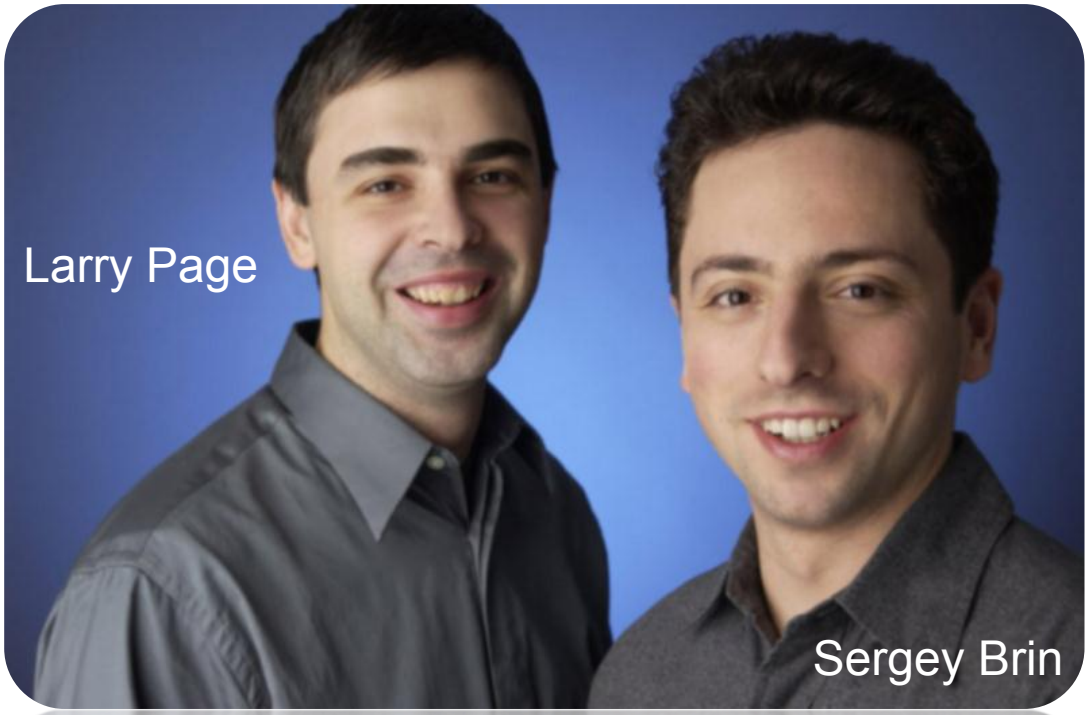
PageRank centrality on directed networks

Seung-Woo Son
Dept. Applied Physics
HYU 한양대학교 HANYANG UNIVERSITY

Co-founders of Google

- Ph.D. students of Stanford University
- Invented **PageRank** algorithm



Early `90s start Internet Era.

- the first Internet search engine, W3Catalog (1993)
- Lycos (1994), Alta Vista (1995), and Yahoo! (1995)
- Google (1998)

Forbes

http://www.forbes.com/profile/larry-page/

Sergey Brin

Net Worth **$18.7 B** As of March 2012

+ Follow (258)

At a Glance

Co-Founder, Google

Age: **38**

Source of Wealth: **Google, self-made**

Residence: **Los Altos, CA**

Forbes Lists

#24 Forbes Billionaires

#13 in United States

#30 Powerful People

#15 Forbes 400

Photos

Like 147 people like this. Be the first of your friends.

OPEN ACCESS Freely available online

PLoS one

Who Is the Best Player Ever? A Complex Network Analysis of the History of Professional Tennis

Filippo Radicchi*

Department of Chemical and Biological Engineering, Northwestern University, Evanston, Illinois, United States of America

Abstract

We considered all matches played by professional tennis players between 1968 and2010, and, on the basis of this data set, constructed a directed and weighted network of contacts. The resulting graph showed complex features, typical of many real networked systems studied in literature. We developed a diffusion algorithm and applied it to the tennis contact network in order to rank professional players. *Jimmy Connors* was identified as the best player in the history of tennis according to our ranking procedure. We performed a complete analysis by determining the best players on specific playing surfaces as well as the best ones in each of the years covered by the data set. The results of our technique were compared to those of two other well established methods. In general, we observed that our ranking method performed better: it had a higher predictive power and did not require the arbitrary introduction of external criteria for the correct assessment of the quality of players. The present work provides novel evidence of the utility of tools and methods of network theory in real applications.

#13 Sergey Brin

Forbes

Real Time Net Worth As of 5/23/16

$35.3 Billion

Cofounder, Director Of Special Projects, Google

Age	42
Source Of Wealth	Google, Self Made
Self-Made Score	9
Residence	Los Altos, CA
Citizenship	United States
Marital Status	Divorced
Children	2
Education	Master of Science, Stanford University; Bachelor of Arts / Science, University of Maryland, College Park

Sergey Brin on Forbes Lists

#13 Billionaires (2016)

- #10 in United States
- #20 in 2015

#30 Powerful People (2015)

#11 Forbes 400 (2015)

#6 Richest In Tech (2015)

Larry Page

Real Time Net Worth As of 5/23/16

$36.1 Billion

CEO, Google

Age	**43**
Source Of Wealth	**Google, Self Made**
Self-Made Score	**8**
Residence	**Palo Alto, CA**
Citizenship	**United States**
Marital Status	**Married**
Children	**1**
Education	**Bachelor of Arts / Science, University of Michigan; Master of Science, Stanford University**

Forbes

Larry Page on Forbes Lists

Global Game Changers (2016)

#12 Billionaires (2016)

- #9 in United States

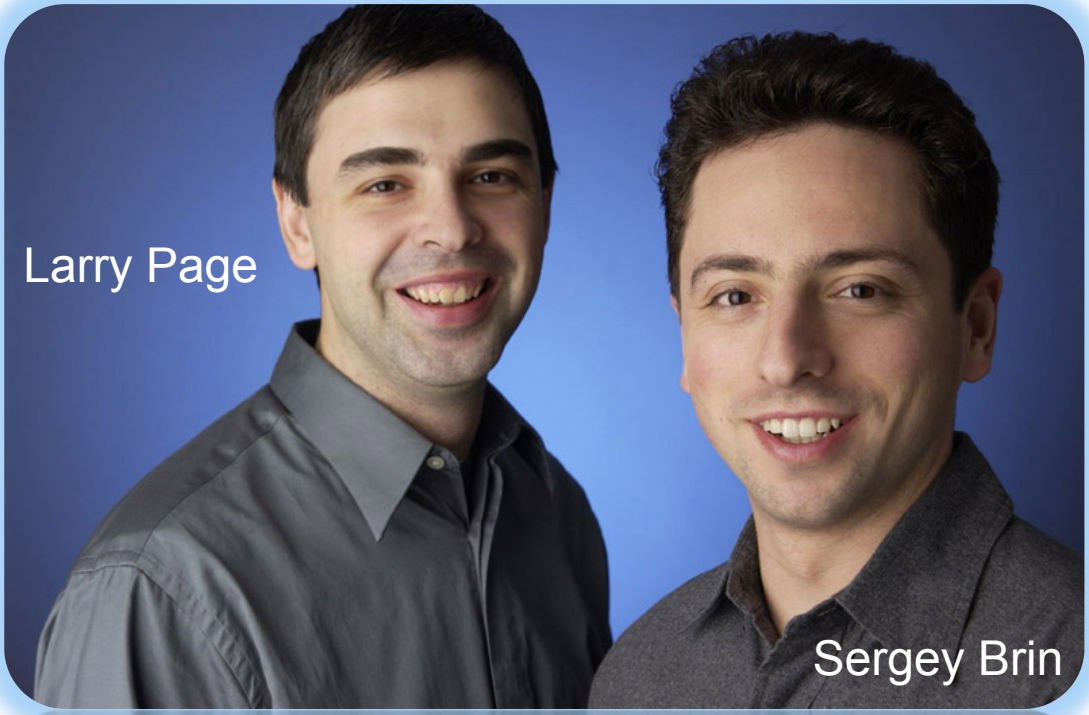
#10 Powerful People (2015)

#10 Forbes 400 (2015)

#5 Richest In Tech (2015)

Co-founders of Google

- Ph.D. students of Stanford University
- Invented **PageRank** algorithm



Early `90s start Internet Era.

- the first Internet search engine, W3Catalog (1993)
- Lycos (1994), Alta Vista (1995), and Yahoo! (1995)
- Google (1998)

Contents based algorithms

V.S.

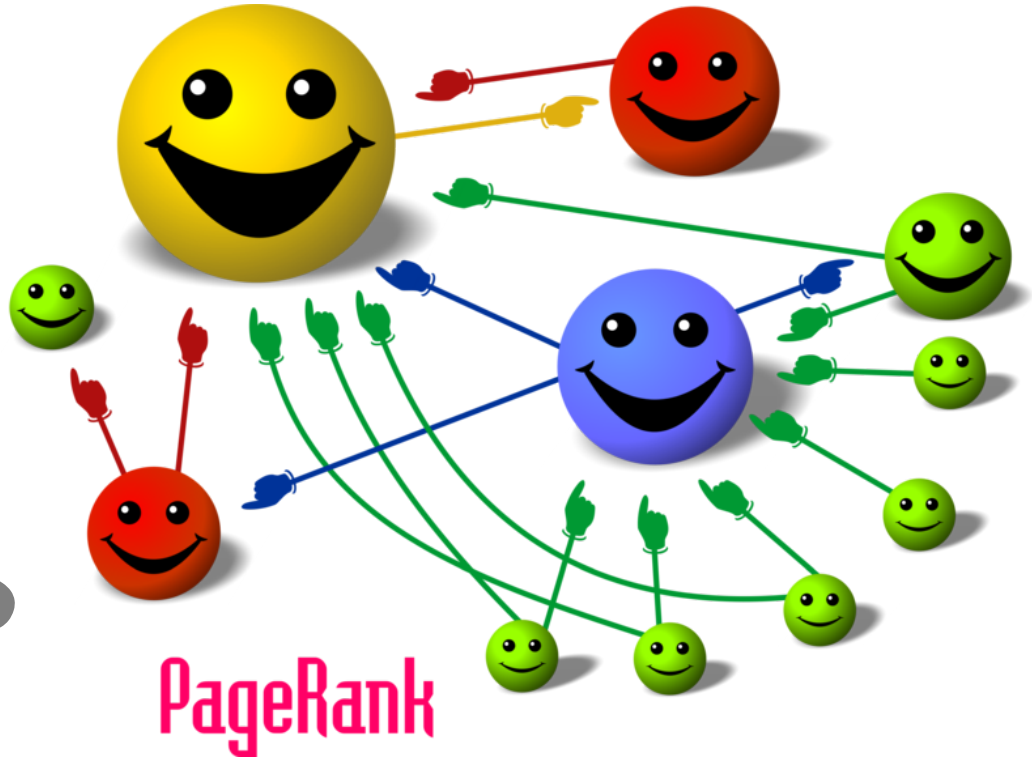
Link analysis algorithms

PageRank (Larry Page) ← HIT algorithm (Jon Kleinberg)

Hyperlink-Induced Topic Search hubs and authorities

- A link analysis algorithm

- One of hundreds of factors Google considering for better search results



$$\pi(t) = dP\pi(t-1) + \frac{(1-d)}{N}1$$?

