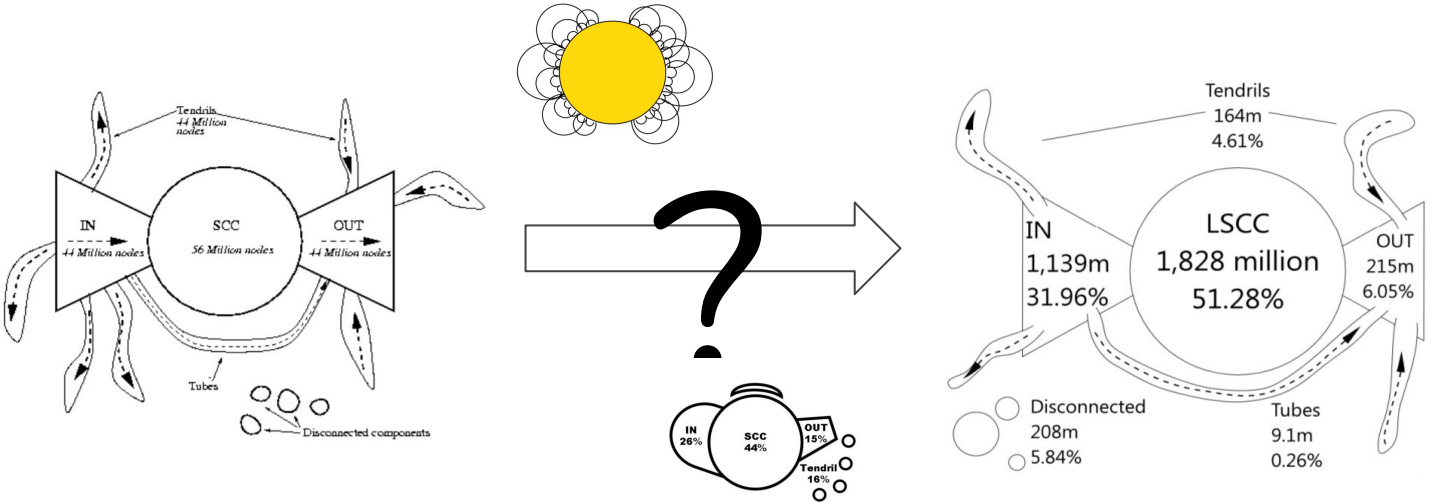
Studying the Evolution of Web Graph

Daniel Kang

Mentor: Michele Catasta

Problem



- **Project**: looking at changes in structure of the web
- **Dataset**: 12 years of data (~57TB)
- **Main challenge**: huge dataset, network structure

Pipeline

AUT, PySpark | PySpark | Webgraph (Java)

