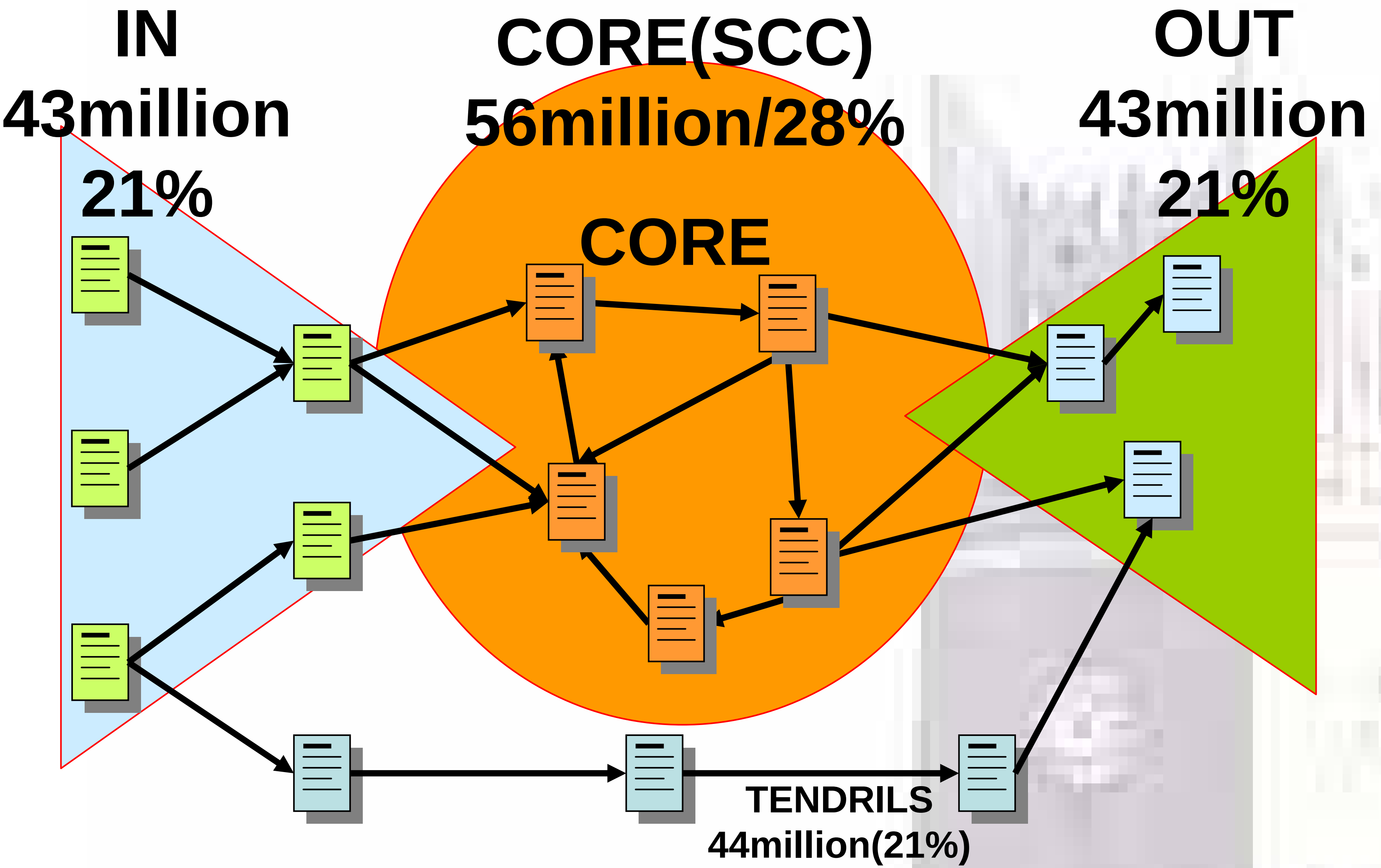
Web Structure in 2005

Yu Hirate and Hayato Yamana

Computer Science Div. Science and Engineering,
Waseda University, Japan

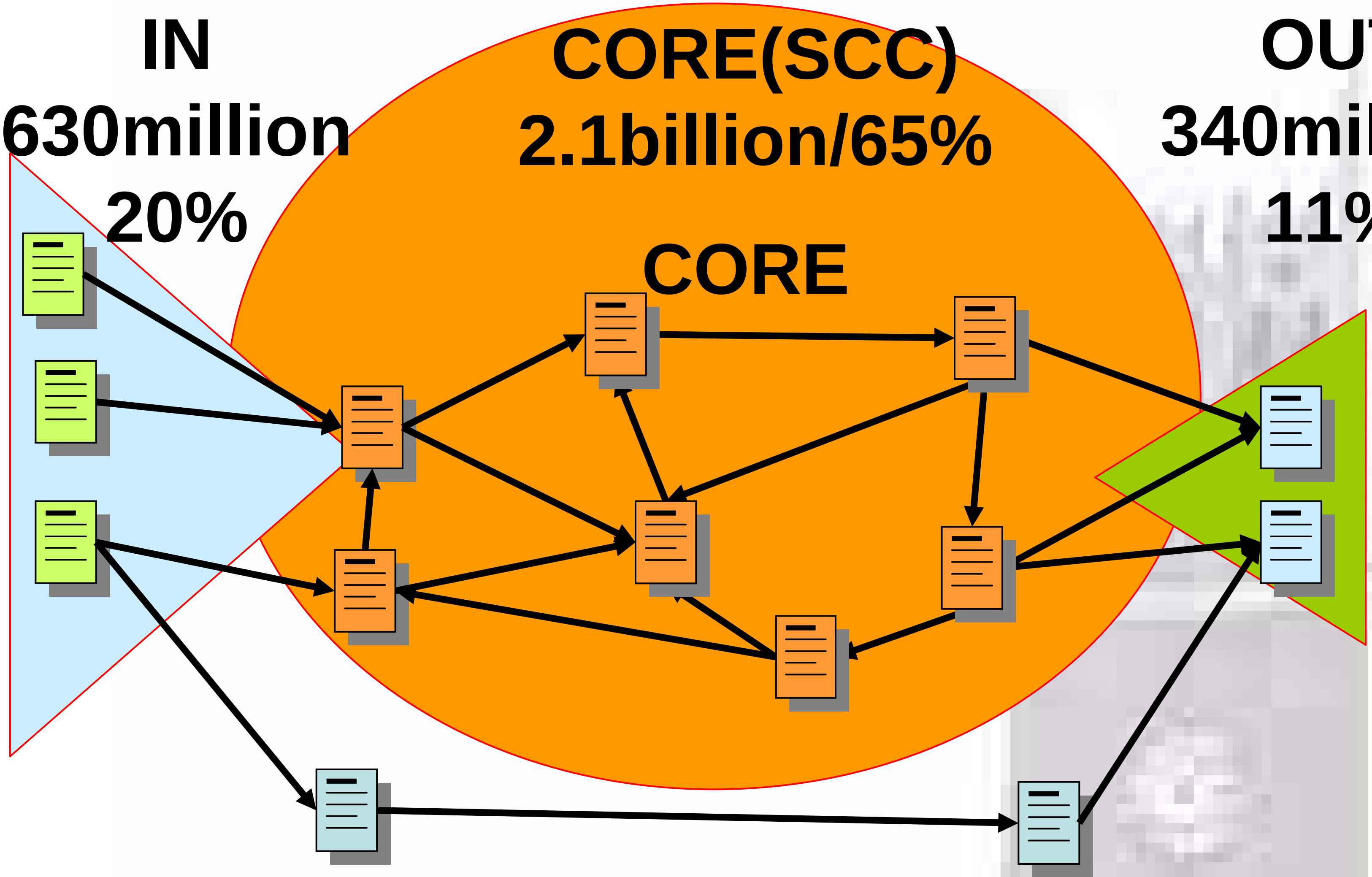
{hirate,yamana}@yama.info.waseda.ac.jp

Web Structure in 1999 [Broder99]



Based on 200 million web pages gathered in 1999

Web Structure in 2005 (Our Investigation)



IN
630million
20%

CORE(SCC)
2.1billion/65%

OUT
340million
11%

TENDRILS
120 million(4%)

Based on 3.2 billion web pages

Agenda

PART1 Background of our research

- The Japanese government founded project : The e-Society project

PART2 How many web pages are there?

PART3 The web structure

- Related works
- The web structure in 2005

PART4 Conclusion

Part1: The e-Society Project

Gathering 14billion Web Pages and
Discovering New Knowledge

The e-Society Project(2003.4-2008.3)

http://www.yama.info.waseda.ac.jp/~yamana/e-society/index_eng.htm

- Founded by the Ministry of Education, Culture, Sports, Science and Technology, JAPAN
- Contractor: Waseda University
- Goal
  - **To realize the largest Web repository in the world**