1. A random surfer (walker), starting from a random web page, chooses the next page to which it will move by clicking at random one among the hyperlinks in the current page. (with probablity d)

2. Otherwise, with probability (1 – d), the surfer jumps to any web page in the network.

3. If a page is a dangling end, meaning it has no outgoing hyperlinks, the random surfer selects an arbitrary web page from a uniform distribution and "teleports" to that page.

the probability to be on page i at time t

damping factor d

random teleportation

$$\pi_i(t) = d \sum_j p_{ij} \pi_j(t-1) + \frac{(1-d)}{N}$$

$$p_{ij} = 1/k_j^{\text{out}}$$

$$p_{ij} = 1/N$$

if page j has an outgoing link to i, otherwise 0

if page j is a dangling end $(k_j^{\text{out}} = 0)$

$$\boldsymbol{\pi}(t) = d\boldsymbol{P}\boldsymbol{\pi}(t-1) + \frac{(1-d)}{N}\mathbf{1}$$

where $\mathbf{1} = [1, \ldots, 1]^T$

For $t \to \infty$, $\boldsymbol{\pi}(t)$ converges to the stationary state

$$\boldsymbol{\pi}(t) = d\boldsymbol{P}\boldsymbol{\pi}(t-1) + \frac{(1-d)}{N}\mathbf{1}$$

$$(\boldsymbol{I} - d\boldsymbol{P})\boldsymbol{\pi} = \frac{(1-d)}{N}\mathbf{1}$$

solution of this linear system

or

Google matrix

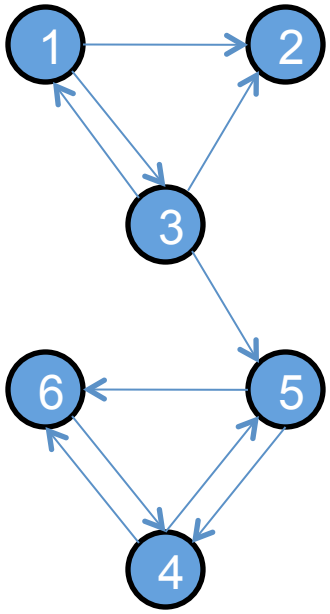
$$\boldsymbol{G} = d\boldsymbol{P} + \frac{(1-d)}{N}\boldsymbol{E},$$

where the matrix $\boldsymbol{E} = \mathbf{1}\mathbf{1}^T$ is 1 for all elements

$$\boldsymbol{\pi}(t) = \boldsymbol{G}\boldsymbol{\pi}(t-1)$$

For d<1, left stochastic, aperiodic, irreducible!

PR vector $\boldsymbol{\pi}$ is the principal eigenvector of the system $\boldsymbol{G}\boldsymbol{\pi} = \boldsymbol{\pi}$



$$A = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{pmatrix}, P = \begin{pmatrix} 0 & 1/6 & 1/3 & 0 & 0 & 0 \\ 1/2 & 1/6 & 1/3 & 0 & 0 & 0 \\ 1/2 & 1/6 & 0 & 0 & 0 & 0 \\ 0 & 1/6 & 0 & 0 & 1/2 & 1 \\ 0 & 1/6 & 1/3 & 1/2 & 0 & 0 \\ 0 & 1/6 & 0 & 1/2 & 1/2 & 0 \end{pmatrix}$$

$$d = 0.9$$

$$G = \begin{pmatrix} 1/60 & 1/6 & 19/60 & 1/60 & 1/60 & 1/60 \\ 7/15 & 1/6 & 19/60 & 1/60 & 1/60 & 1/60 \\ 7/15 & 1/6 & 1/60 & 1/60 & 1/60 & 1/60 \\ 1/60 & 1/6 & 1/60 & 1/60 & 7/15 & 11/12 \\ 1/60 & 1/6 & 19/60 & 7/15 & 1/60 & 1/60 \\ 1/60 & 1/6 & 1/60 & 7/15 & 7/15 & 1/60 \end{pmatrix}$$

$$\pi^T = (.03721 \quad .05396 \quad .04151 \quad .3751 \quad .206 \quad .2862)$$

If $\pi_i > \pi_j$, then page i ranks above page j on the search results.

Public Belief vs. Truth

PageRank is more reliable when d is close to 1, since network structure is more reflected.

No. when d goes to 1, only sink component emerges on the top.

When d decreases, it is only fade-out of network structure gradually. There is only smooth transition into uniform.

No. it is rather drastic than a gradual transition. Why? ……

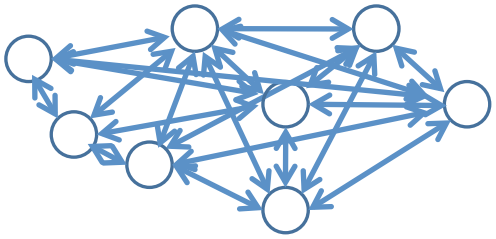
When d is small enough, we can deal this perturbation approach. Let's go…

Well… I don't think so.
Small enough d?
It probably depends on the network size.

Actually there is not enough knowledge about the **PageRank**.

Dilemma of PageRank

d=0

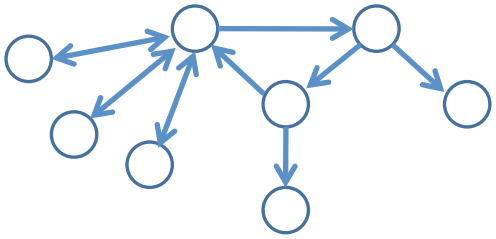


damping factor

tuning

d

d=1



All-to-all. randomly, uniformly

Following a network structure

When d is 0, trivial uniform distribution.
As d goes to 1, the network becomes more important.
What is the best value of the damping factor d? Is getting close to 1 ideal?

No, contrarily to popular belief, when d goes to 1, PageRank gets concentrated in the OUT components (“rank-sinks”) .

SCC decomposition

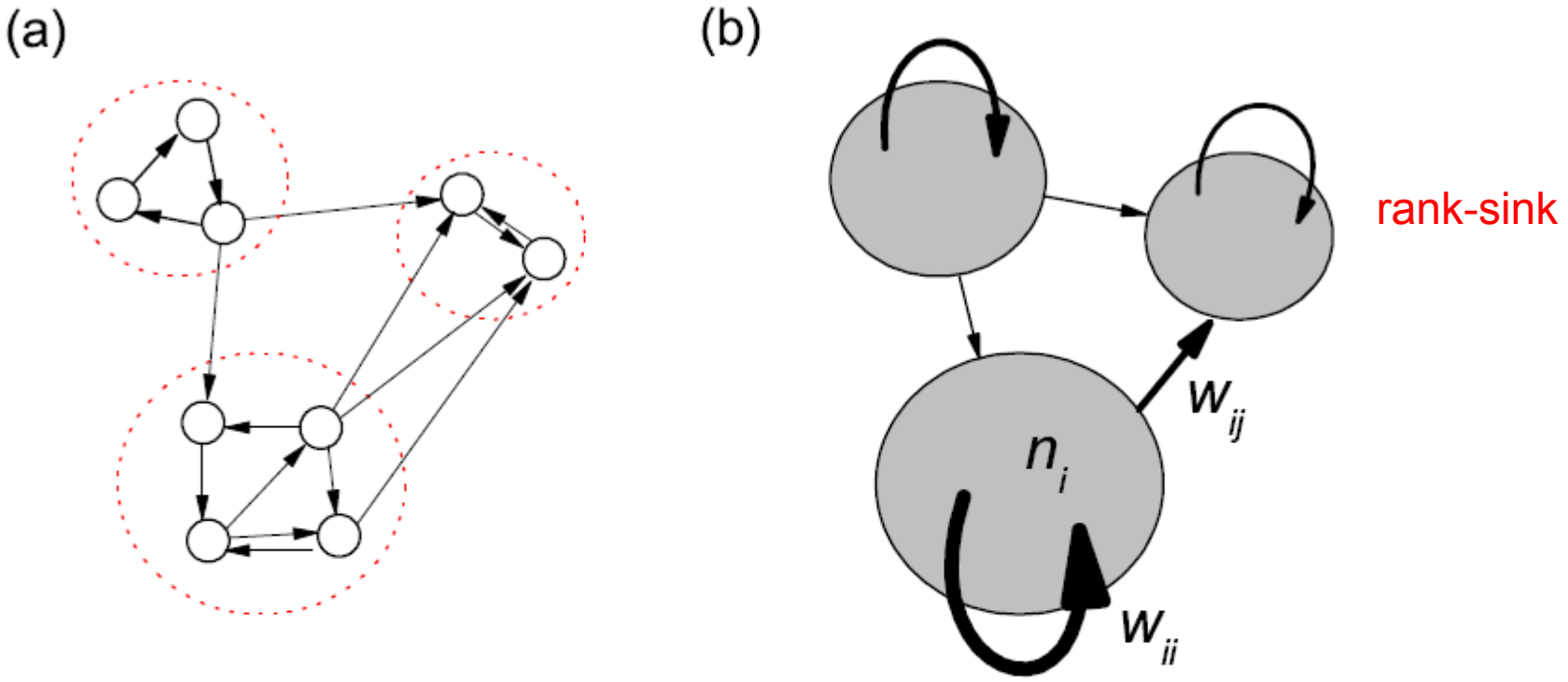


Figure 1. (Color online) SCC diagram. (a) A directed network can be decomposed into several SCCs. Each red-dotted circle in (a) delineates a SCC. (b) A directed network can also be abstracted into a SCC diagram through coarse-graining. This abstracted SCC diagram is an acyclic weighted network, containing self-links (denoted by w_{ii}) and size-heterogeneous nodes (n_i).

