International Seminar on Standardization of
IR Usage Statistics

On unsolved points in the methodology for IR usage analysis: some lessons learned from the ROAT project

Yoshinori Sato
Tohoku Gakuin University

Jan 11, 2011

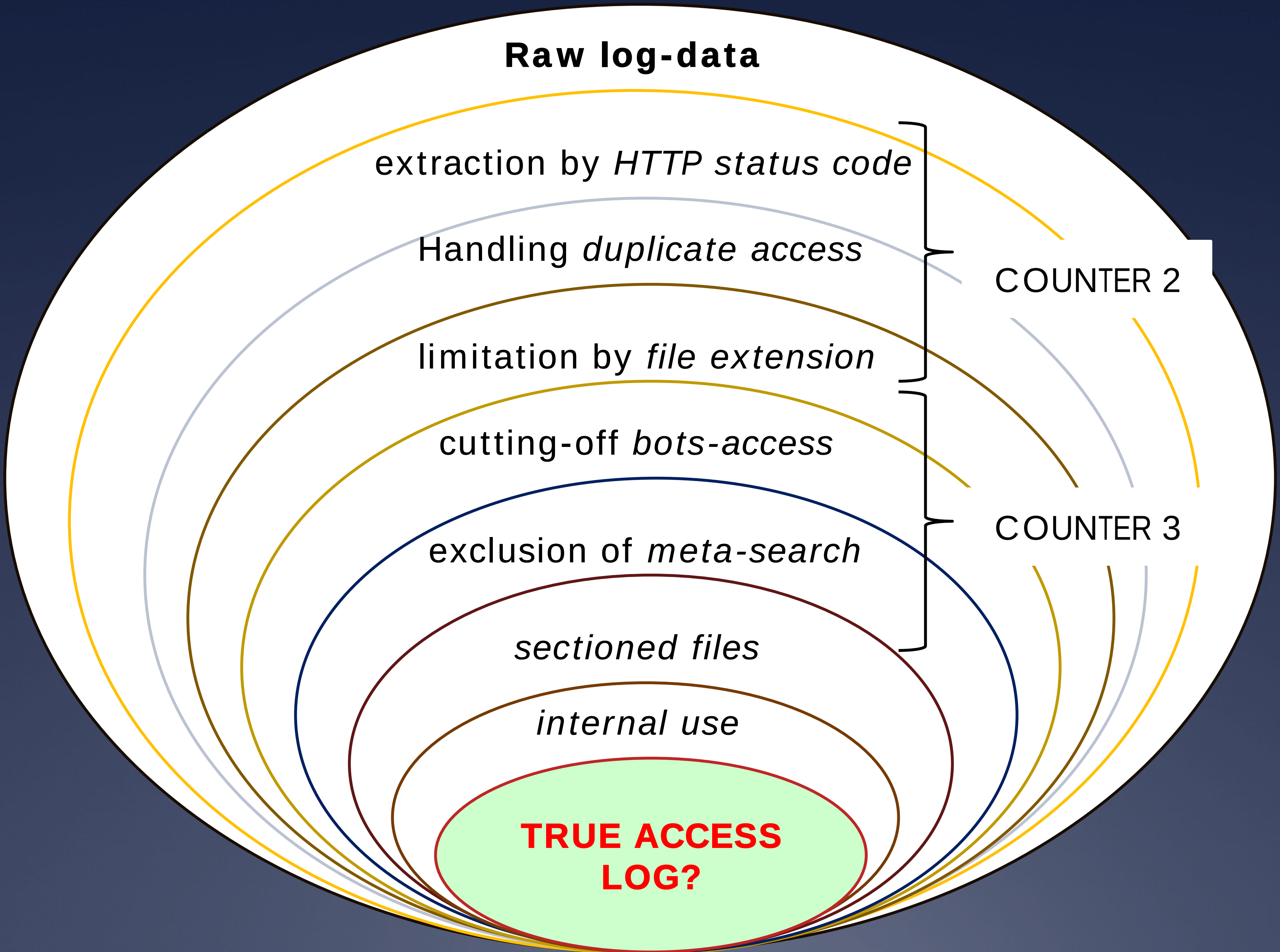
Methodological Issues

1) Reexamination of COUNTER Code of Practice
   a. Maintenance of bots database
   b. How to handle “duplicate access”

2) From page-view to session, and to user behavior

3) Author and title identification

Log-data Purification and the COUNTER



Raw log-data

extraction by *HTTP status code*

Handling *duplicate access*

limitation by *file extension*

cutting-off *bots-access*

exclusion of *meta-search*

sectioned files

internal use

TRUE ACCESS LOG?

COUNTER 2

COUNTER 3

Standardization

- *The COUNTER Code of Practice. Journals and Databases: Release 3.* 2008.8, 38 p.
- *ISO 2789 4th ed. International Library Statistics. Annex A(normative). Measuring the use of electronic library services.*

Ways of utilization and the Internet connection itself change continuously, so it is critical to keep monitoring the changes and reexamining the real data.