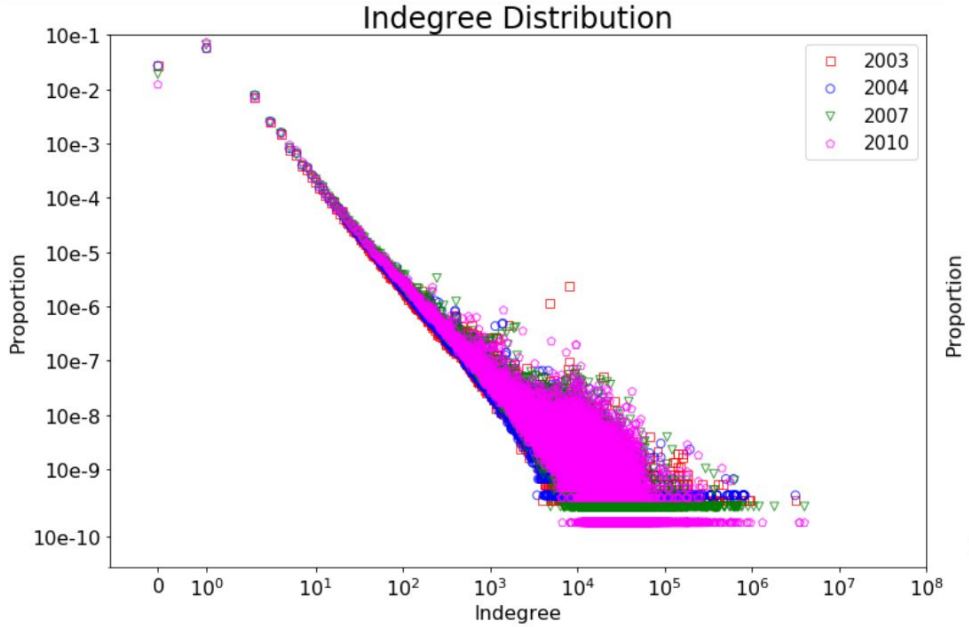
Change in dataset sizes

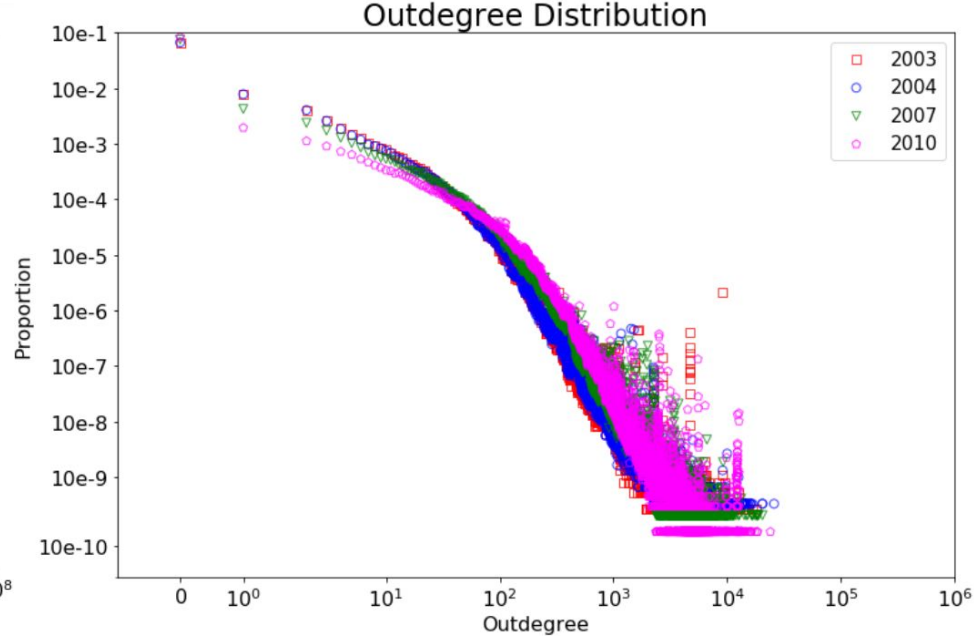
1.7 TB	28 GB	47 GB	5.3 GB	(2003)
2.2 TB	21 GB	37 GB	4.4 GB	(2004)
3.0 TB	44 GB	71 GB	9.5 GB	(2007)
5.4 TB	86 GB	111 GB	14 GB	(2010)

Network Statistics

	2003	2004	2007	2010	Broder00	Donato04	Meusel14
# Nodes (in million)	376	305	491	677	203	185	3,563
# Arcs (in million)	2,071	1,602	3,274	5,198	1,466	1,500	128,736
Avg degree	5.50	5.25	6.67	7.68	7.5	-	36.8