Big picture of a directed network

Bow tie diagram

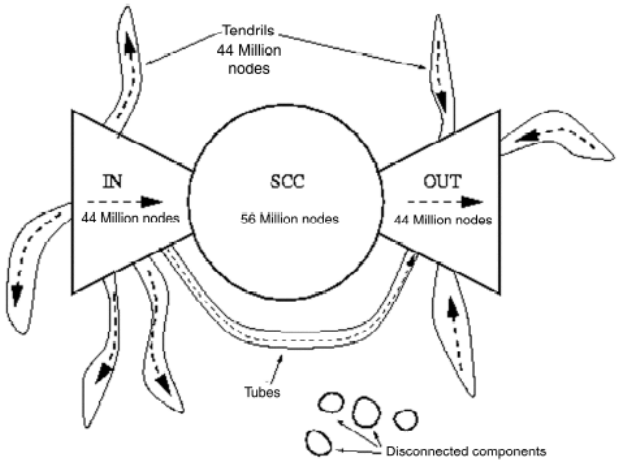
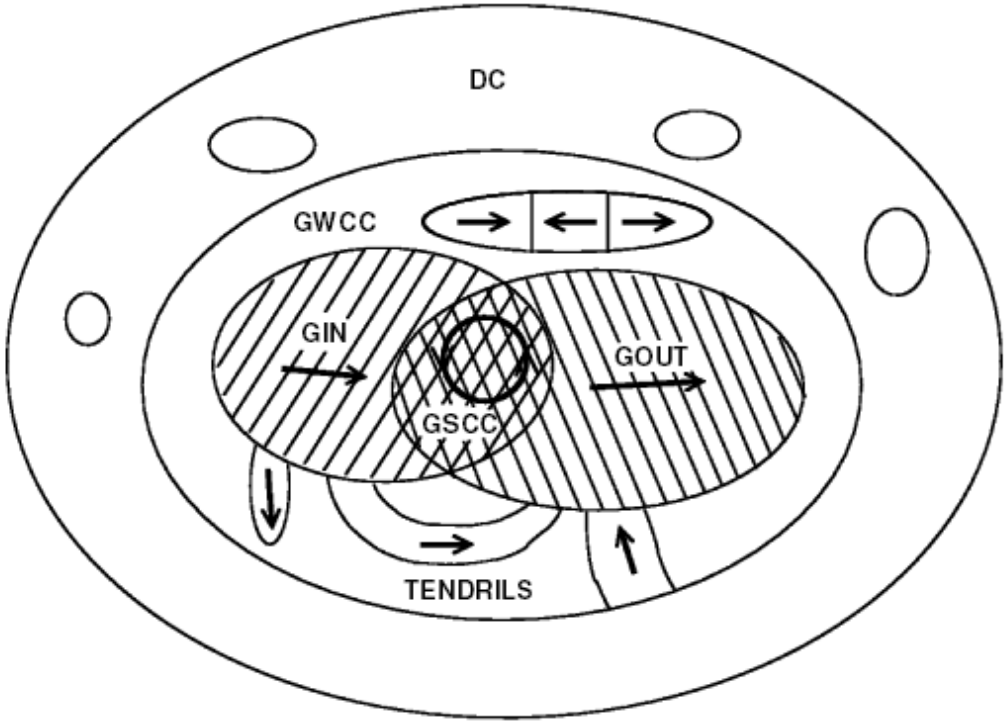


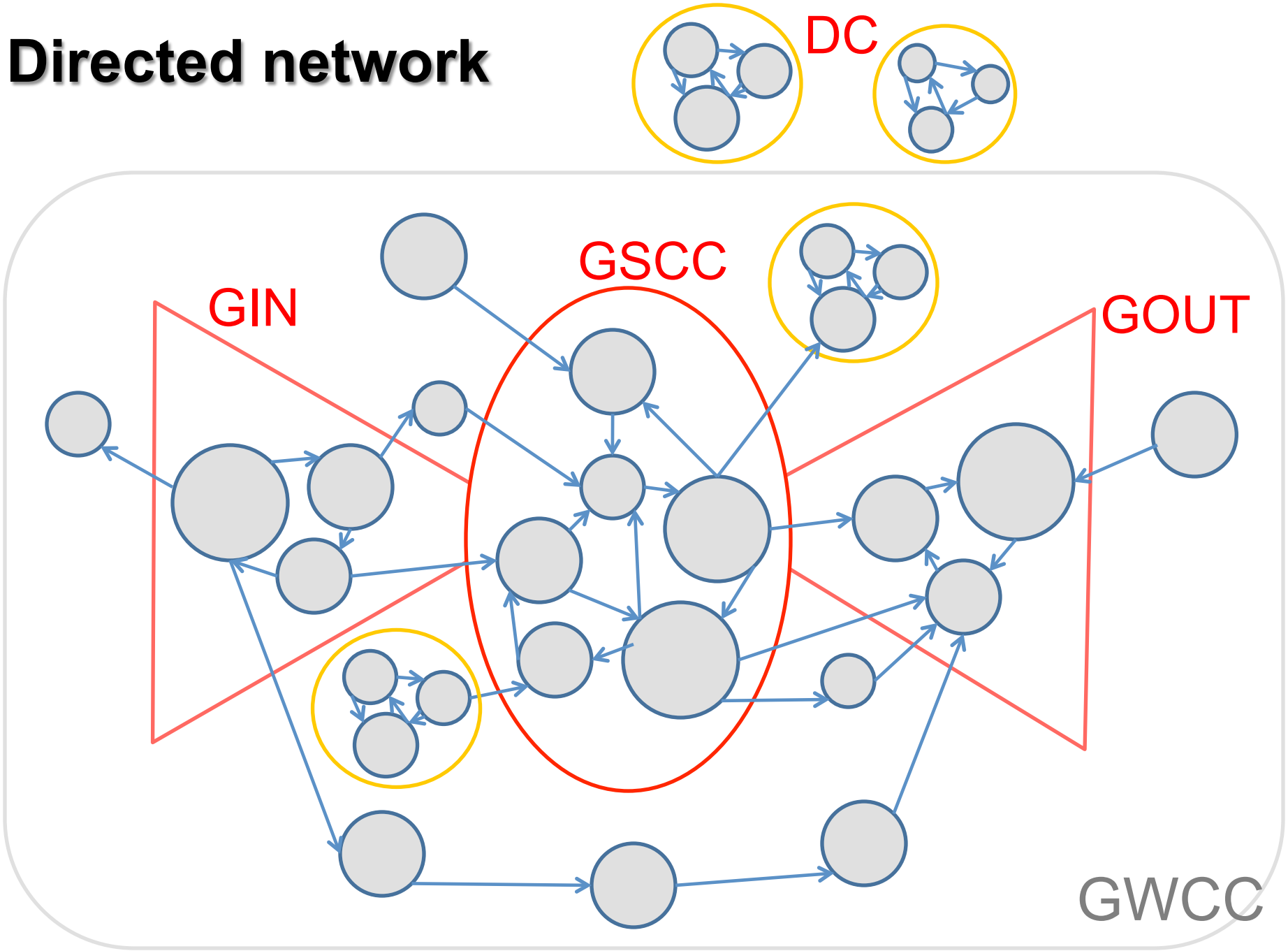
Fig. 9. Connectivity of the Web: one can pass from any node of IN through SCC to any node of OUT. Hanging off IN and OUT are TENDRILS containing nodes that are reachable from portions of IN, or that can reach portions of OUT, without passage through SCC. It is possible for a TENDRIL hanging off from IN to be hooked into a TENDRIL leading into OUT, forming a TUBE: i.e., a passage from a portion of IN to a portion of OUT without touching SCC.



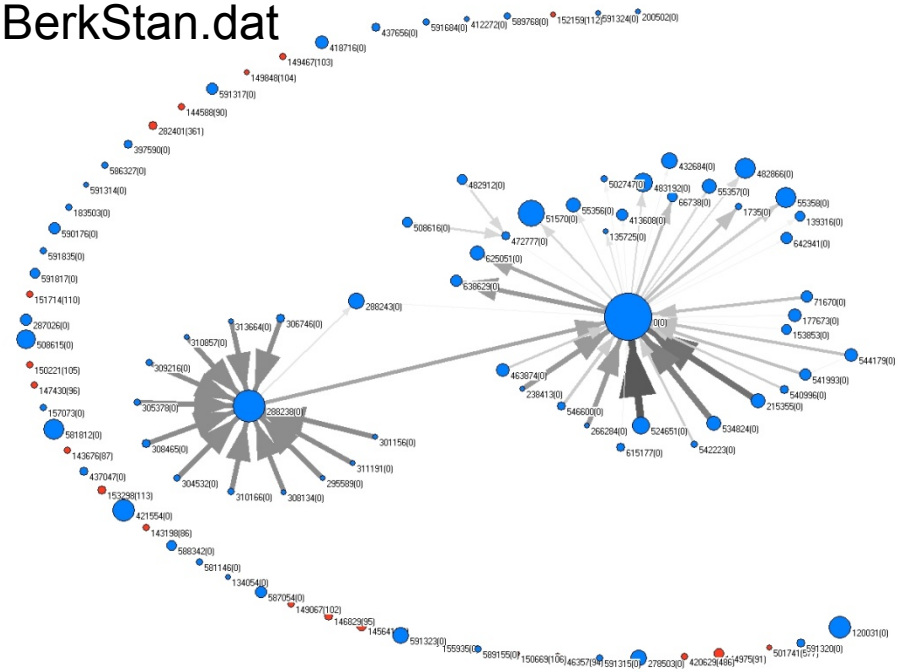
6. Structure of a directed graph when the giant strongly connected component is present [112] (see text). Also, the structure of the WWW (compare with figure 9 of reference [6]). If one ignores the directedness of edges, the network consists of the *giant weakly connected component* (GWCC)—actually, the usual percolating cluster—and disconnected components (DC). Accounting for the directedness of edges, the GWCC contains the following components:

A. Broder et al., "Graph structure in the Web," Computer Networks **33**, 309 (2000).
S. N. Dorogovtsev and J. F. F. Mendes, "Evolution of networks," Advances in Physics **51**, 1079 (2002).