Observation of bots access to IRs

- Data: Access logs from15 IRs, from Jan to Dec 2008

- Approaches to detect bots-access;

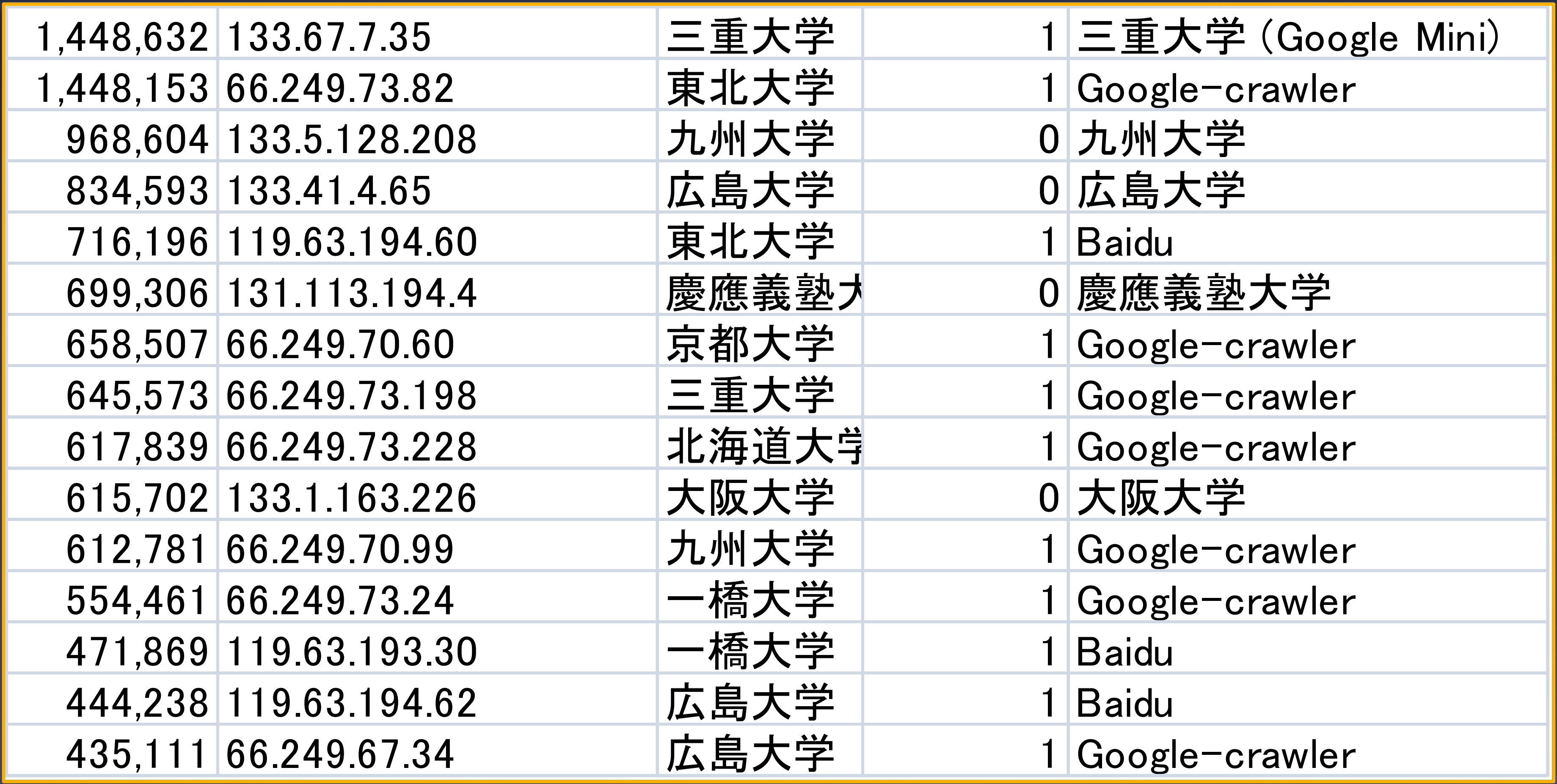
1. Counting IP address frequency in the requests for PDF files in 15 IRs

   → Listing up the frequently appearing IP address (over 800 times in each IR) and scrutinizing them