    - Gathering web pages from all over the world (over 14 billion pages)
    - Keeping up with the average age of the gathered Web pages smaller than 1 month
  - **To realize distributed Web Mining Scheme**
    - Finding out the useful information such as hidden communities in the Web
    - Proposing a new mining concept for the Web data

Systems for the e-Society Project

Web page crawling system (1 location)

Data analyzing system (PC Cluster)

We have 4 crawling locations

-2 locations in Waseda University

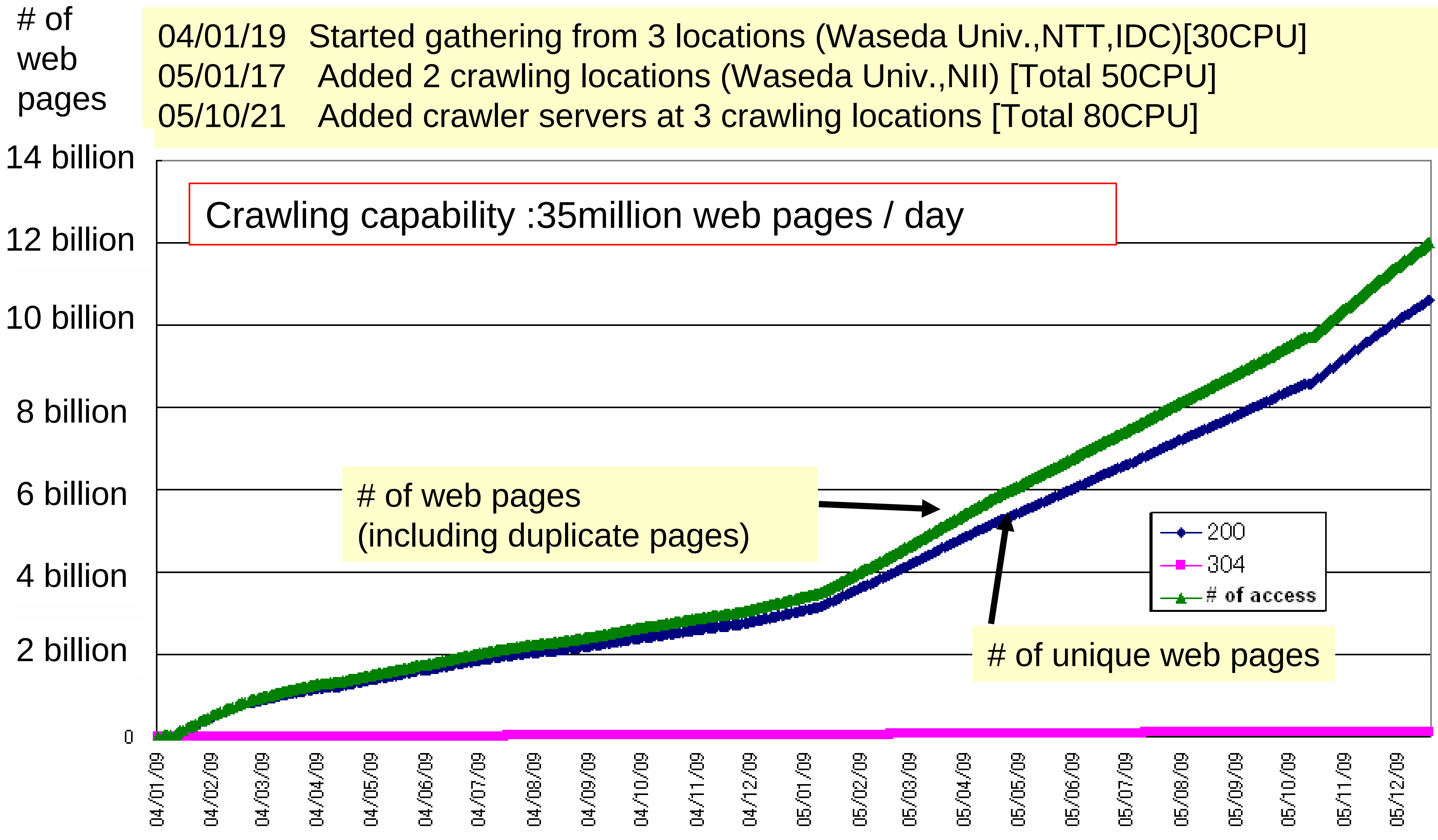
-1 location in NII (National Institute of Informatics)

-1 location in IDC Data Center

128 nodes (Pentium 4 2.4GHz, 1GB Memory, 800GB HDD)

Status of gathering web pages

04/01/19 Started gathering from 3 locations (Waseda Univ.,NTT,IDC)[30CPU]
05/01/17 Added 2 crawling locations (Waseda Univ.,NII) [Total 50CPU]
05/10/21 Added crawler servers at 3 crawling locations [Total 80CPU]



Crawling capability :35million web pages / day

# of web pages
(including duplicate pages)

# of unique web pages

We have gathered 14 billion web pages. (Oct. 2006)

Part2: How many web pages are there?