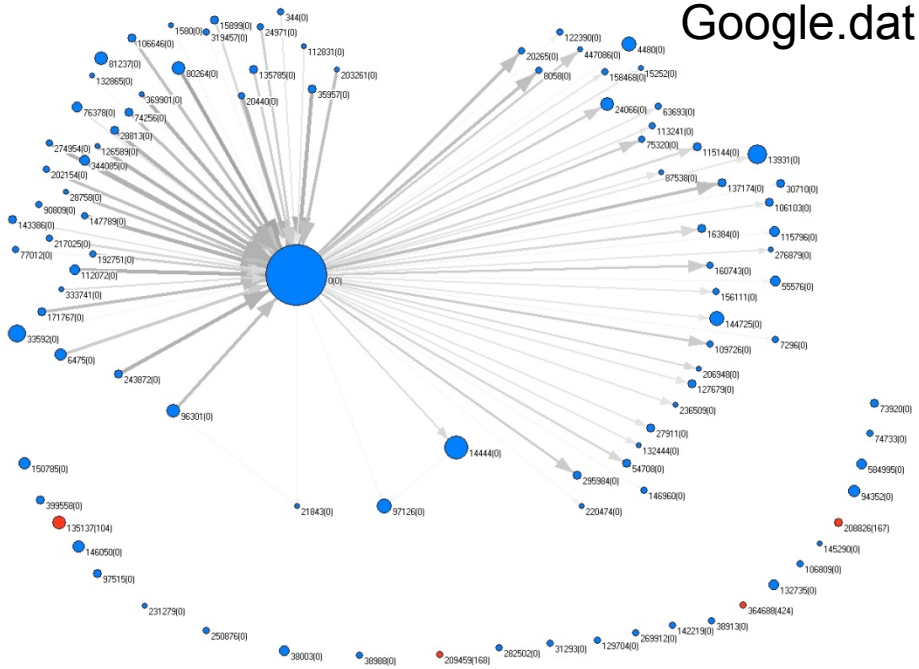
Directed network

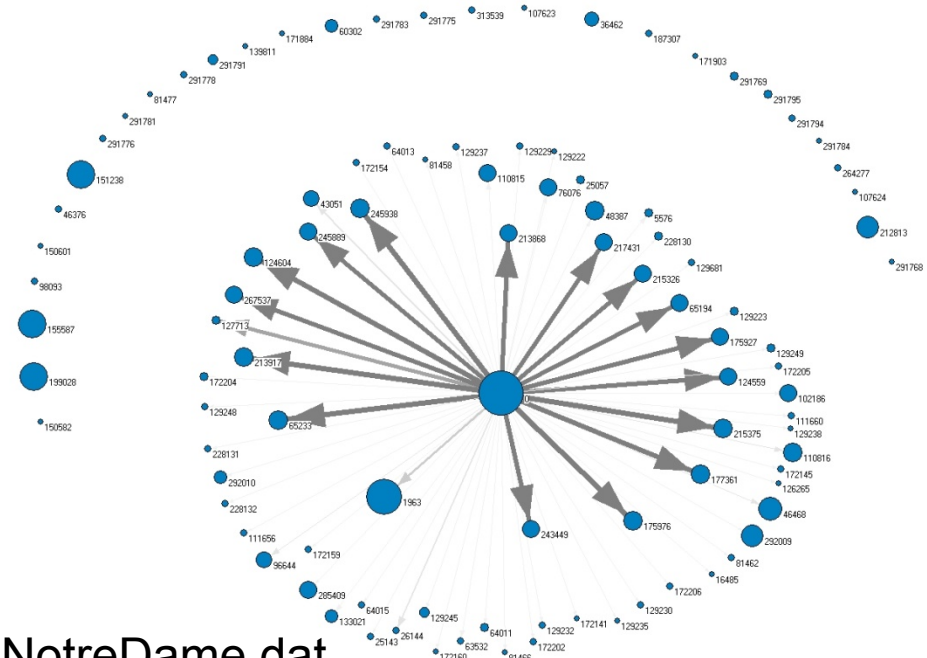


BerkStan.dat

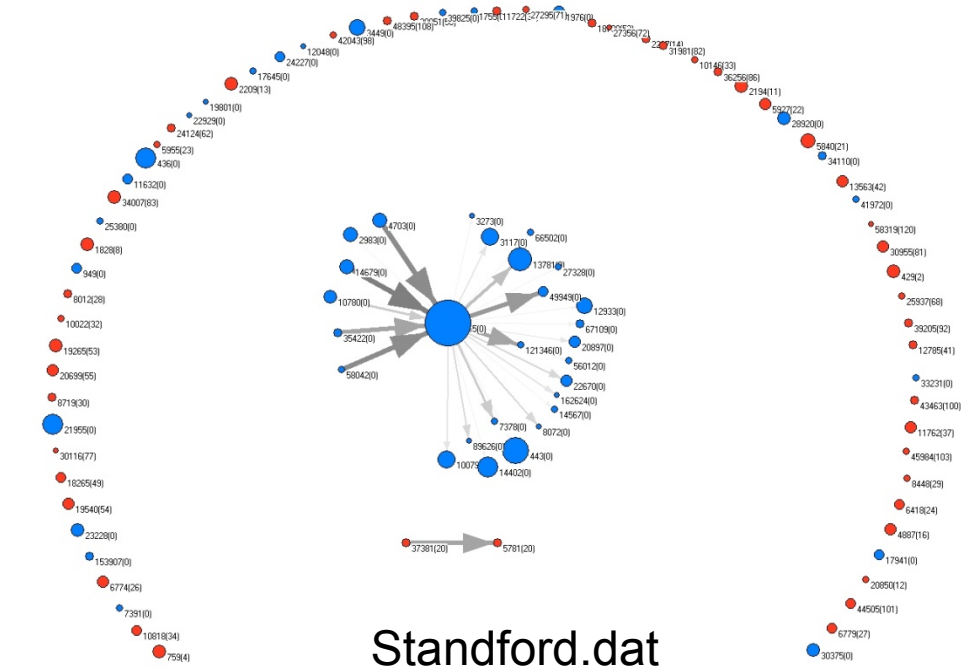


Google.dat



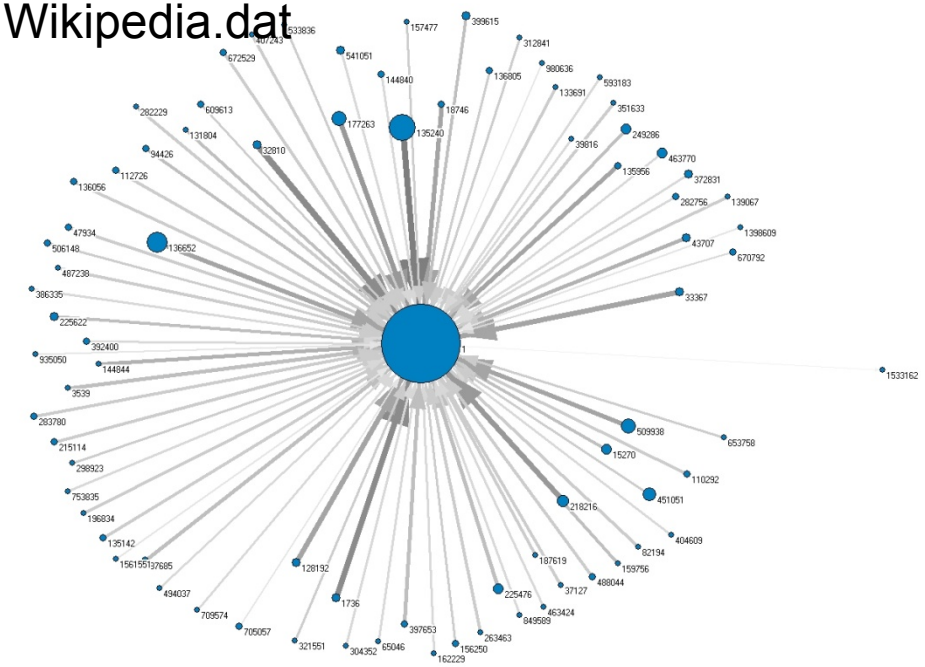


NotreDame.dat

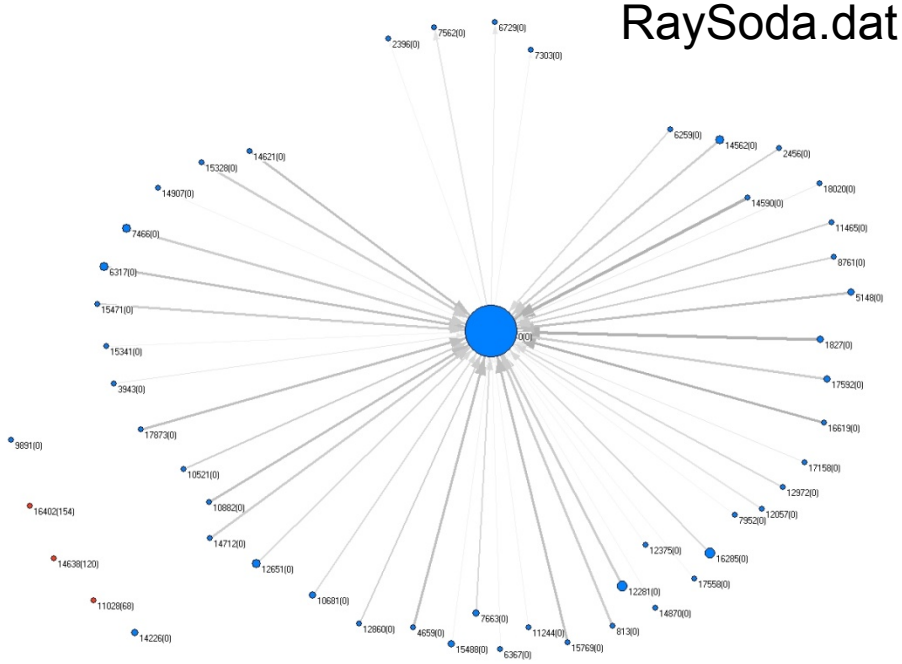


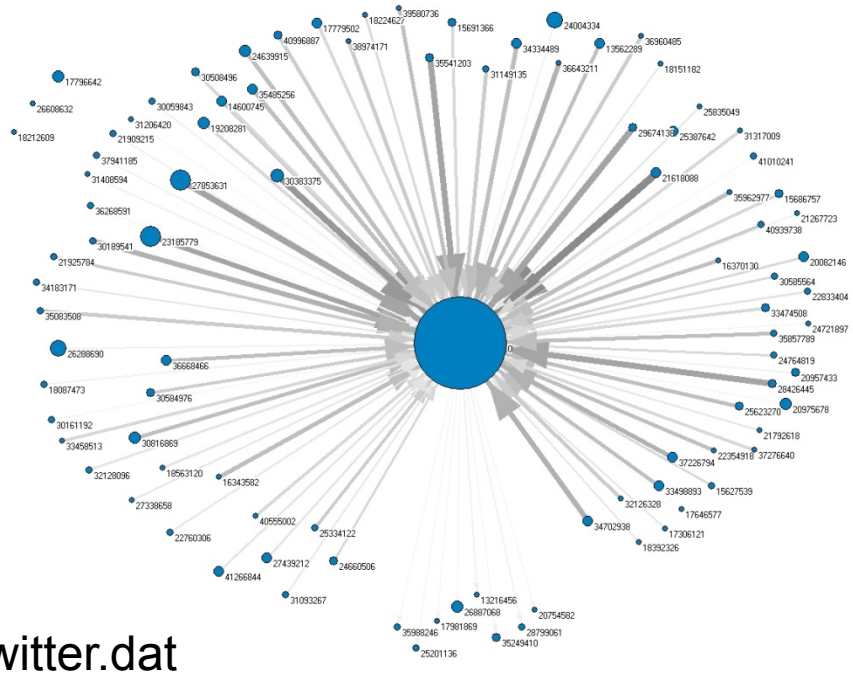