2. Counting user-agent frequency in the requests for PDF files in 15 IRs

   → Extracting unknown or newer user-agents

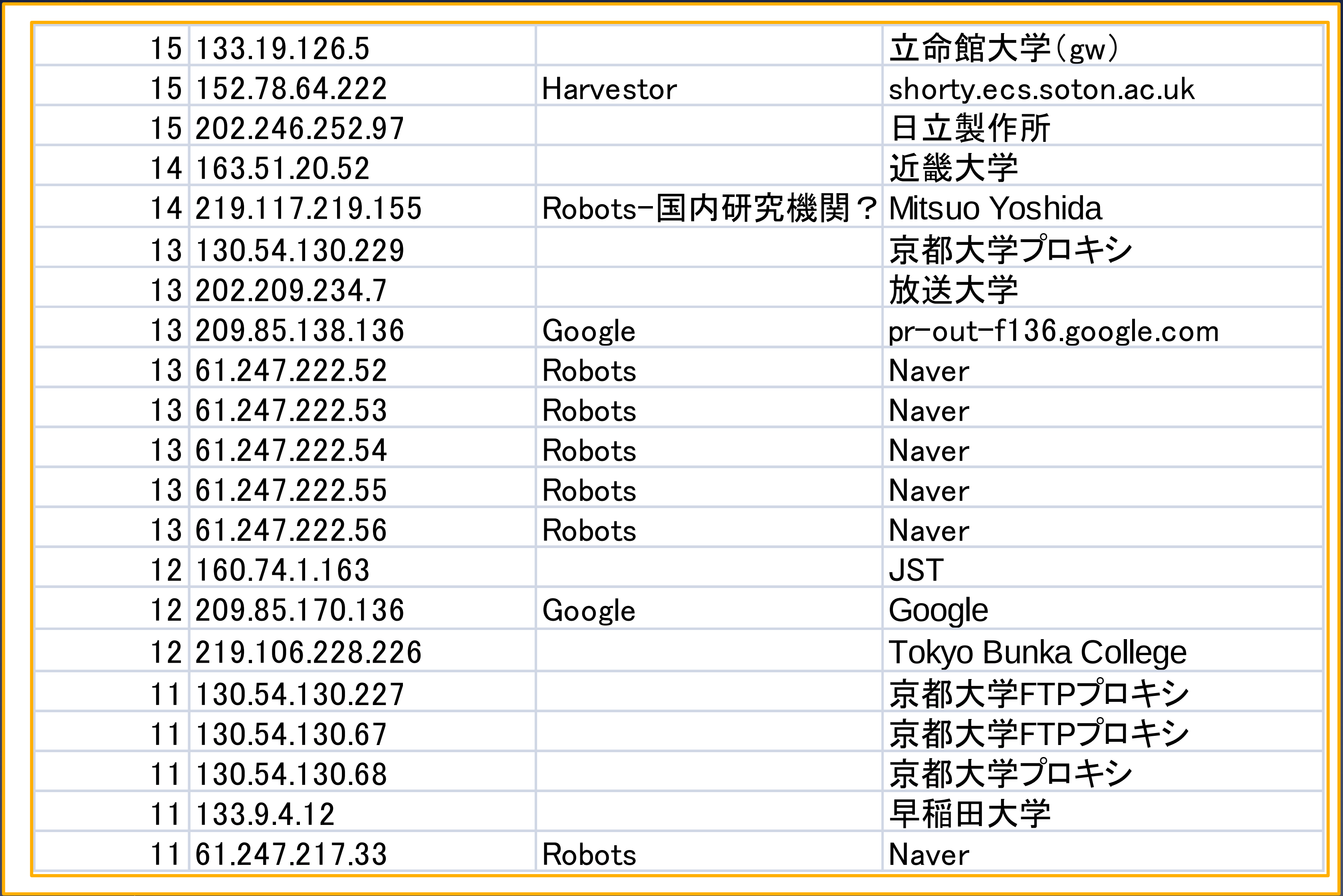
Number of access by IP address



1,448,632	133.67.7.35	三重大学	1	三重大学（Google Mini）
1,448,153	66.249.73.82	東北大学	1	Google-crawler
968,604	133.5.128.208	九州大学	0	九州大学
834,593	133.41.4.65	広島大学	0	広島大学
716,196	119.63.194.60	東北大学	1	Baidu
699,306	131.113.194.4	慶應義塾大	0	慶應義塾大学
658,507	66.249.70.60	京都大学	1	Google-crawler
645,573	66.249.73.198	三重大学	1	Google-crawler
617,839	66.249.73.228	北海道大学	1	Google-crawler
615,702	133.1.163.226	大阪大学	0	大阪大学
612,781	66.249.70.99	九州大学	1	Google-crawler
554,461	66.249.73.24	一橋大学	1	Google-crawler
471,869	119.63.193.30	一橋大学	1	Baidu
444,238	119.63.194.62	広島大学	1	Baidu
435,111	66.249.67.34	広島大学	1	Google-crawler