Degree distributions





Component Statistics

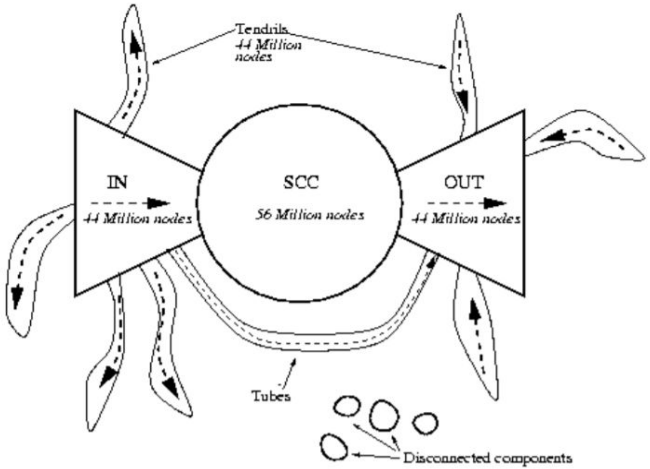
	2003	2004	2007	2010	Broder00	Donato04	Meusel14
Max SCC	9.2 (2.43)	8.5 (2.79)	17.2 (3.50)	22.0 (3.25)	56.4 (27.74)	44.7 (32.9)	1,827.5 (51.28)
Max WCC	357.5 (94.86)	291.1 (95.41)	478.3 (97.40)	661.7 (97.78)	186 (91.62)	166.5 (90+)	3,349.2 (94)
IN	47.5 (12.61)	37.8 (12.39)	48.5 (9.87)	47.5 (7.02)	43.3 (21.29)	14.4 (10.6)	1,138.9 (31.96)
OUT	81.6 (21.64)	69.2 (22.67)	175.3 (35.70)	295.5 (43.67)	43.2 (21.21)	53.3 (39.3)	215.4 (6.05)
Dangling	249.1 (66.09)	202.6 (66.40)	366.1 (74.55)	557.0 (82.31)			

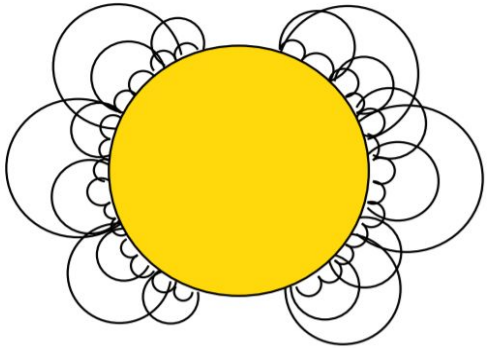
* all numbers in million ** number in parentheses: proportion of all nodes, in percent

BFS Statistics

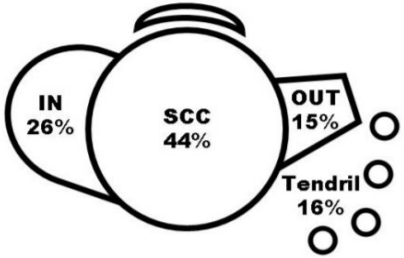
	2003	2004	2007	2010	Broder00	Donato04	Meusel14
BFS depth	765 ~ 775	567 ~ 574	1330 ~ 1334	8173 ~ 8180	475 ~ 503	580 (OUT)	> 5,282
Reverse BFS depth	130 ~ 139	808 ~ 817	799 ~ 804	196 ~ 202	430 ~ 444	8 (IN)	-

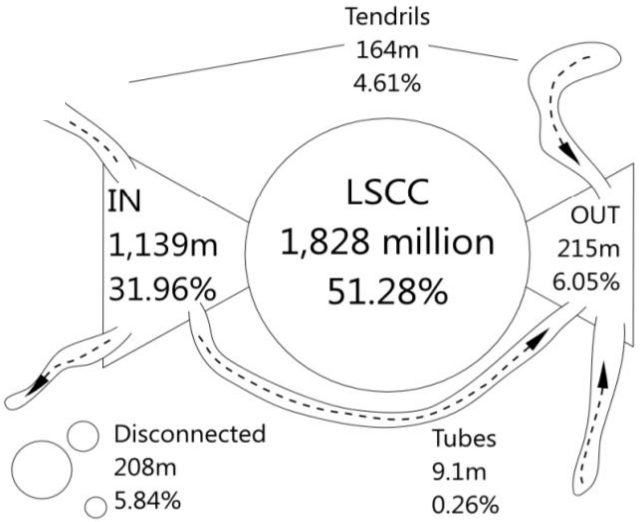
Structure?