Standford.dat

Wikipedia.dat



RaySoda.dat

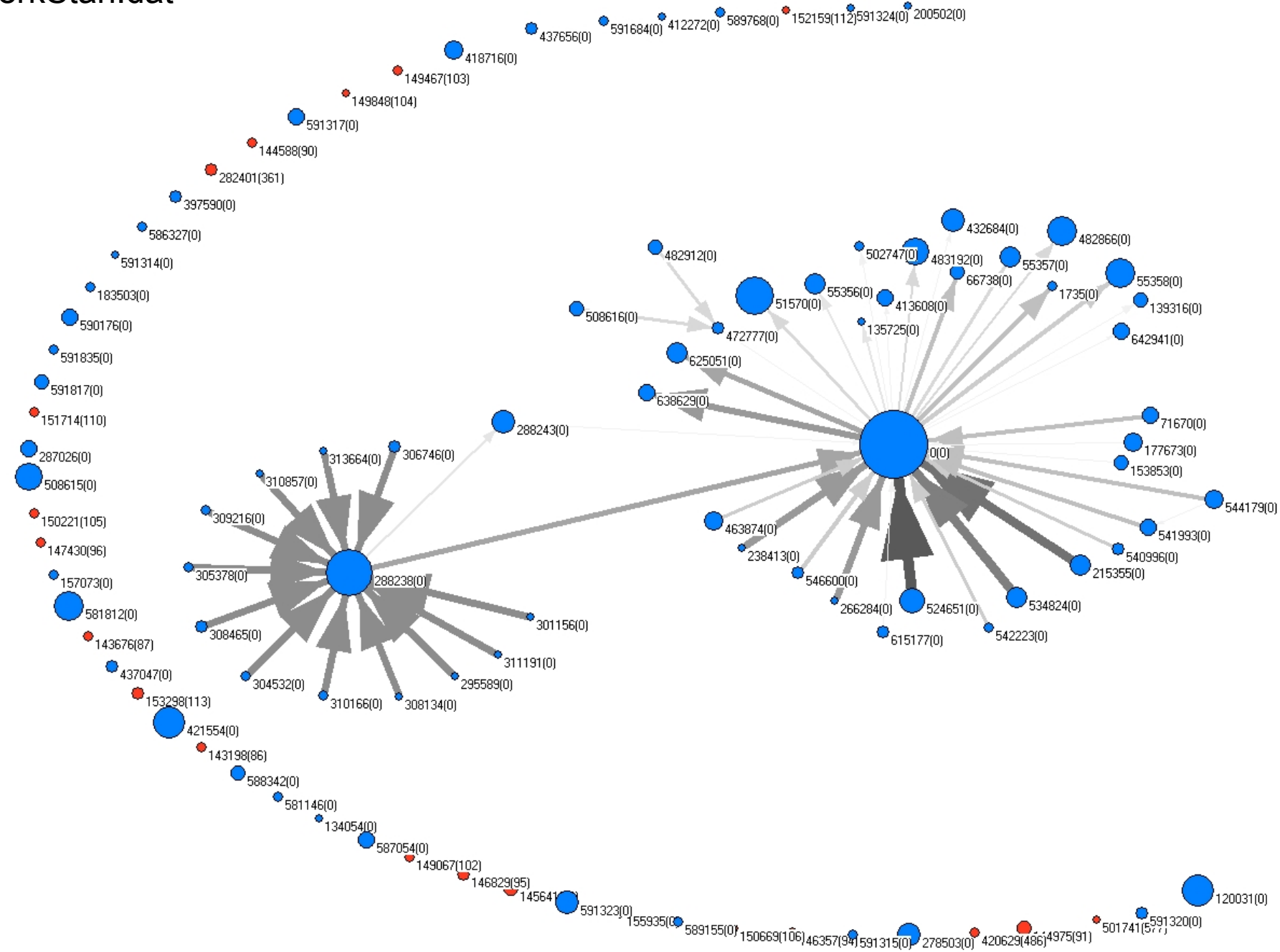




Twitter.dat

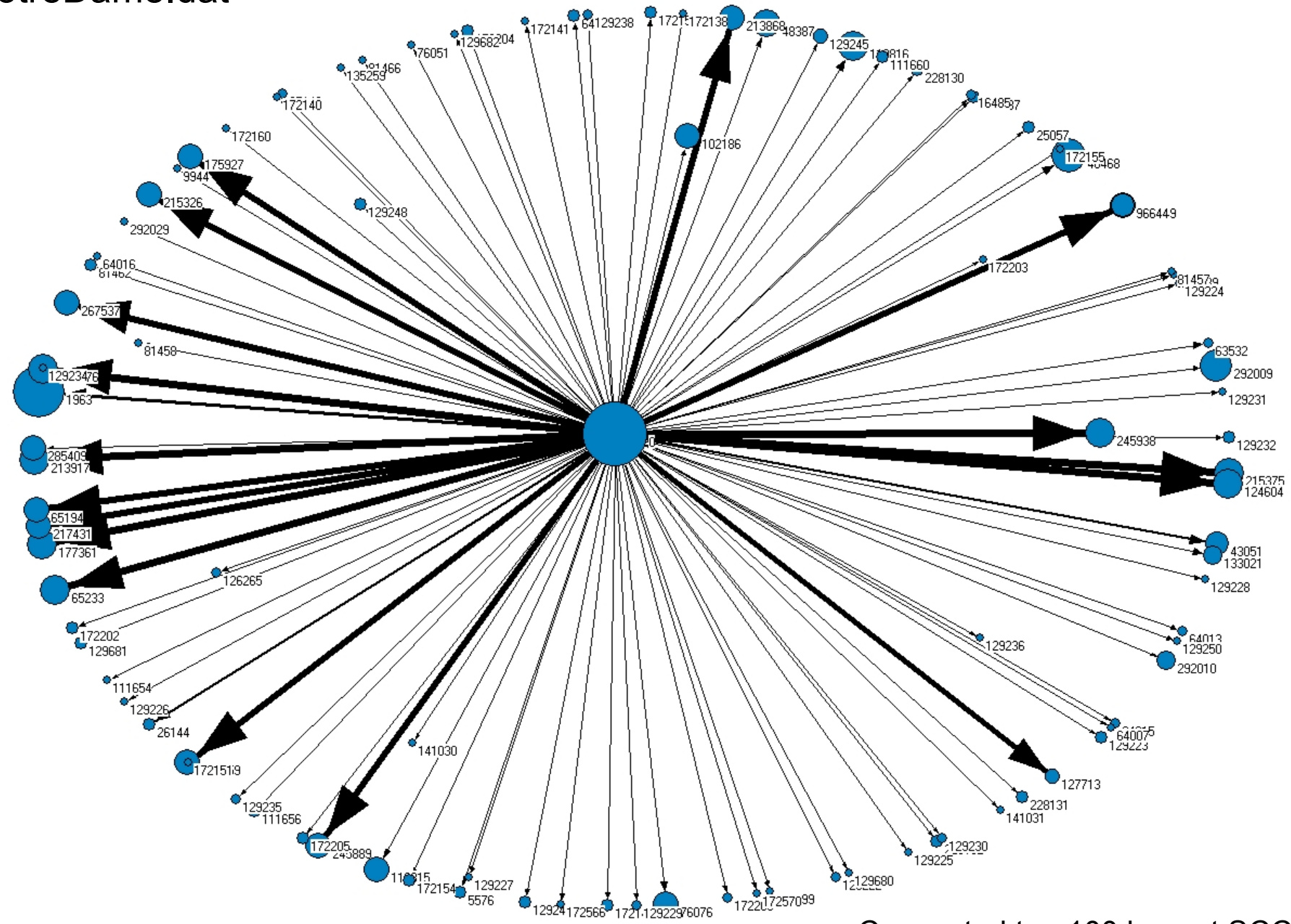
Top 100 largest SCC
and connections between them

BerkStan.dat



NotreDame.dat

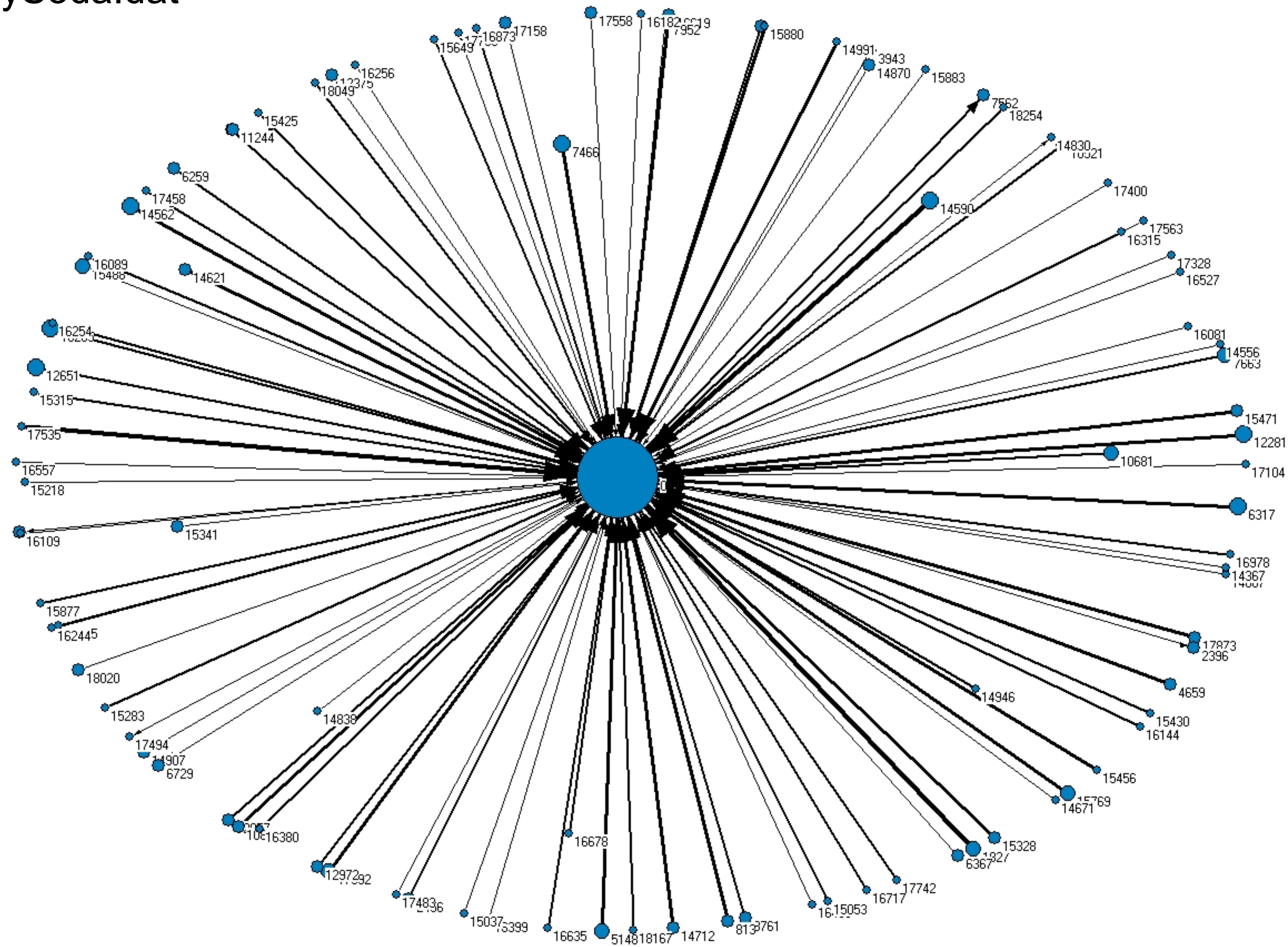
One GSCC and all SCC is OUT since it is a crawled data.