How many web pages are there?

The average # of web pages per one web site:

- 190 pages[1]
- 186 pages[2]
- 202 pages[3]

# of web sites of all over the world : **101,435,253** servers [4]

(as of Nov. 2006)

The estimated # of web pages of all over the world:

20.3 billion web pages

[1] S.Lawrence, C.L.Giles:"Searching the World Wide Web", Science, Vol.280, No.5360, pp.98-100 ,1998.

[2] S.Lawrence, C.L.Giles:"Accessibility of Information on the Web", Nature, Vol.400, pp.107-109, 1999.

[3] Institute for Information and Communications Policy:
Statistics Investigation Report for contents on the World-Wide Web,
http://www.soumu.go.jp/iicp/chousakenkyu/seika/houkoku.html 2004.

[4] Netcraft November 2006 Web Server Survay,
http://news.netcraft.com/archives/2006/11/01/november_2006_web_server_survey.html

How many Web Pages are there? (Our investigation)

We have gathered 8,507,237,370 web pages from 16,035,801 servers. (as of Oct. 2005)

The average # of web pages / server : $\dfrac{8{,}507{,}237{,}370}{16{,}035{,}801} \cong 530$

# of web servers of all over the world : **101,435,253** servers [4]

(as of Nov. 2006)

↓

The estimated # of web pages of all over the world:

53.7 billion web pages

[4] Netcraft November 2006 Web Server Survay, http://news.netcraft.com/archives/2006/11/01/november_2006_web_server_survey.html

Why the difference was occurred ?

20.3 billion pages << 53.7 billion pages

Estimated by
Conventional Research

Estimated by
Our Investigation

Increasing dynamic web pages

- CGI pages based on Databases
- Blogs
- Portal Sites
- EC Sites

How many web pages does google indexed?

- Possible to estimate by requesting “site: ******”

  (e.g.) “site:.net”

TLD	Percentage of dominance	# of pages (result)	description
net	42.3%	1,170,000,000	Networks
com	17.5%	14,040,000,000	Commercial
jp	6.4%	986,000,000	Japan
it	3.0%	349,000,000	Italy
de	2.7%	1,130,000,000	Germany
edu	2.3%	2,820,000,000	Educational
fr	2.1%	472,000,000	France
Total	**99.6%**	**34,711,000,000**	

Google Indexes approximately 34,7 billion web pages

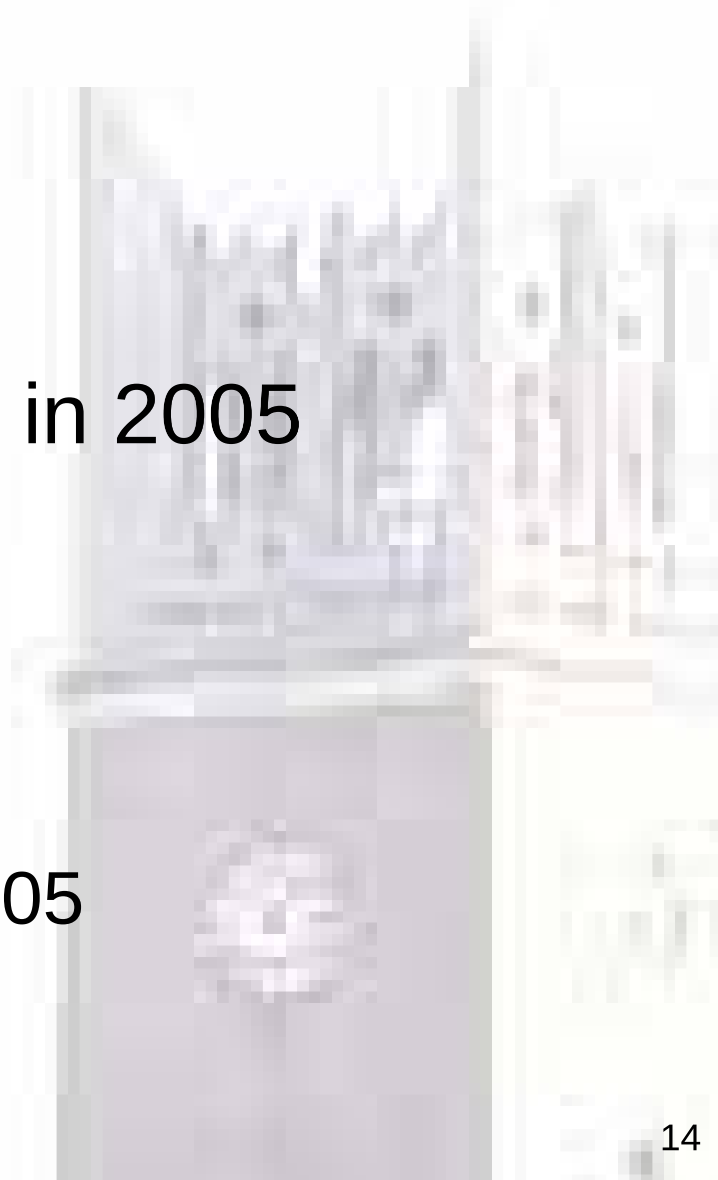
Our estimation : 53.7 billion web pages

Part3: Web structure in 2005

(1) Related works

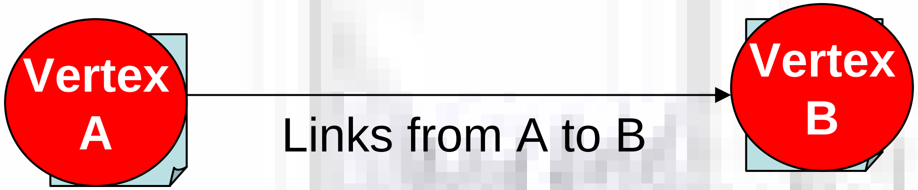
(2) Our approach

(3) The web structure in 2005



Graph Structure in the Web[Broder99]