Tendrils
44 Million
nodes
IN
44 Million nodes
SCC
56 Million nodes
OUT
44 Million nodes
Tubes
Disconnected components

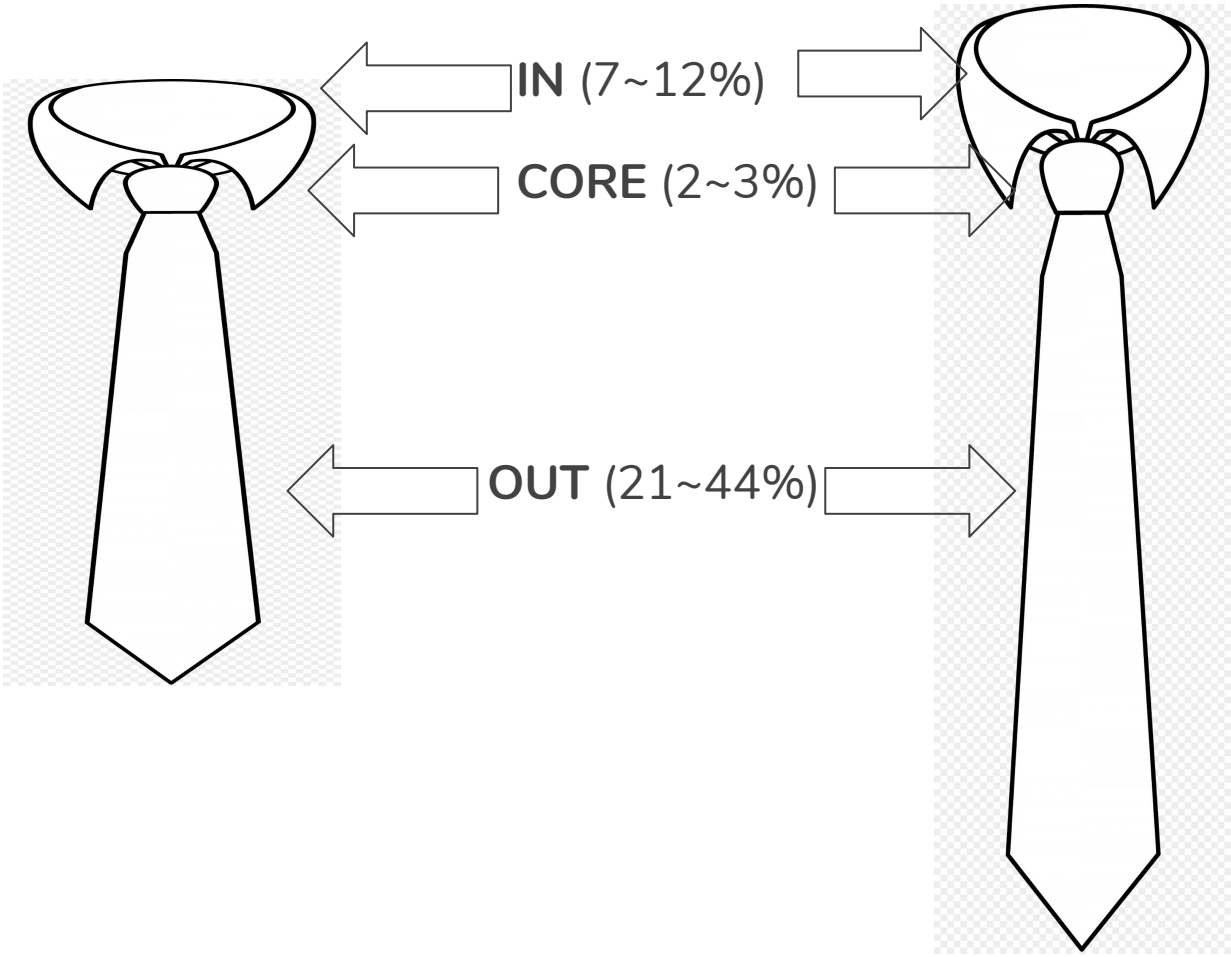


?