IPS_over800.xlsx

Redundantly appeared IP addresses in different sites



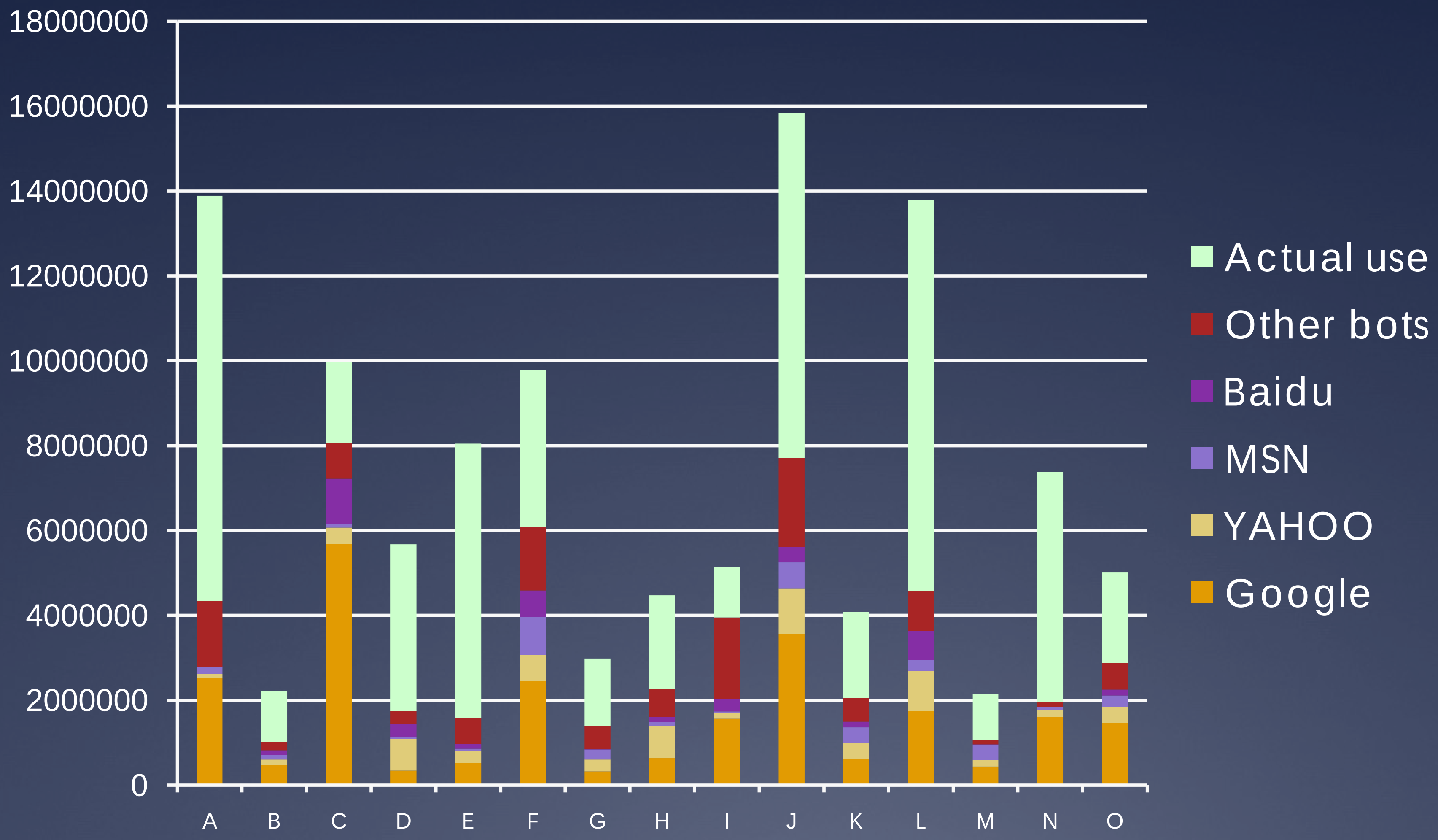
15	133.19.126.5		立命館大学(gw)
15	152.78.64.222	Harvestor	shorty.ecs.soton.ac.uk
15	202.246.252.97		日立製作所
14	163.51.20.52		近畿大学
14	219.117.219.155	Robots-国内研究機関？	Mitsuo Yoshida
13	130.54.130.229		京都大学プロキシ
13	202.209.234.7		放送大学
13	209.85.138.136	Google	pr-out-f136.google.com
13	61.247.222.52	Robots	Naver
13	61.247.222.53	Robots	Naver
13	61.247.222.54	Robots	Naver
13	61.247.222.55	Robots	Naver
13	61.247.222.56	Robots	Naver
12	160.74.1.163		JST
12	209.85.170.136	Google	Google
12	219.106.228.226		Tokyo Bunka College
11	130.54.130.227		京都大学FTPプロキシ
11	130.54.130.67		京都大学FTPプロキシ
11	130.54.130.68		京都大学プロキシ
11	133.9.4.12		早稲田大学
11	61.247.217.33	Robots	Naver

ファイル： IP_temp3.xlsx

IP address appearance in different sites

- Only 3 IPs appeared at all of 15 sites
- The number of IPs observed at more than 8 sites – 87 IPs
- Search engines used numerous IPs, many of which did not appear commonly
  - Yahoo – 532 IPs
  - Google – 466 IPs
  - MSN – 421 IPs