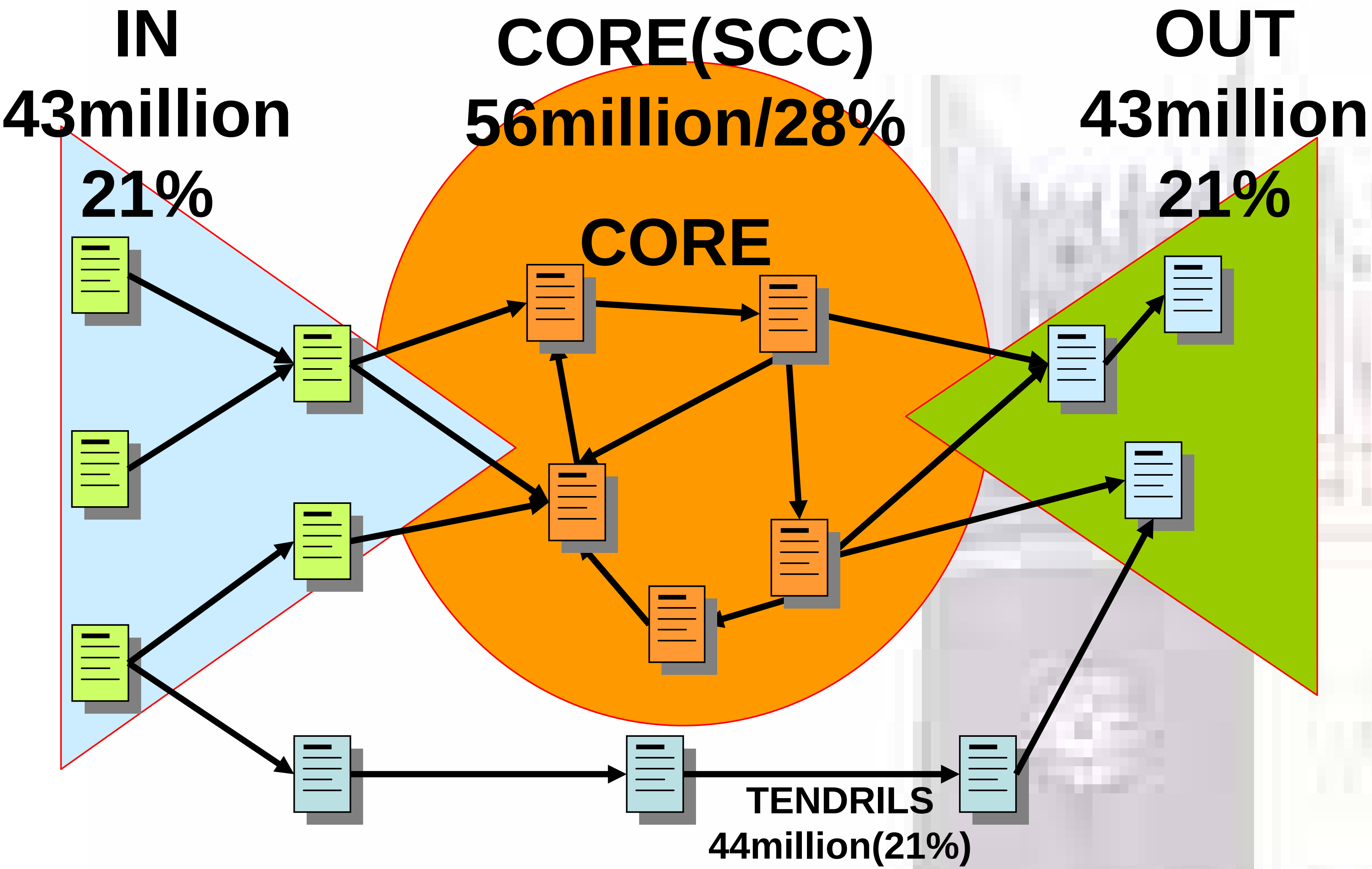
- Consider web as directed graph structure
  - Web pages = Nodes
  - Web Links = Edges



Pages are categorized in 5 groups

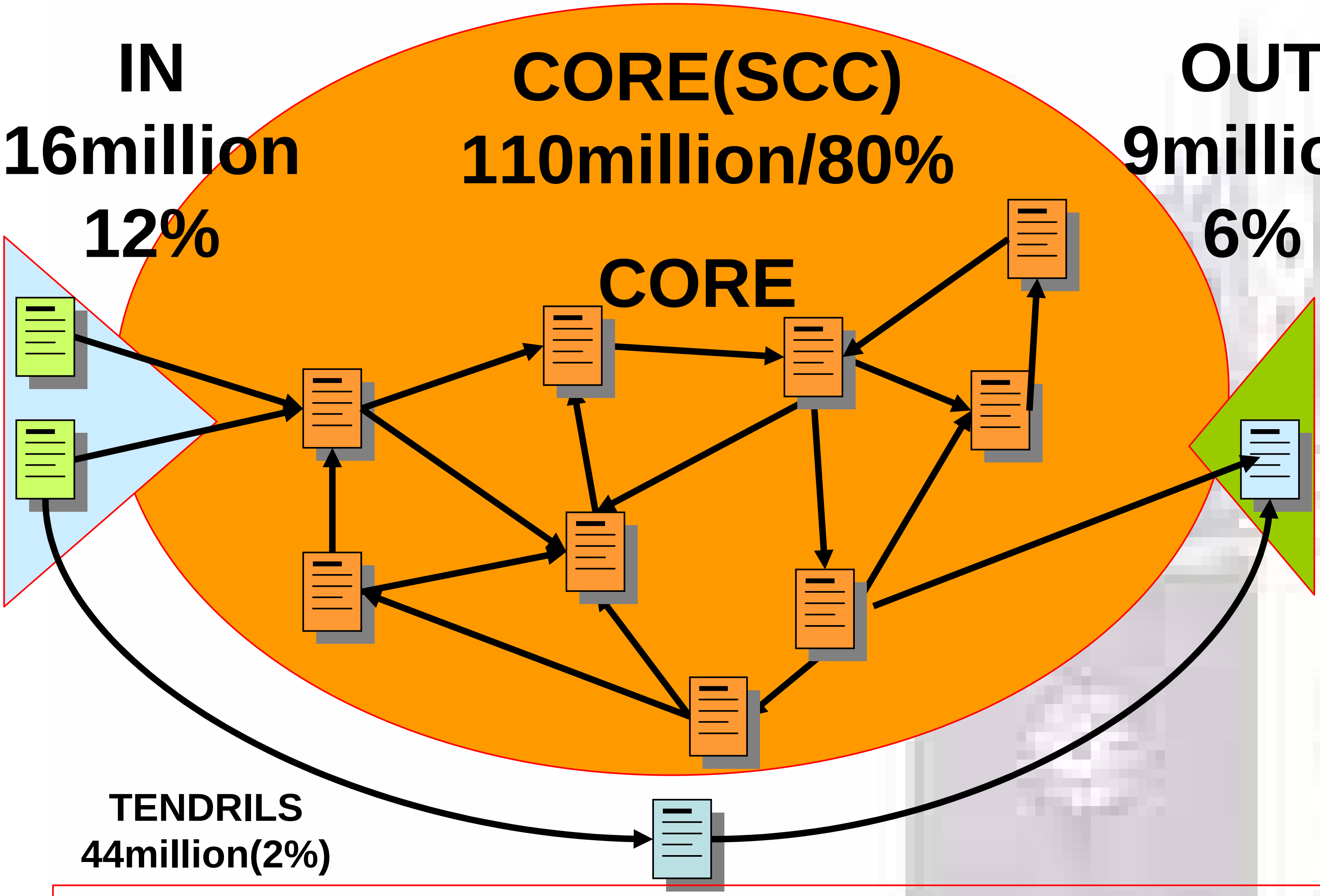
CORE	Pages included in SCC(=Strongly Connected Component)
IN	Pages that have links to CORE, but do not have Links from CORE
OUT	Pages that have links from CORE, but do not have Links to CORE
Tendrils	Pages included in Path from IN to OUT
Disconnected Components	Pages that do not have links to either CORE, IN, OUT, and Tendrils.

