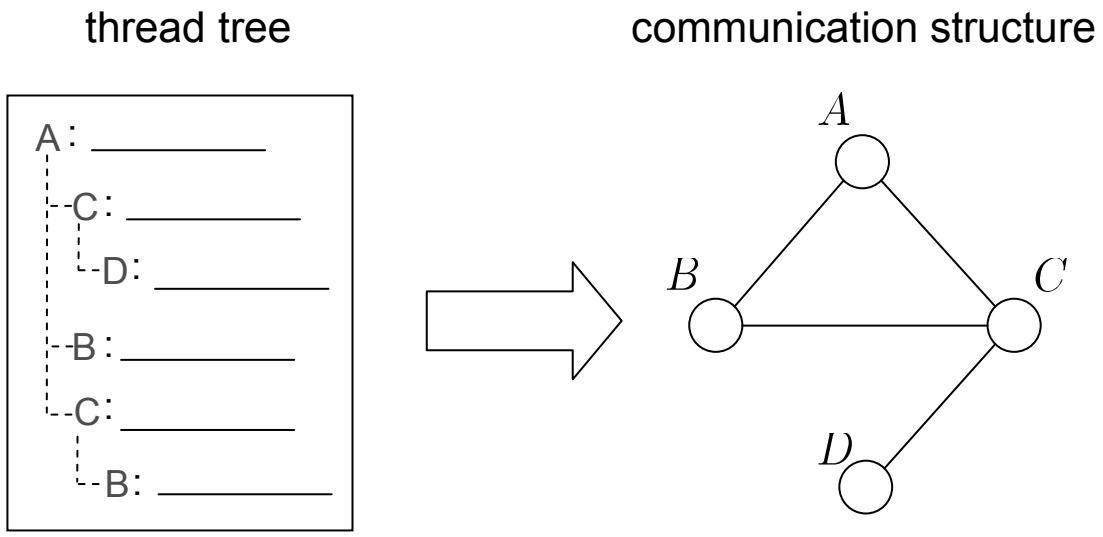
Why did we use the betweenness centrality?



- We analyze the participants which play a role of a intermediator in the network

Communication structure in online media

- We define a communication structure as sender-receiver relationships in online media
  - e.g., if participants *B* and *C* replay to a message sent by participant *A*, edges are linked from nodes *B* and *C* to node *A*



Contents of messages -Information transfer-

- If bi-participants transferred information to a developer group that only users had, we considered such action of bi-participants coordination.

Information transfer: users --> developers

sender	dialog	notes
developer	*Some versions do not seem very popular.*	The developer is suggesting in the developer group that some versions of Apache should be stopped from being made public because they are not popular.
$P3_{d \cap u}$	*Since I have personally received emails from users regarding their versions, I think the versions are still popular.*	$P3_{d \cap u}$ is telling the developers that the versions mentioned by the developers are still popular. He is motivating the developers to continue Apache development by describing its popularity among users.

Contents of messages -Request for participation-

- If bi-participants took a kind of coordinative action, which was a request for participation, we considered such action of bi-participants coordination.

Request for participation: developers --> users

sender	dialog	Notes
$P3_{d \cap u}$	*It is short of testers for some minor OS. We need your contributions as testers.*	Due to the shortage of testers for some versions of Apache compiled for some minor OS, $P3_{d \cap u}$ asks the user group to participate in the tester group. Such coordination would contribute to the development by reducing the burden of developers.

- 媒介中心性の高い参加者にコーディネーションの形跡が見られたからと言って，コミュニティにおけるコーディネーションの構造が把握できるとは言えない

- MLで，送受信関係ってｗｗｗ
  - ちゃんと送受信関係になってますのん？
  - ほかの人も見れているのとちゃいますのん？