Impacts of crawlers, robots, etc. on 15 IRs



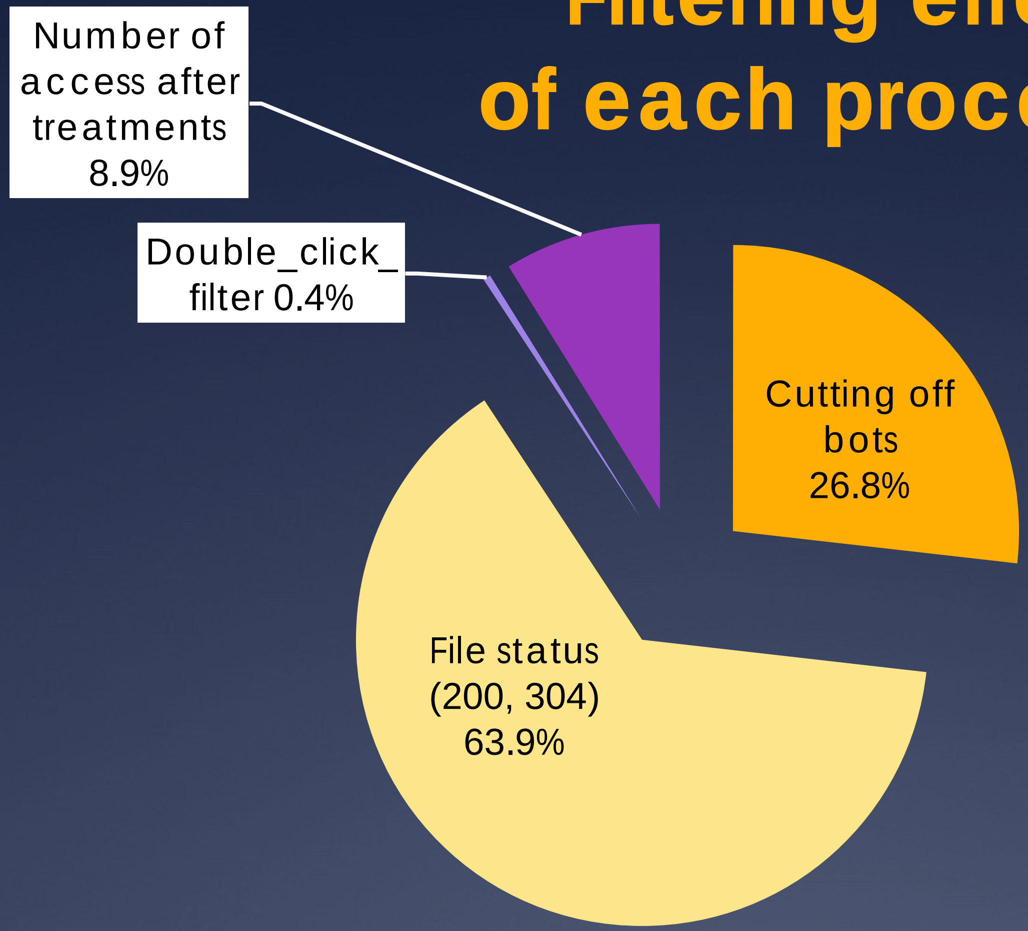
Validation of filtering effects

- Data: Access log in Chiba University's Curator, from Jan to Mar, 2010

- Procedure 1:
  1. Extracting only 'access to PDF files'
  2. Running filtering programs in following sequence
     i. Cutting off bots' access
     ii. Extracting by file status (200, 304)
     iii. Handling duplicate access
  3. Comparing the number of records before and after the treatments

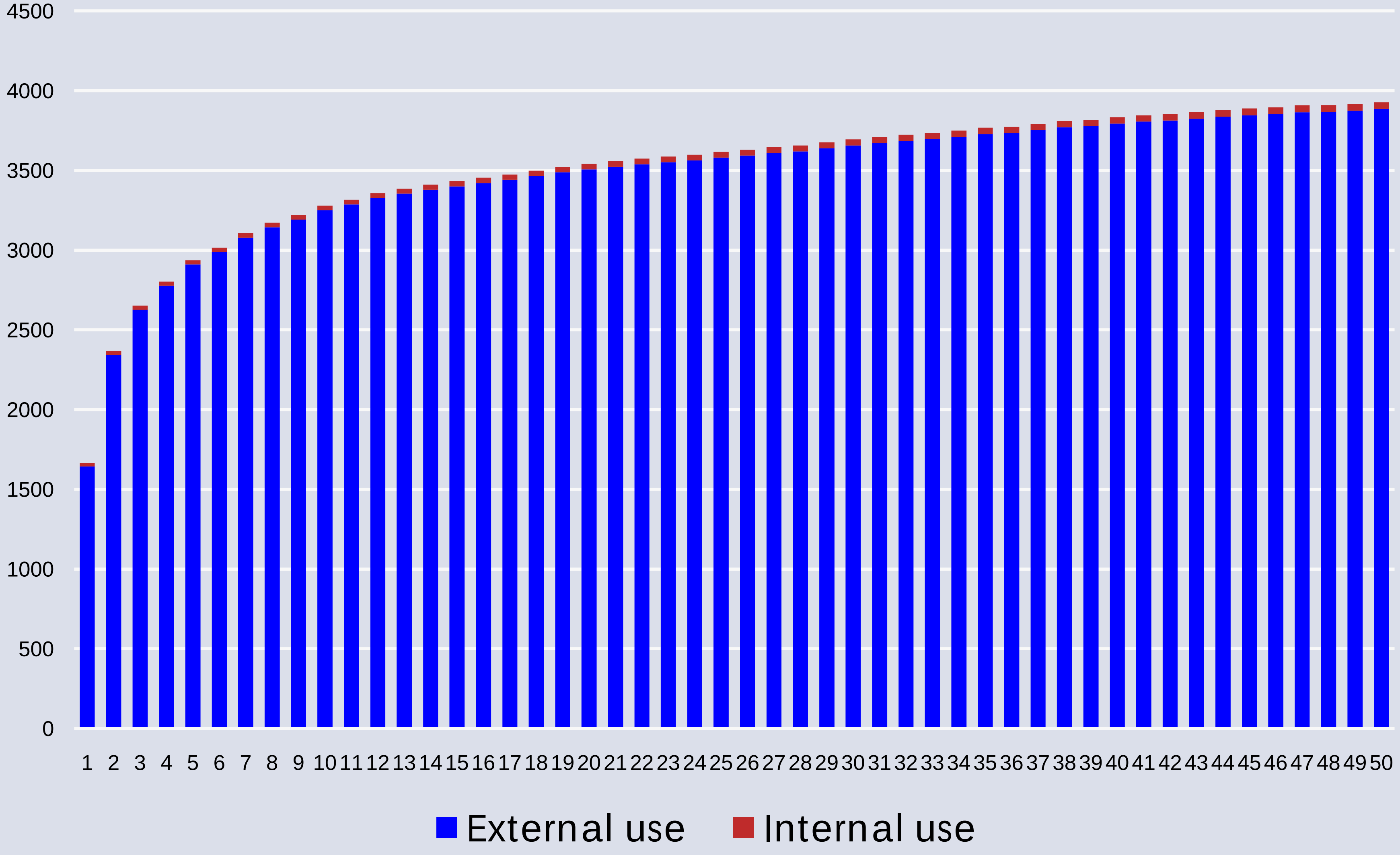
- Procedure 2:
  1. Monitoring the effect of double-click-filter by varying the range from 0 to 50 seconds