Connected top 100 largest SCC

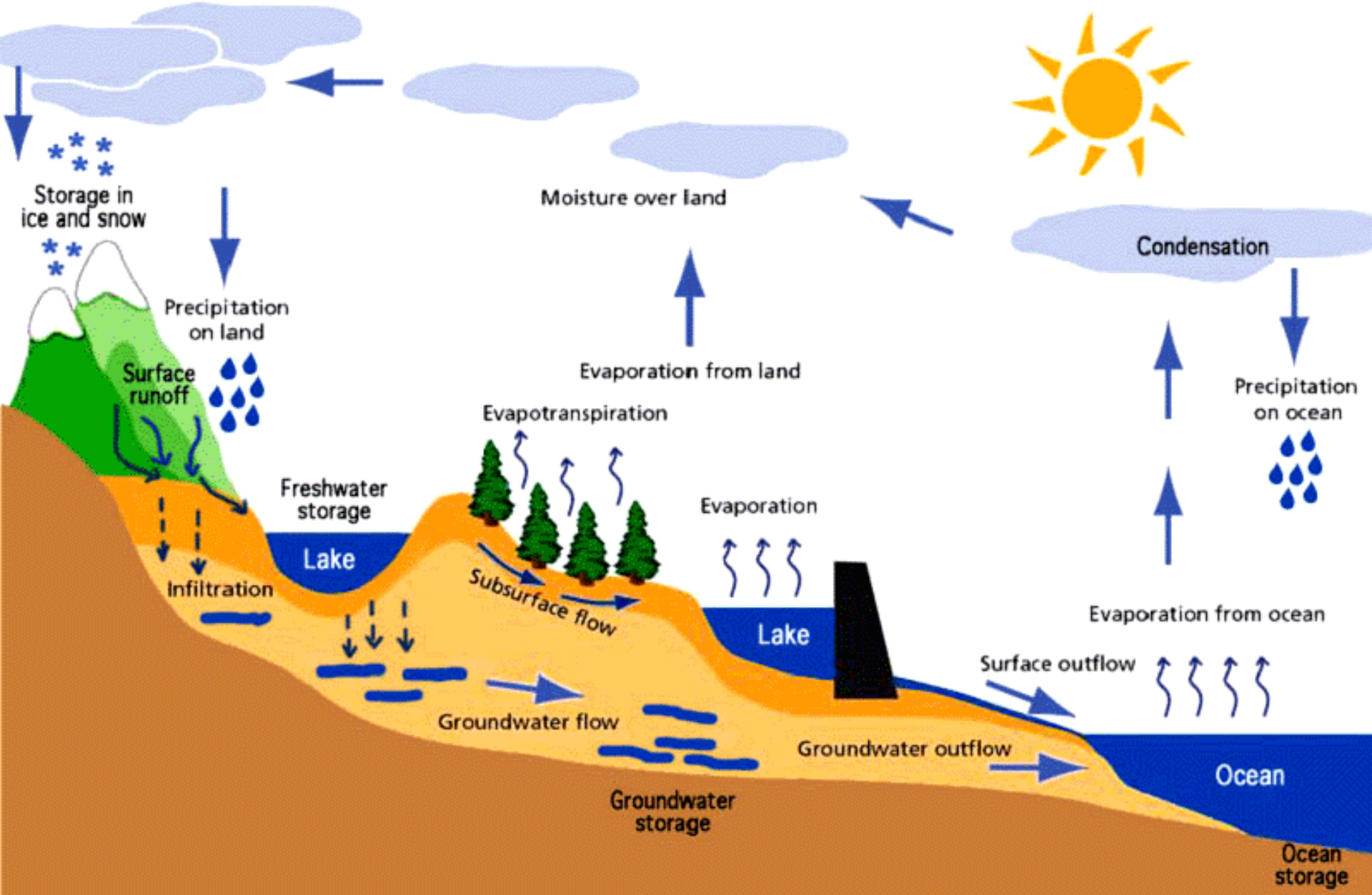
So, this definitely reflects there is serious sampling bias.

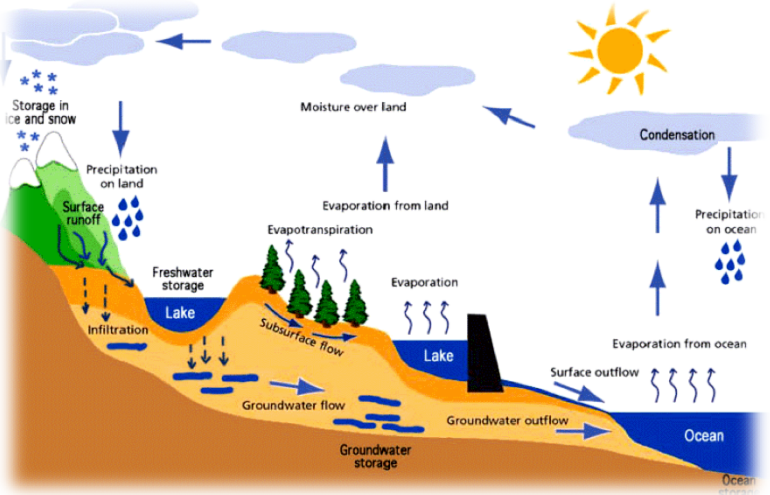
RaySoda.dat



17558
15880
17158
16873
15649
14991
3943
14870
15883
16256
18049
15425
11244
7466
18254
14830
17400
6259
17458
14562
14590
17563
16315
16089
14621
17328
16527
16254
16081
14556
12651
15315
15471
17535
12281
10681
16557
17104
15218
6317
16109
15341
16978
14367
15877
18020
17873
2396
4659
15283
14838
14946
15430
16144
17494
6729
15456
14671
16380
15328
16678
6367
12972
17742
16717
17483
15053
15037
16635
14712
3761

Hydrologic Cycle (물의 순환)





Evaporation & precipitation:
random teleportation

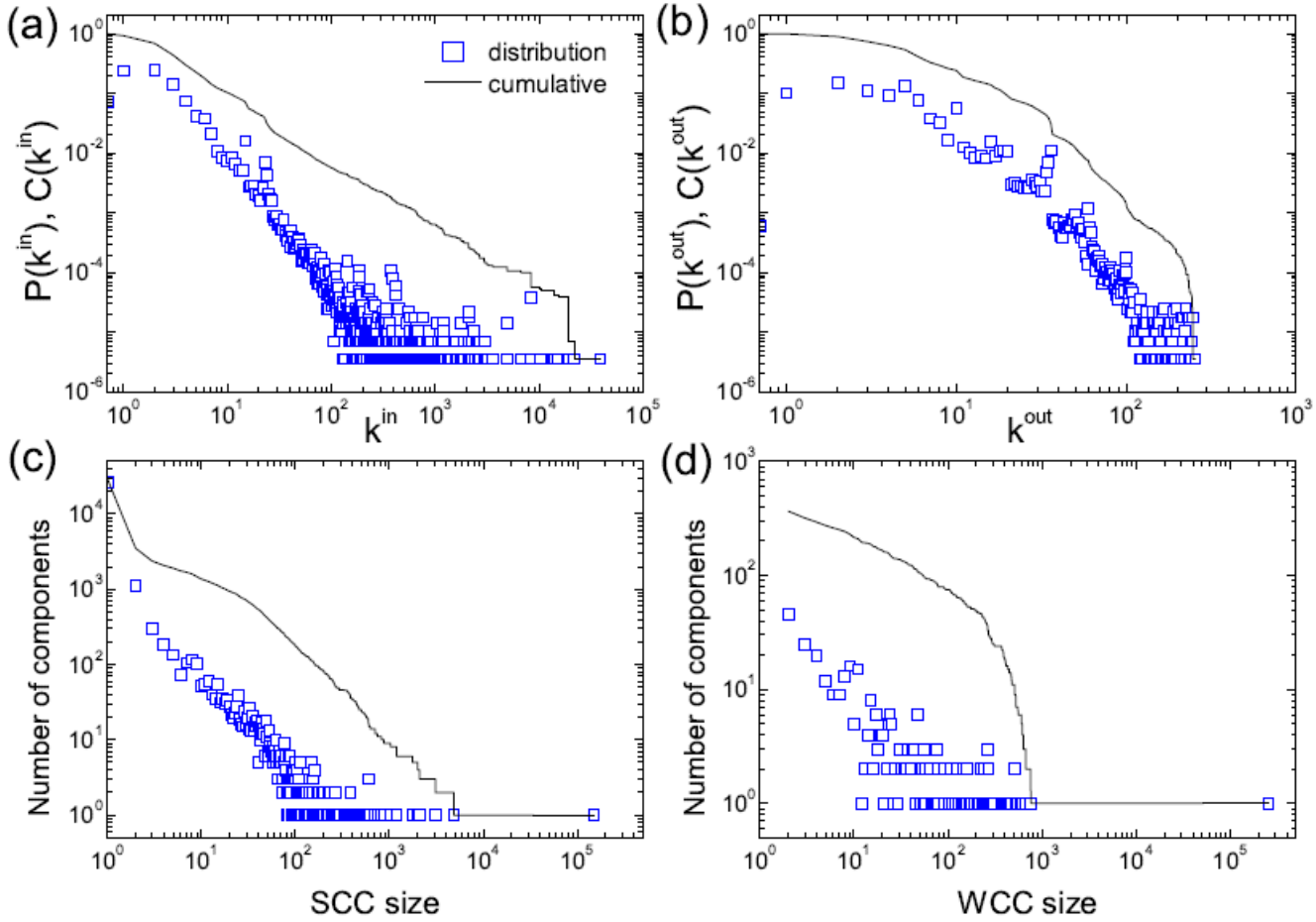
a Strongly Connected Component

Effect of a damping factor ?