IN
26%
SCC
44%
OUT
15%
Tendril
16%


Tendrils
164m
4.61%
IN
1,139m
31.96%
LSCC
1,828 million
51.28%
OUT
215m
6.05%
Disconnected
208m
5.84%
Tubes
9.1m
0.26%

Tie?....



Domain Graph

	2003	2004	2007	2010
# Nodes (in million)	7.69	7.24	12.16	9.25
# Arcs (in million)	41.62	34.77	34.54	25.33
Avg degree	5.41	4.79	2.80	2.74
Max SCC (in thousands)	72.59 (0.94)	49.78 (0.69)	40.21 (0.33)	22.78 (0.25)
Max WCC (in million)	7.68 (99.99)	7.24 (99.99)	12.16 (99.99)	9.24 (99.99)
BFS depth	7	5~6	5~6	5~6
Reverse BFS depth	8	6~8	6~7	6~7

Frequent Source Domains

	2003	2004	2007	2010
1	da.ru	tripod.lycos.com	tripod.lycos.com	tripod.lycos.com
2	directory.google.com	webbound.com	ljudmila.org	gy.com
3	anywho.com	cyber.law.harvard.edu	google.com	mybloglog.com
4	dominion-web.com	anywho.com	paginegialle.it	directory.google.com
5	tollfree.att.net	tollfree.att.net	choicehotels.com	google.com
6	att.net	salon.com	gy.com	at-la.com
7	suchmaschine.com	att.net	directory.google.com	mister-wong.de
8	dmoz.org	directory.google.com	educationplanet.com	hotsheet.com
9	newhoo.com	gy.com	hotsheet.com	rhymeswithright.mu.nu
10	asia.dir.yahoo.com	excite.co.uk	stumbleupon.com	choicehotels.com

2003


europa.eu.int
nytimes.com
pbs.org
w3.org
cnn.com
netscape.com
yahoo.com
indiana.edu
lcweb.loc.gov
zdnet.com
real.com
columbia.edu
bbc.co.uk
home.earthlink.net
microsoft.com
nsf.gov
senate.gov
ibm.com
dir.yahoo.com
ed.gov
cdc.gov
epa.gov
lycos.com
adobe.com
stanford.edu
nps.gov
un.org
utexas.edu
groups.yahoo.com
census.gov
altavista.com
excite.com
geocities.com
sun.com
thomas.loc.gov
web.mit.edu
mapquest.com
whitehouse.gov
google.com
house.gov
loc.gov
msnbc.com
usatoday.com
apple.com
news.bbc.co.uk

2007


members.aol.com
news.bbc.co.uk
groups.yahoo.com
macromedia.com
ncbi.nlm.nih.gov
validator.w3.org
maps.google.com
msnbc.msn.com
loc.gov
apple.com
latimes.com
youtube.com
en.wikipedia.org
go.microsoft.com
usdoj.gov
mapquest.com
google.com
geocities.com
ed.gov
epa.gov
nsf.gov
web.mit.edu
census.gov
cdc.gov
pbs.org
nature.com
npr.org
un.org
who.int
indiana.edu
nasa.gov
w3.org
adobe.com
state.gov
sfgate.com
europa.eu.int
bbc.co.uk
usatoday.com
senate.gov
microsoft.com
amazon.com
real.com
nps.gov
stanford.edu
myspace.com
guardian.co.uk
house.gov
thomas.loc.gov
nytimes.com
angelfire.com
columbia.edu
home.earthlink.net
cnn.com
yahoo.com
whitehouse.gov
members.tripod.com