Filtering effect of each procedure



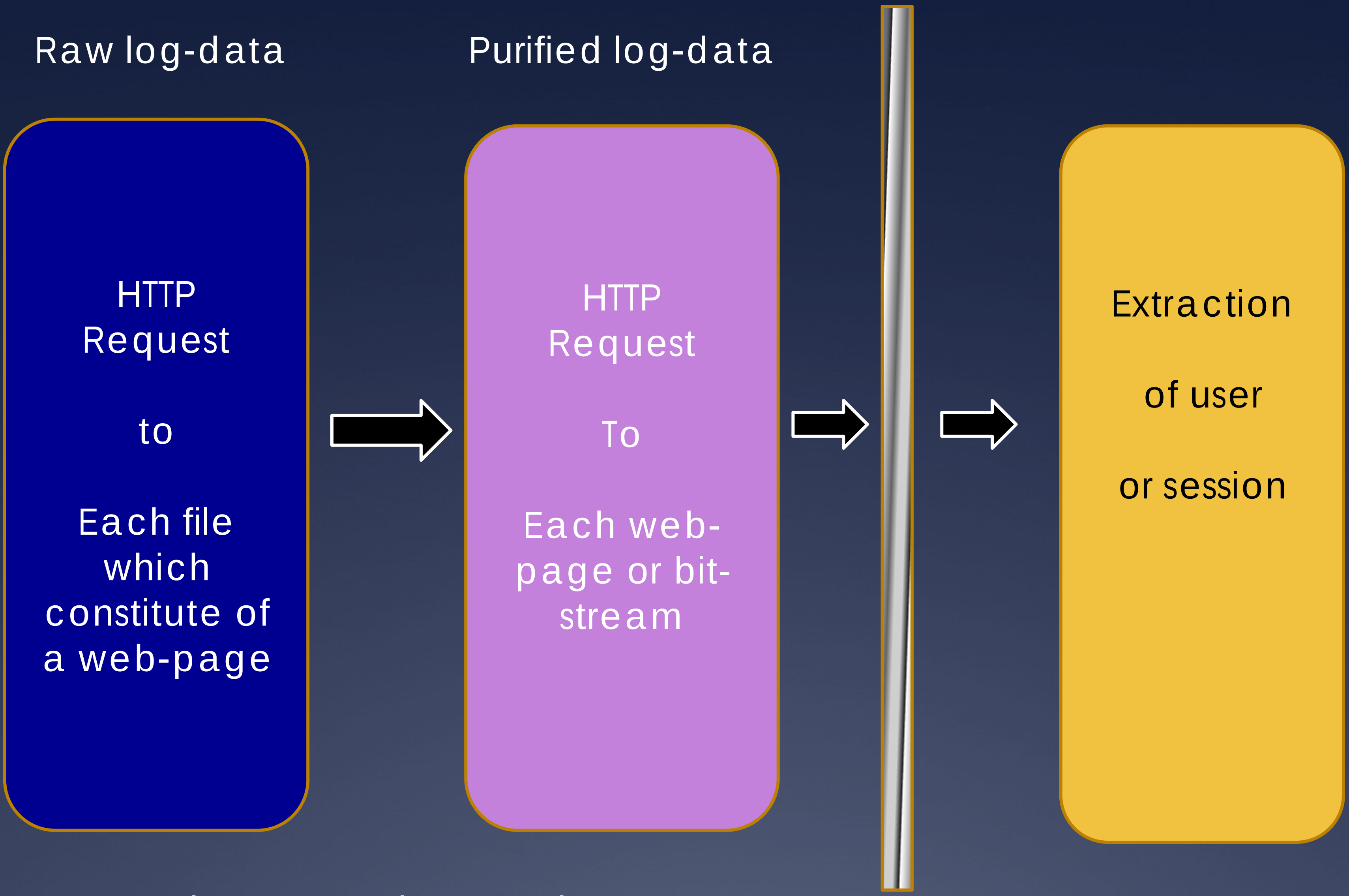
	Number of access	Number of access rejected	Rejection ratio
Pretreatment (only for PDF files)	872,956	–	–
Cutting off bots	638,900	234,056	26.8%
File status (200, 304)	80,967	557,933	63.9%
Double_click_filter	77,270	3,697	0.4%

Double-click-filtering within the number of seconds



From page-view to session, and to user behavior

- To what extent can we understand “user behavior” in IRs, by seeing not merely the number of hits and/or page-views?