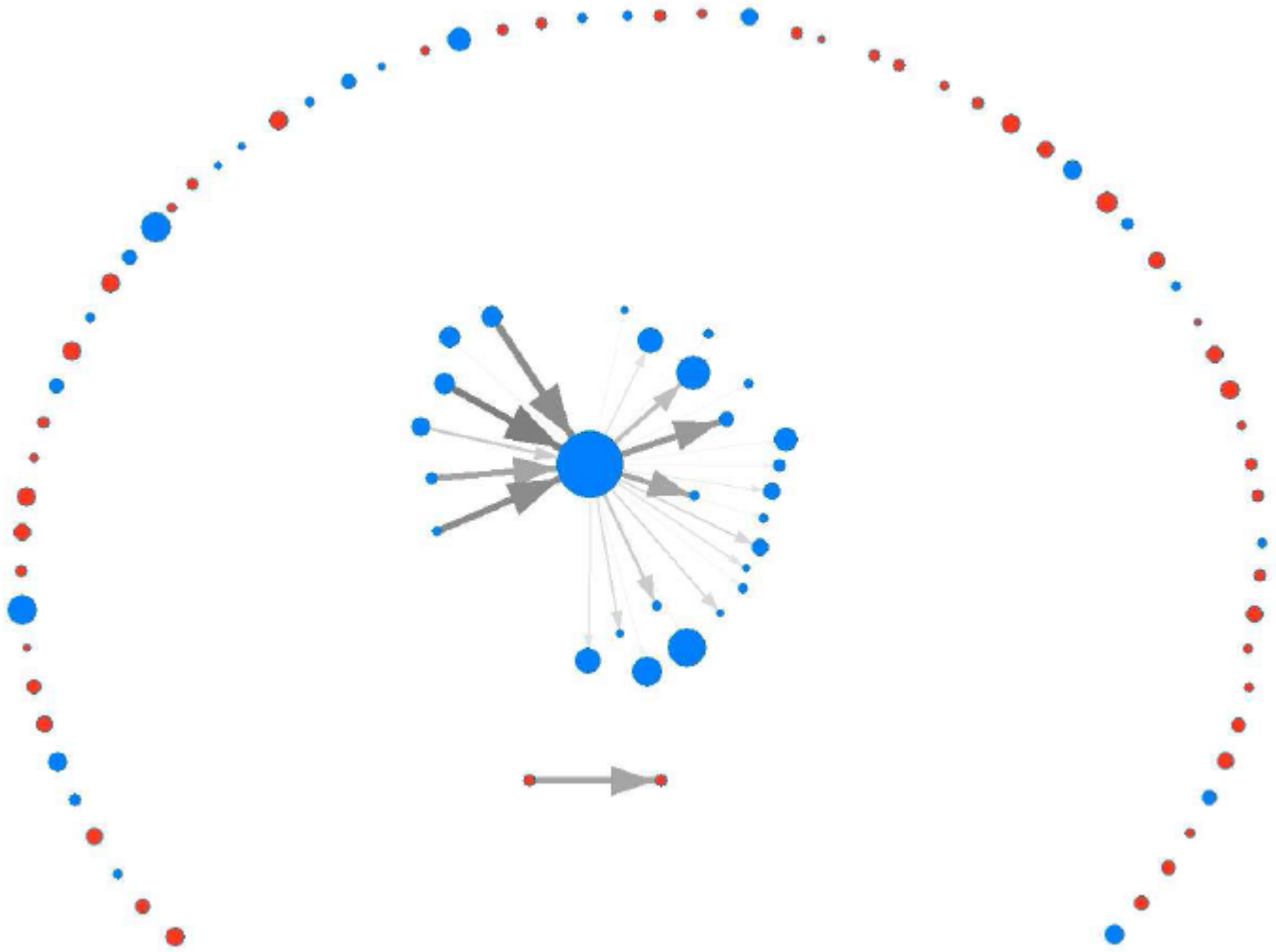
Stanford Web data

Table 1. Summary of the Stanford Web data.

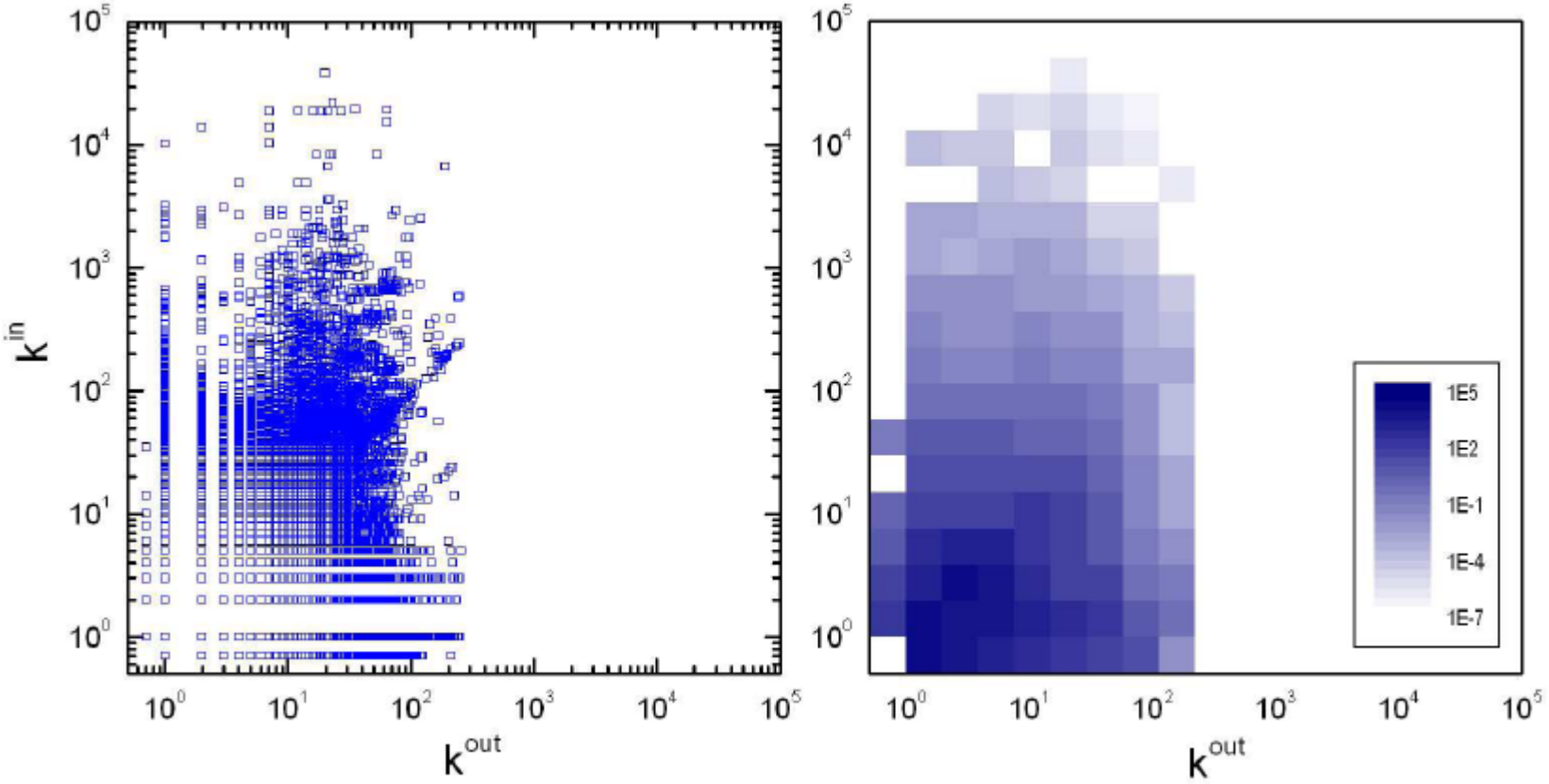
Number of nodes	Number of links	Average degree	Number of SCCs	Size of the giant SCC	Degree-degree auto-correlation
281,903	2,312,497	8.20	29,914	150,532 (0.534)	0.047



SCC decomposition diagram



Degree correlation



Pearson correlation : 0.047
Spearman correlation: 0.258
Kendall correlation: 0.206

Correlation coefficients

1. Pearson correlation

$$r = \frac{\sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N} (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^{N} (Y_i - \bar{Y})^2}}$$

2. Spearman rank correlation

$$r_S = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{N} (y_i - \bar{y})^2}} = 1 - \frac{6 \sum_{i=1}^{N} (x_i - y_i)^2}{N(N^2 - 1)}$$

3. Kendall's tau rank correlation

$$\tau = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} \operatorname{sgn}\left[(x_i - x_j)(y_i - y_j)\right]}{N(N-1)}$$

$\pi(d)$

How the PRs look like…

Rank-reversal

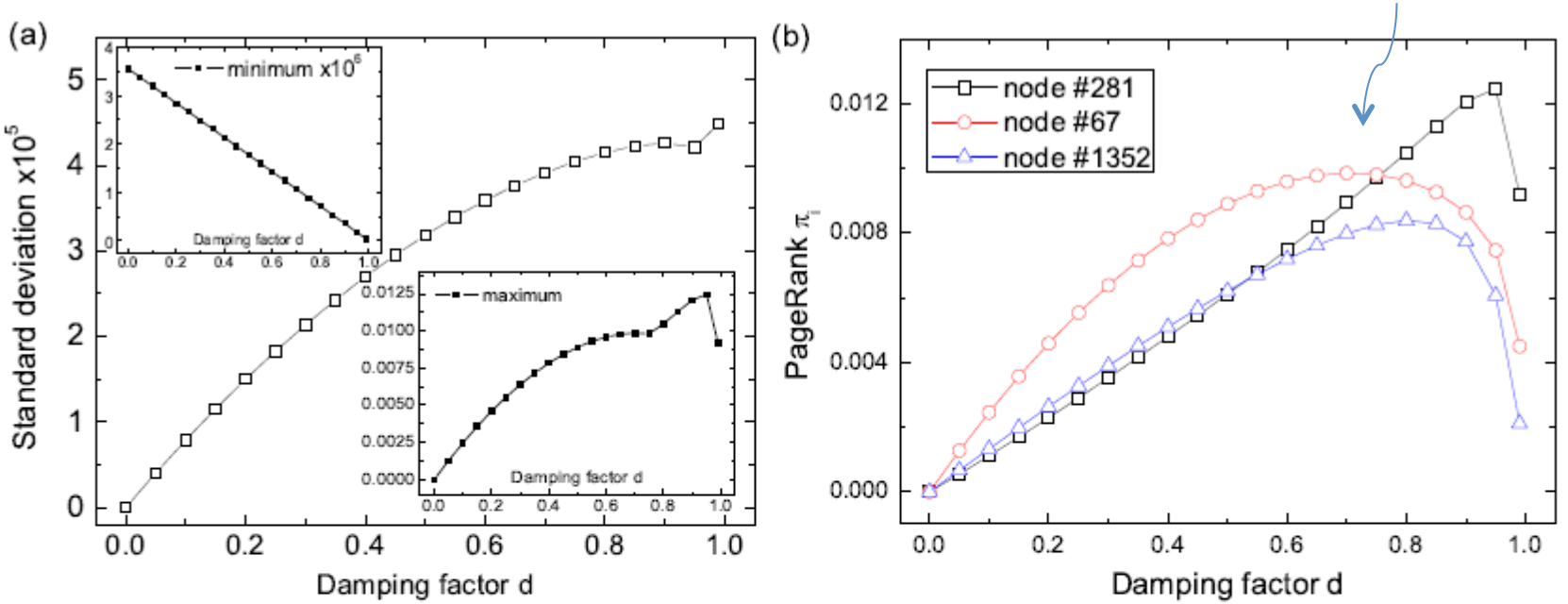


Figure 2. (Color online) (a) Standard deviation of the PR of the Stanford network, along with its minimum and maximum. The standard deviation of the PR gradually increases as the damping factor increases. In the insets: the minimum of the PR follows a trivial linear relation, while the maximum is nontrivially correlated with the value of the damping factor. (b) PR values of the three nodes having the highest PR are traced as the value of the damping factor is changed. Rank reversals occur around $d = 0.55$ and $d = 0.75$.