Web Structure in 1999 [Broder99]



Based on 200 million web pages gathered in 1999

China Web Structure in 2003[Lie05]



Based on 140 million web pages gathered in 2003

Constructing Web Structure in 2005

Target web pages :

3.2 billion web pages gathered by e-society project

Generated 3 types of web structure

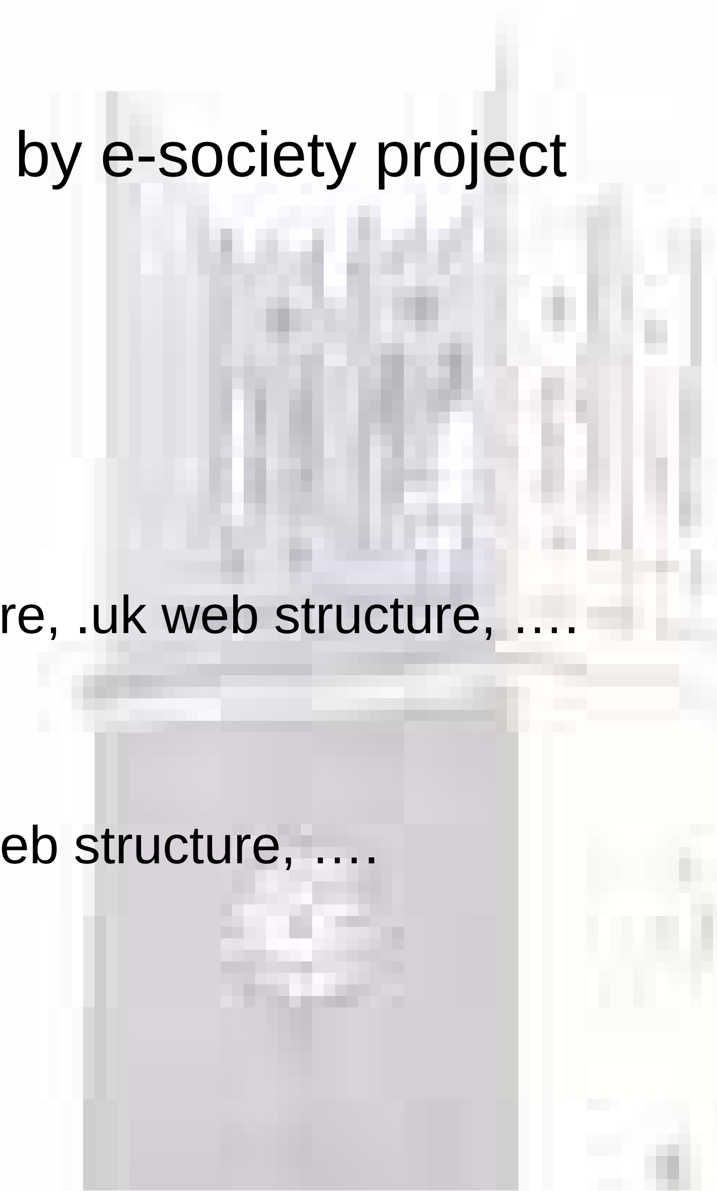
(1) whole

(2) by TLD (=Top Level Domain)

(e.g.) .com web structure, .jp web structure, .uk web structure, ….

(3) by written Language

(e.g.) English web structure, Japanese web structure, ….