Raw log-data

HTTP Request

to

Each file which constitute of a web-page

Purified log-data

HTTP Request

To

Each web-page or bit-stream

Extraction

of user

or session

issues:

What procedures and standardization do we need for inter-operable and inter-comparable statistics?

What conditions are required for counting users and sessions?

Research themes

- Empirical examination of session identification which is required for counting the number of users

  → Comparing the estimated numbers of users/sessions between cookie setting and non-cookie (IP only) setting

  → Consideration of time-out value in IR setting

Adoption of cookie and its efficacy

- User identification in Counter Code of Practice
  - IP address, cookie, and user-account are usually adopted
- Comparison of the methods for user identification

Identification approach	Target	Workload	Accuracy	Flexibility
IP address	Terminal machine	Small	Low	○
Cookie	Browser	Medium	Medium	○
User account	User	Big	High	×

※Where more than one user use a terminal continuously, e.g. in libraries, user identification can be accomplished only by adopting individual user account. A terminal in a library, however, may be determined by its access pattern left on the log-file.

Checking cookie's efficacy

- Access log-file in Chiba Univ's Curator
  - From Mar 3 to Apr 14, 2010
  - Cookie's expiration date: 1 month (30 days)
  - (for reference): session cookie – expires when the session is closed; user cookie - expires when the validated period is over
- Pretreatment: data purification
  - Elimination of the access by bots
  - Removal of the records which don't have HTTP status code 200 or 304

Acceptance rate of cookie

- Percentage of total requests having cookie

  77.6 % = 215,105 / 277,093

Correspondence between Cookie and IP address					
with Cookie	7,264	100%	Cookie:IP address = 1:1	5,087	70%
			Cookie:IP address = n:1 (n>1)	1,288	18%
			Cookie:IP address = 1:n (n>1)	456	6%
			Cookie:IP address = n:m (n>1, m>1)	433	6%

Cookie : IP address = n:1 (n>1)

- An IP corresponds to multiple cookies
  - Up to 24 cookies
- Possible situations;
  - Where different terminals use the same (global) IP address, e.g. Proxy server – NAT (Network address translation)
    - Identification by cookie is effective
    - Identification by IP address will lead to underestimation
  - Where a user removes the cookie intentionally after every session

Cookie : IP address = 1:n (n>1)

- A cookie corresponds to multiple IPs
  - Up to 66 IPs
- Possible situations;
  - Where ISP configures IP address dynamically in each session
    - Identification by cookie is effective
    - Identification by IP address will lead to overestimation
  - Where a user (browser) access more than once beyond the expiration date of 1 month

Cookie : IP address = n : m (n, m>1)

Multiple cookies correspond to multiple IPs

→ Combination of two phenomena mentioned above

Estimation of time-out duration 1

- All requests with cookie -

Time interval from previous request	Frequency	Cumulated frequency	Cumulated ratio	Rate of increase
within 5 minutes	20,693	20,693	83.79%	
within 10 minutes	518	21,211	85.89%	0.10%
within 15 minutes	230	21,441	86.82%	0.06%
within 20 minutes	147	21,588	87.41%	0.04%
within 25 minutes	80	21,668	87.74%	0.04%
within 30 minutes	101	21,769	88.15%	0.06%
within 35 minutes	136	21,905	88.70%	0.03%
within 40 minutes	55	21,960	88.92%	0.03%
within 45 minutes	45	22,005	89.10%	0.01%
within 50 minutes	27	22,032	89.21%	0.01%
within 55 minutes	25	22,057	89.31%	0.02%
within 1 hour	32	22,089	89.44%	0.01%
Over 1 hour	2,607	24,696	100.00%	10.56%