2010


whitehouse.gov
ncbi.nlm.nih.gov
flickr.com
reuters.com
addthis.com
online.wsj.com
microsoft.com
web.mit.edu
usatoday.com
latimes.com
abcnews.go.com
surveymonkey.com
en.wikipedia.org
npr.org
myspace.com
linkedin.com
twitter.com
nps.gov
loc.gov
who.int
bbc.co.uk
news.bbc.co.uk
cnn.com
w3.org
sfgate.com
members.aol.com
facebook.com
ed.gov
nsf.gov
wired.com
macromedia.com
un.org
time.com
bit.ly
pbs.org
tinyurl.com
youtube.com
businessweek.com
census.gov
boston.com
yahoo.com
nature.com
cdc.gov
epa.gov
google.com
get.adobe.com
guardian.co.uk
msnbc.msn.com
amazon.com
stanford.edu
apple.com

Frequent Dest Domains

	2003	2004	2007	2010
1	adobe.com	adobe.com	adobe.com	facebook.com
2	microsoft.com	microsoft.com	microsoft.com	youtube.com
3	geocities.com	geocities.com	google.com	twitter.com
4	amazon.com	amazon.com	geocities.com	adobe.com
5	members.aol.com	google.com	amazon.com	google.com
6	google.com	members.aol.com	en.wikipedia.org	en.wikipedia.org
7	yahoo.com	yahoo.com	apple.com	microsoft.com
8	cnn.com	cnn.com	nytimes.com	amazon.com
9	apple.com	nytimes.com	cnn.com	nytimes.com
10	nytimes.com	apple.com	members.aol.com	maps.google.com

Why Adobe?

	2003	2007
1	adobe.com/products/acrobat/readstep2.html	adobe.com/products/acrobat/readstep2.html
2	adobe.com/products/acrobat/readstep.html	adobe.com/products/acrobat/readstep.html
3	adobe.com/prodindex/acrobat/readstep.html	adobe.com/prodindex/acrobat/readstep.html
4	adobe.com	adobe.com
5	adobe.com/	adobe.com/
6	access.adobe.com	adobe.co.jp/products/acrobat/readstep2.html
7	adobe.com/acrobat/readstep.html	adobe.com/shockwave/download/download.cgi?P1_Prod_Version=ShockwaveFlash
8	adobe.com/prodindex/acrobat/readstep.html#reader	adobe.com/shockwave/download/download.cgi?P1_Prod_Version=ShockwaveFlash&promoid=BIOW
9	adobe.com/products/acrobat/readermain.html	adobe.com/products/acrobat/readermain.html
10	adobe.co.jp/products/acrobat/readstep.html	adobe.com/go/getflashplayer

Why Microsoft?

	2003	2007
1	microsoft.com/ContentRedirect.asp?prd=iis&sbp=&pver=5.0&pid=&ID=404&cat=web&os=&over=&hrd=&Opt1=&Opt2=&Opt3=	go.microsoft.com/fwlink/?linkid=8180
2	hardwarecentral.dealtime.com/xKW-microsoft_visual_studio_net_professional/FN-Programming_Tools/DL-0/NS-1/linkin_id-3011677/GS.html	microsoft.com/ContentRedirect.asp?prd=iis&sbp=&pver=5.0&pid=&ID=404&cat=web&os=&over=&hrd=&Opt1=&Opt2=&Opt3=
3	microsoft.com/windows/ie/default.asp	microsoft-watch.com
4	microsoft.com	microsoft.com
5	microsoft.com/windows/ie/	microsoft.com/windows/ie/default.asp
6	microsoft.com/windows/ie/default.htm	microsoft.com/windows/ie/default.htm
7	microsoft-watch.com?kc=MWZD10111TTX1B0000538	microsoft.com/
8	microsoft.com/	microsoft.com/windows/ie/default.mspx
9	microsoft.com/ie	msdn.microsoft.com/msdnmag/
10	microsoft.com/downloads/search.asp?	vtc.com/modules/content/microsoft.php

Overview

- **Summary:**
  - Parsed 4 years of data
  - Created web graphs, looked at characteristics of the graphs and compared with previous studies
  - Created domain graphs
- **Challenges:**
  - Size of data, parsing data
  - Inconsistency with other results (especially SCC)
- **Moving forward:**
  - Identify the source of inconsistency, what can we learn from it? Do we need to match the results?