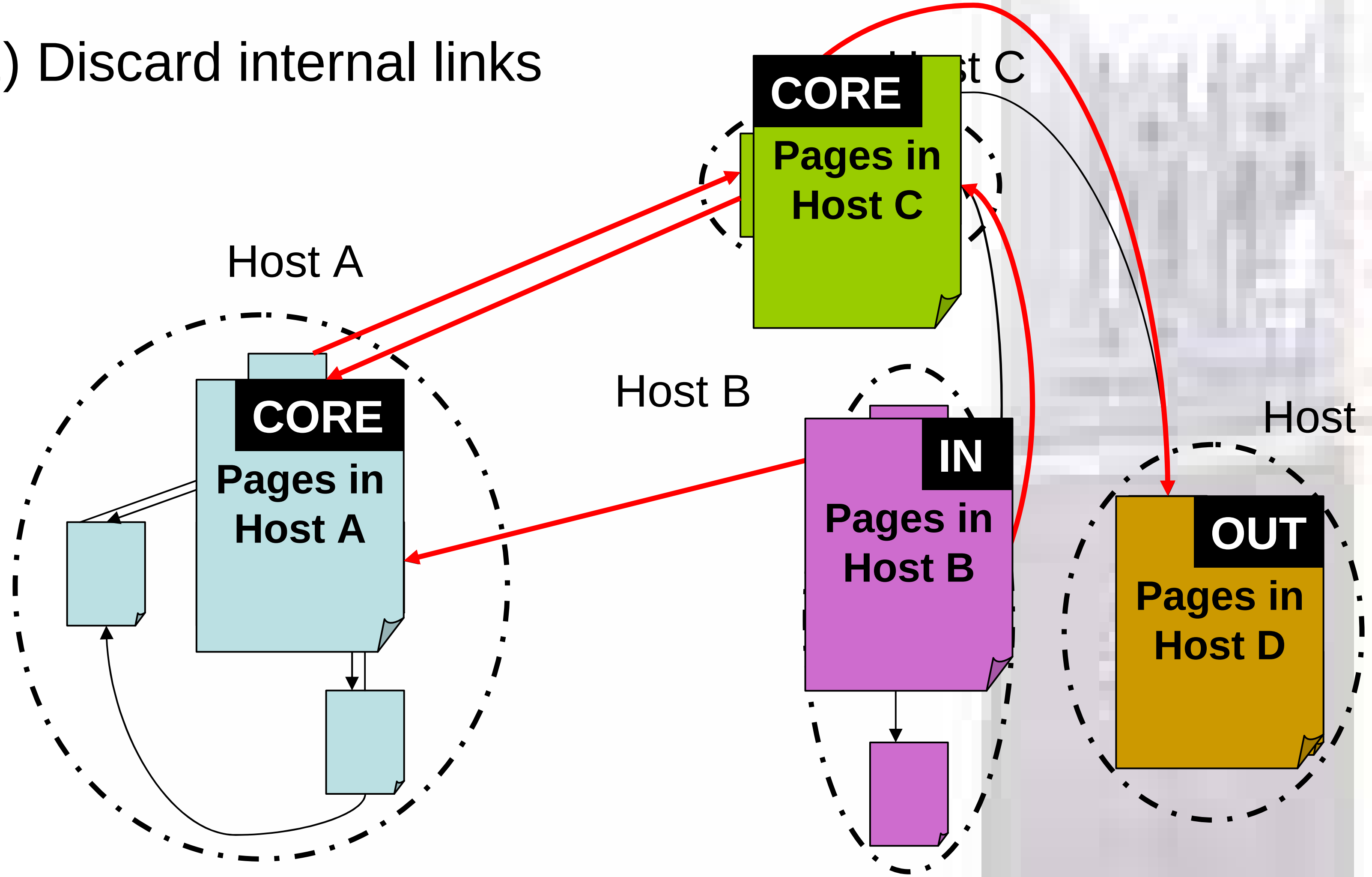


Our Approach

Host Level Reduction:

(1) Consider web pages in the same host as one node.