Maximum duration: 718h 40m 50s

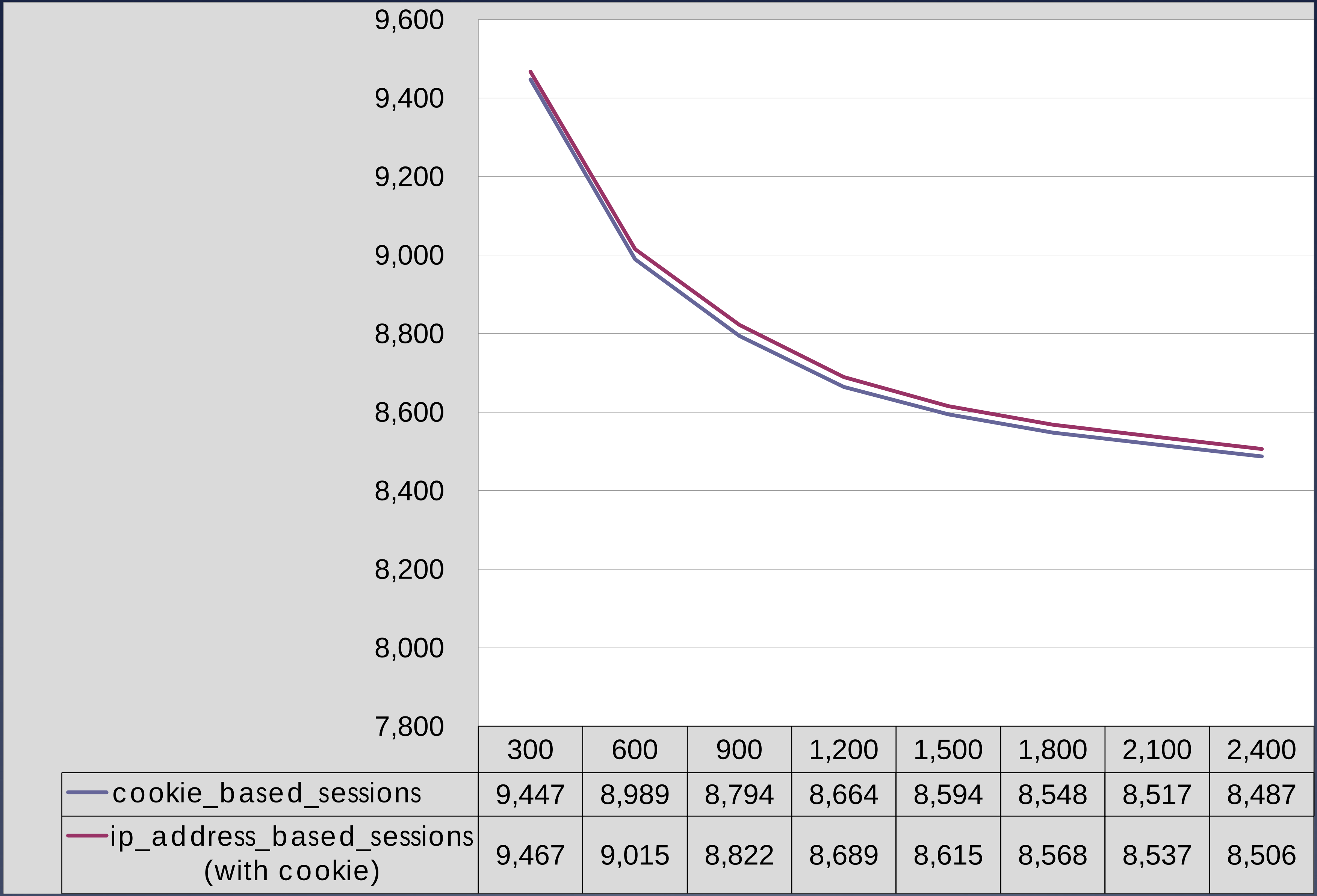
Estimation of time-out duration 2

- cases where preceding request is to PDF -

Time interval from previous request	Frequency	Cumulated frequency	Cumulated ratio	Rate of increase
within 5 minutes	2,835	2,835	66.72%	66.72%
within 10 minutes	217	3,052	71.83%	5.11%
within 15 minutes	103	3,155	74.25%	2.42%
within 20 minutes	79	3,234	76.11%	1.86%
within 25 minutes	40	3,274	77.05%	0.94%
within 30 minutes	22	3,296	77.57%	0.52%
within 35 minutes	17	3,313	77.97%	0.40%
within 40 minutes	17	3,330	78.37%	0.40%
within 45 minutes	16	3,346	78.75%	0.38%
within 50 minutes	17	3,363	79.15%	0.40%
within 55 minutes	7	3,370	79.31%	0.16%
within 1 hour	10	3,380	79.55%	0.24%
Over 1 hour	869	4,249	100.00%	20.45%

Maximum duration: 706h 9m 3s

Estimated number of sessions: Cookie-based vs. IP-based



	300	600	900	1,200	1,500	1,800	2,100	2,400
cookie_based_sessions	9,447	8,989	8,794	8,664	8,594	8,548	8,517	8,487
ip_address_based_sessions (with cookie)	9,467	9,015	8,822	8,689	8,615	8,568	8,537	8,506

3. Author and title identification

- Using URL for matching the info with meta-data (from JAIRO) at this moment
- However, it is problematic because of the instability of URL, even in the case where a system use a persistent mechanism (e.g. handle)
- Dynamic URL also raises a difficult problem
- A possible solution is to introduce “Object Identifiers” for Author and Title (work/expression/manifestation/item?) to JAIRO.

Conclusion

- On the filters defined in *Counter Code of Practice* 3rd ed.
  - To eliminate bots’ access;
    - Collective effort is crucial, as a variety of unknown robots, crawlers, spams etc. is being created and it is hard to detect newer ones
  - About cutting off “Duplicate access”
    - The procedure may be dispensable
- On user/session identification
  - Implementation of cookie is effective, but further investigation is necessary

Next tasks

- Clarification of peculiar patterns, looking at a terminal in library and proxy server
    - Paying attention to access frequency by the cookie

- In-depth understanding of the inflation/deflation of unbalanced cookie and IPs with a larger sample

⇒ Estimation of the number of users/sessions even when only IPs are available

Thank you for your attention!