(2) Discard internal links



Dataset Property

web pages : 3,208,139,905 pages

hyper-links : 93,397,065,743 links

- one web page has 58 links in average

↓ Host level reduction

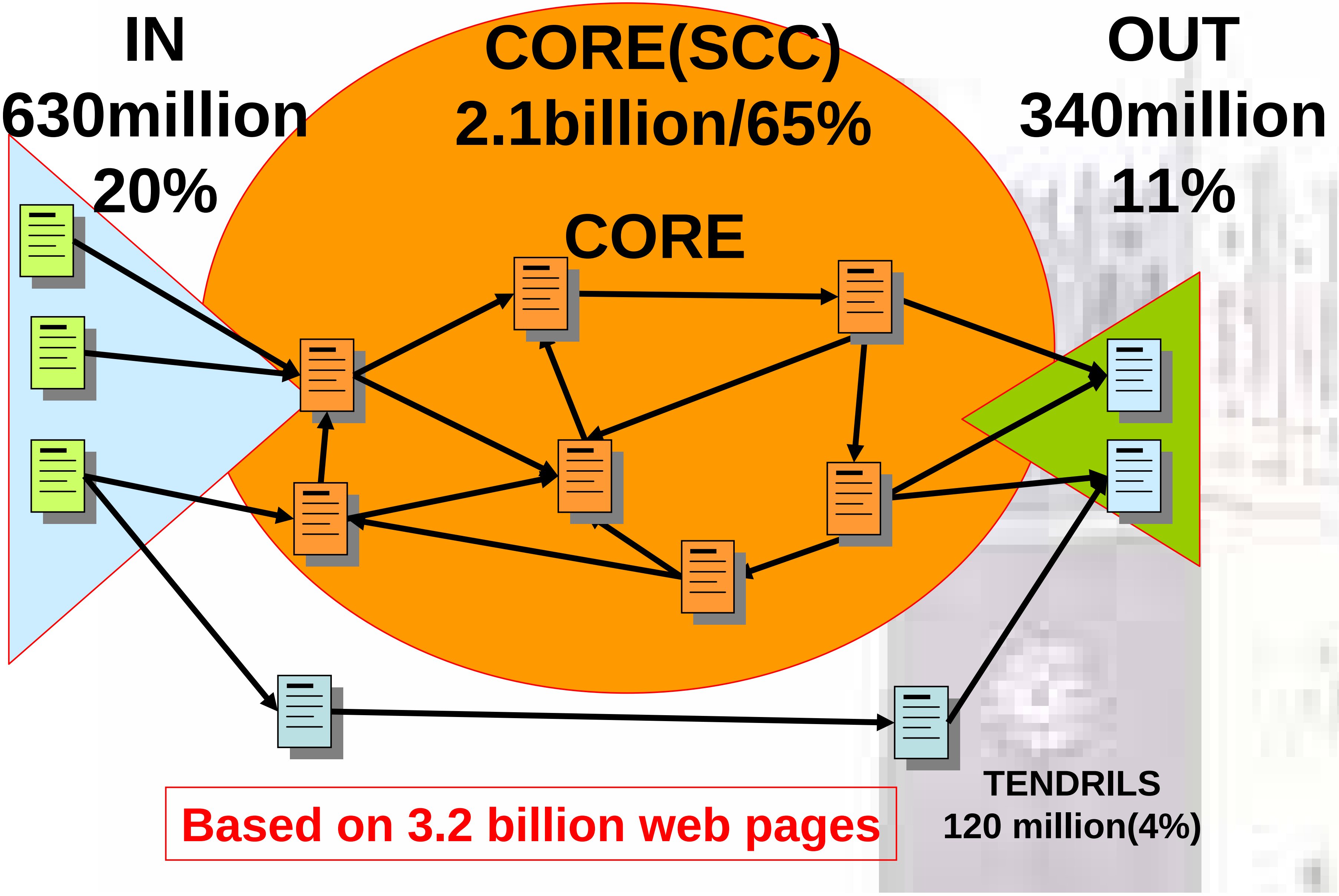
# of hosts : 1,719,134 hosts

# of inter-host links : 91,084,879 links

Computation Environment

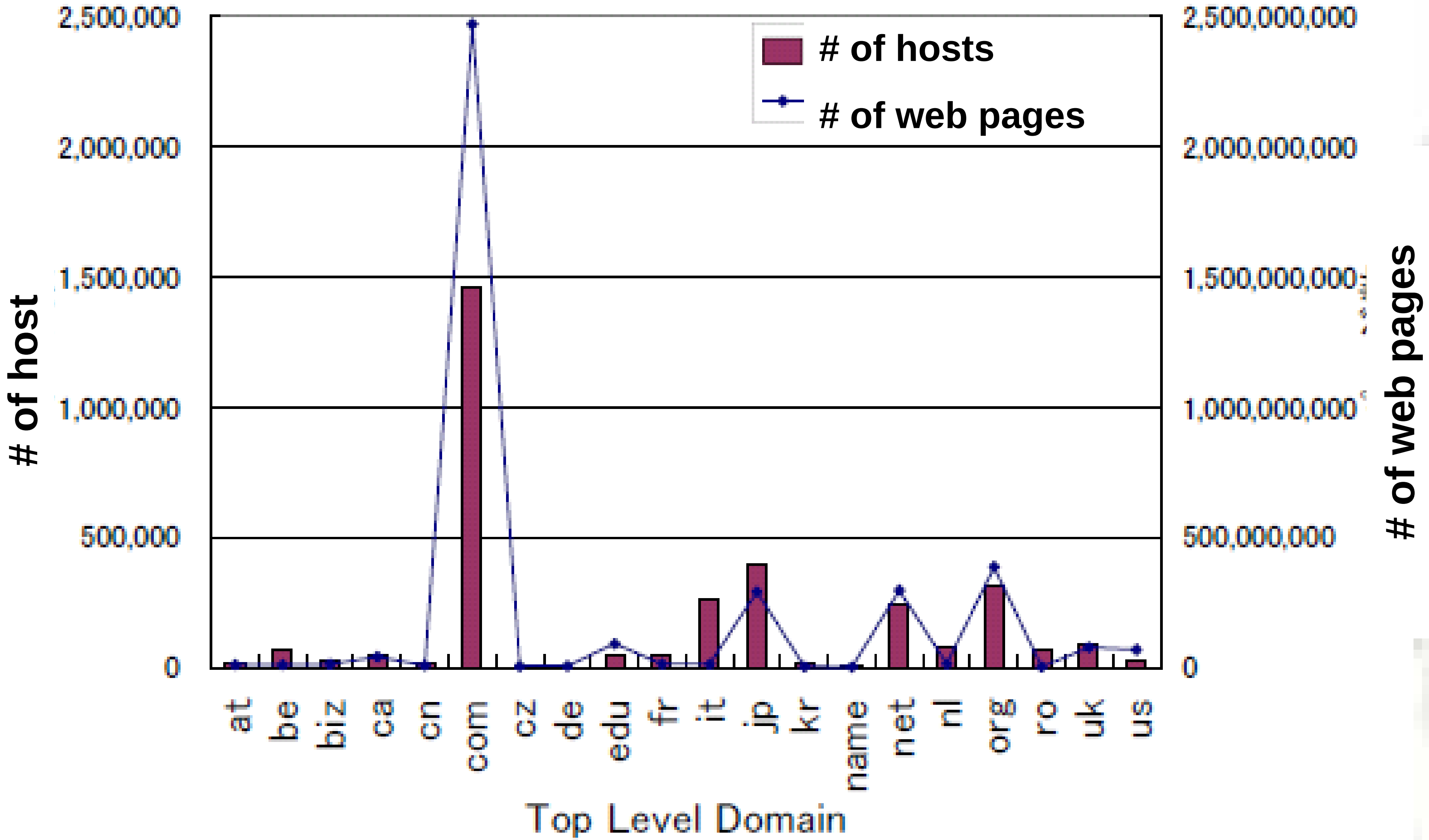
CPU: Opteron 2.4GHz x 2
Memory 16GB
HDD: 300GB x 12(RAID5+spare) x 2 = 4.7TB
boost graph library

(1) Web Structure in 2005 (whole)



Based on 3.2 billion web pages

Dataset Property(2) – TLD distribution



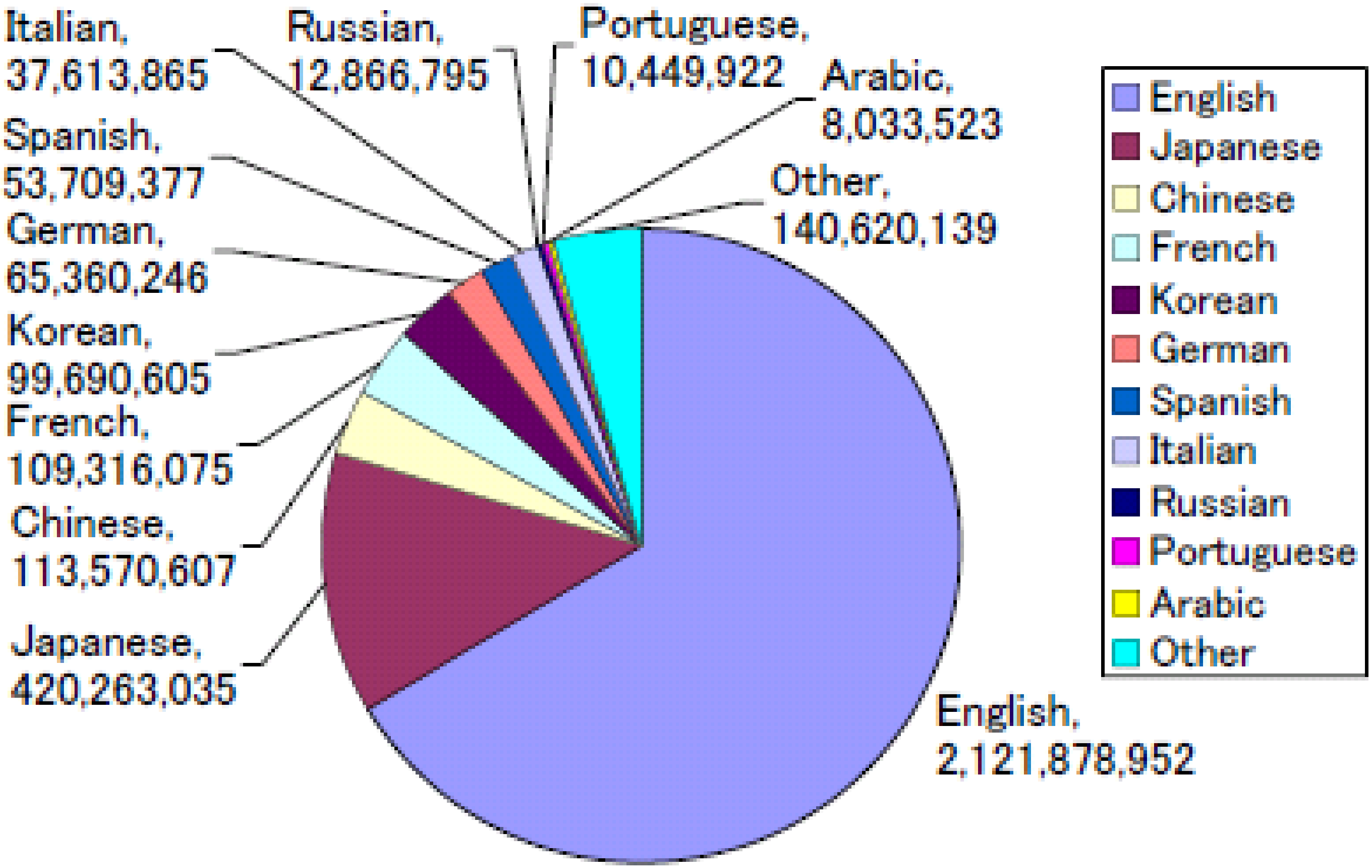
(2) Web Structures by TLD

TLD	CORE	IN	OUT	Other
com	53.65%	19.73%	22.25%	4.37%
jp	26.46%	1.77%	71.32%	0.46%
de	0.25%	0.05%	78.36%	21.34%
edu	0.05%	0.00%	14.44%	85.51%
fr	0.01%	0.02%	25.33%	74.63%
it	0.11%	0.04%	0.04%	99.81%
kr	0.00%	0.00%	1.09%	98.91%
net	0.52%	0.17%	35.42%	63.89%
org	0.61%	0.38%	64.25%	34.76%
ru	0.77%	0.05%	0.49%	98.70%

Web pages cannot be divided by TLD

whole	65%	20%	11%	4%

Dataset Property(3) – written language distribution



Identified by basis technology language identifier

(3) Web Structures by written language

Language	CORE	IN	OUT	Other
English	66.9%	9.0%	16.4%	7.7%
Japanese	71.1%	25.9%	2.5%	0.5%
Arabic	61.4%	10.2%	18.6%	9.8%
Chinese	76.9%	10.0%	10.6%	2.5%
French	61.9%	9.2%	[illegible]	[illegible]
German	26.6%	8.2%	42.2%	23.0%
Italian	23.7%	17.1%	29.5%	29.7%
Korea	54.3%	17.1%	19.4%	9.2%
Portuguese	26.6%	4.9%	42.2%	26.3%
Russian	35.8%	18.2%	18.4%	27.6%
Spanish	64.9%	5.3%	23.6%	6.2%

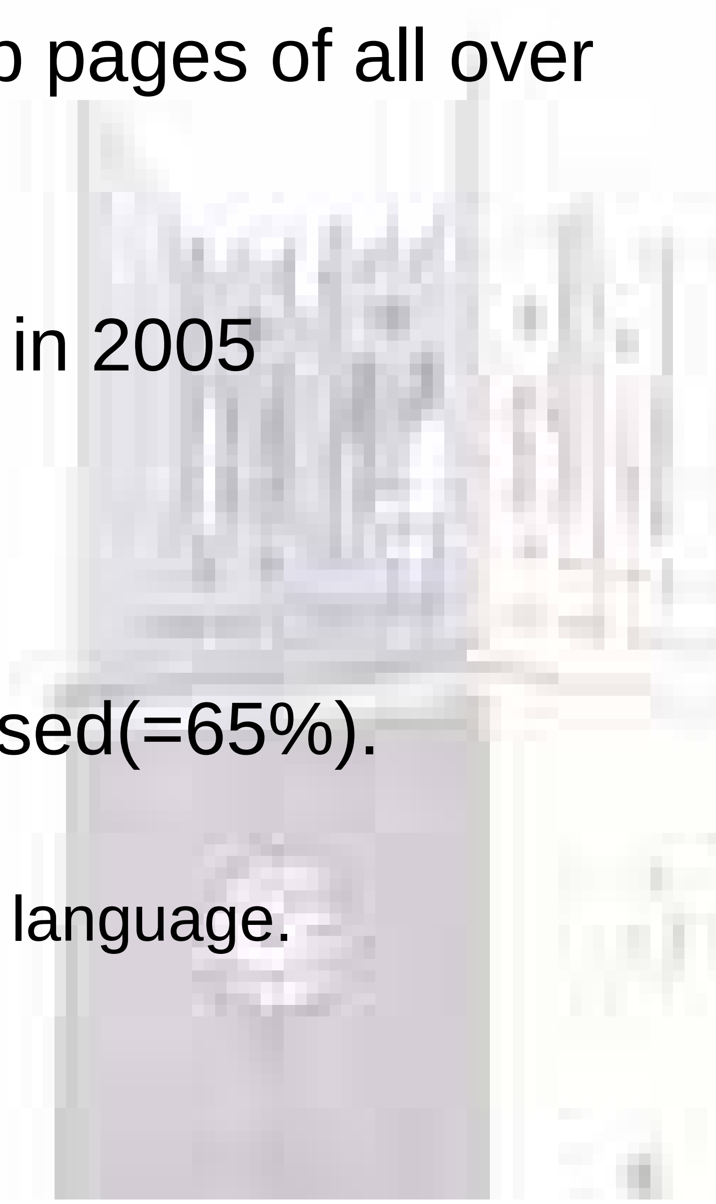
all	65%	20%	11%	4%

Similar to Lie's china web structure
Lie's china web structure: 80%

Part4: Conclusion

Conclusion

- We estimated the number of web pages of all over the world
  - 53.7 billion web pages
- We generated the web structure in 2005
  - The whole web structure
  - Web structures by TLD
  - Web structures by languages
- The percentage of CORE increased(=65%).
  - Web pages are well connected.
  - Web pages can be divided by their language.



Ongoing & Future Works

- Generate web structures based on 14 billion web pages
  - Need to develop parallel processing
- Compute Page Ranks based on 14 billion web pages