Relation between in-degree and PR

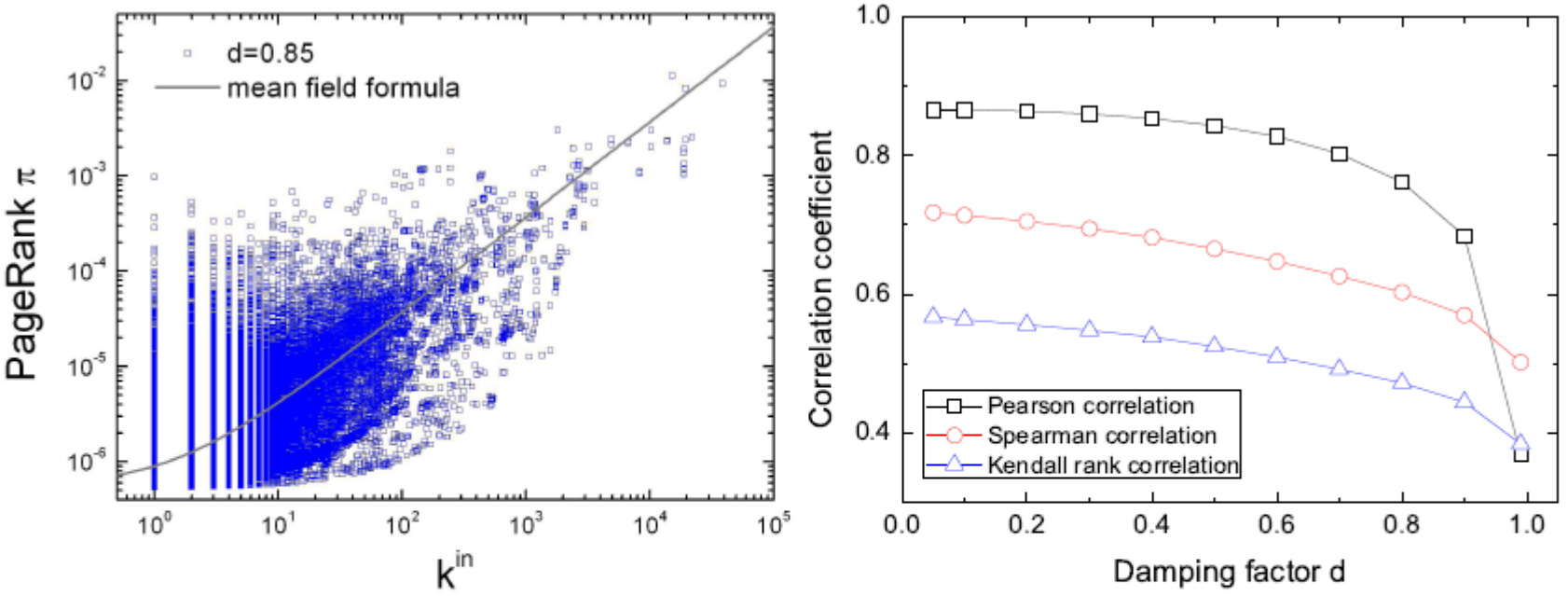
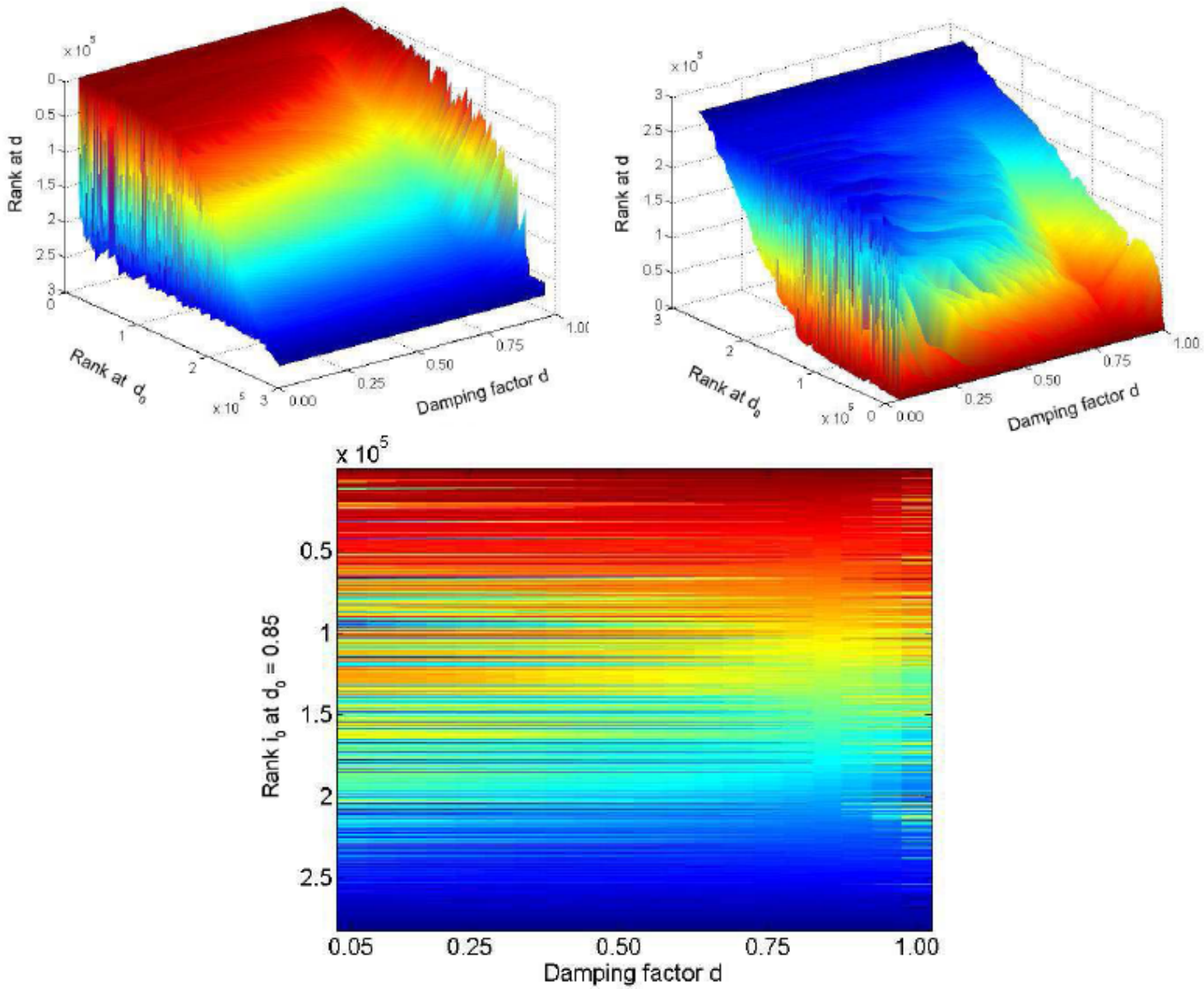


Figure 3. (Color online) The relationship between incoming degree and PR. The left panel clearly shows a positive correlation between the incoming degree and PR for $d_0 = 0.85$. It agrees well with the mean field result of Ref. [27], indicated by the solid line. The right-side panel shows the three different correlation coefficients between incoming degree and PR at different values of d. The correlation increases as the damping factor decreases.

S. Fortunato, M. Boguna, A. Flammini, and F. Menczer, "**On Local Estimations of PageRank: A Mean Field Approach**," Internet Mathematics **4**, 245 (2007).

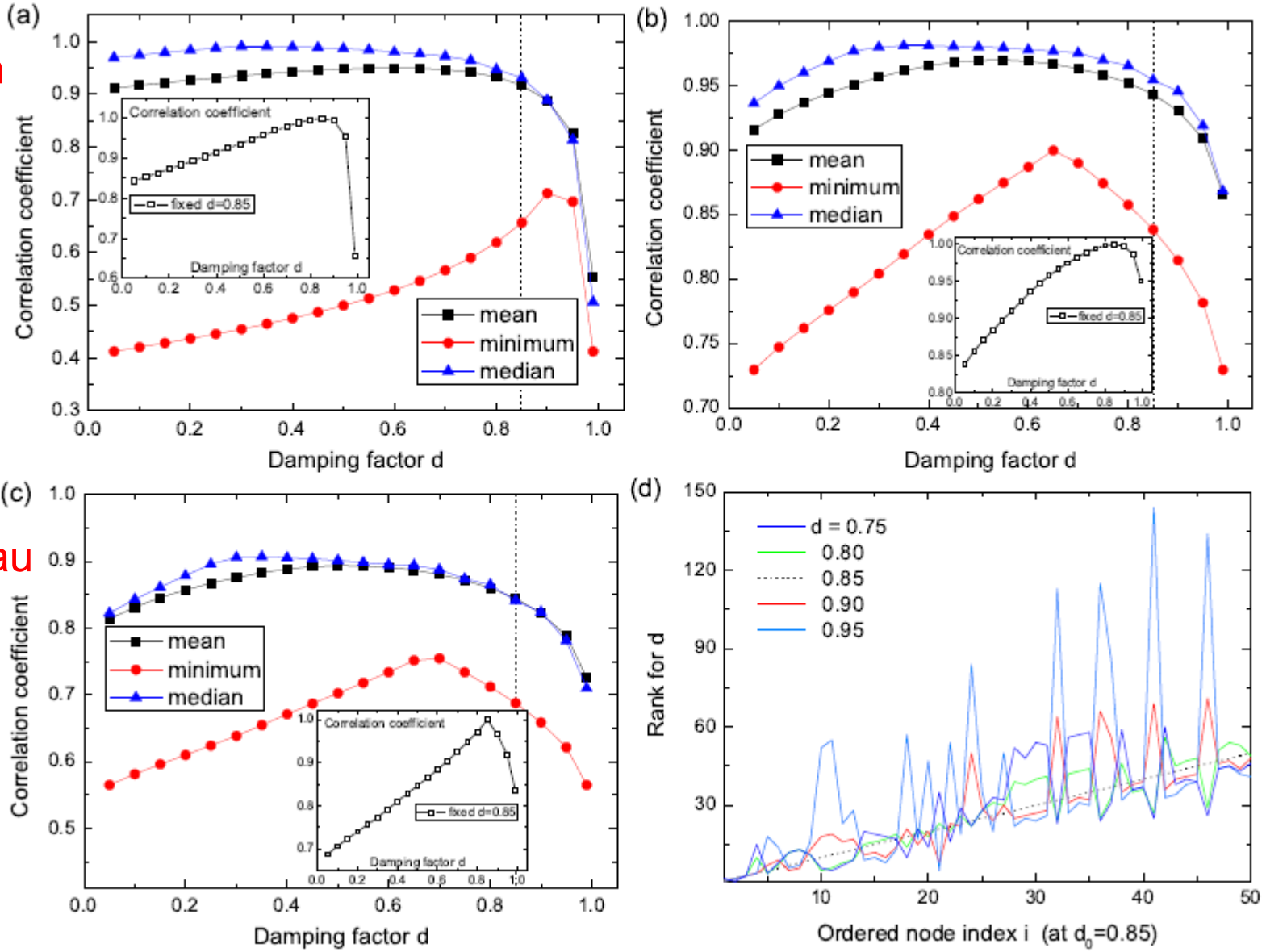
Rank changes depending on damping factor d



PageRank correlations

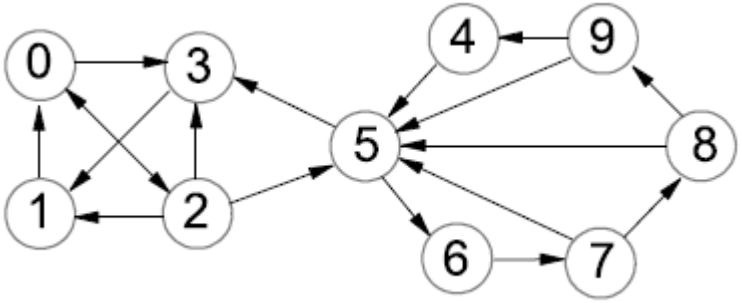
Spearman

Pearson



Kendall's tau

On a single SCC ?



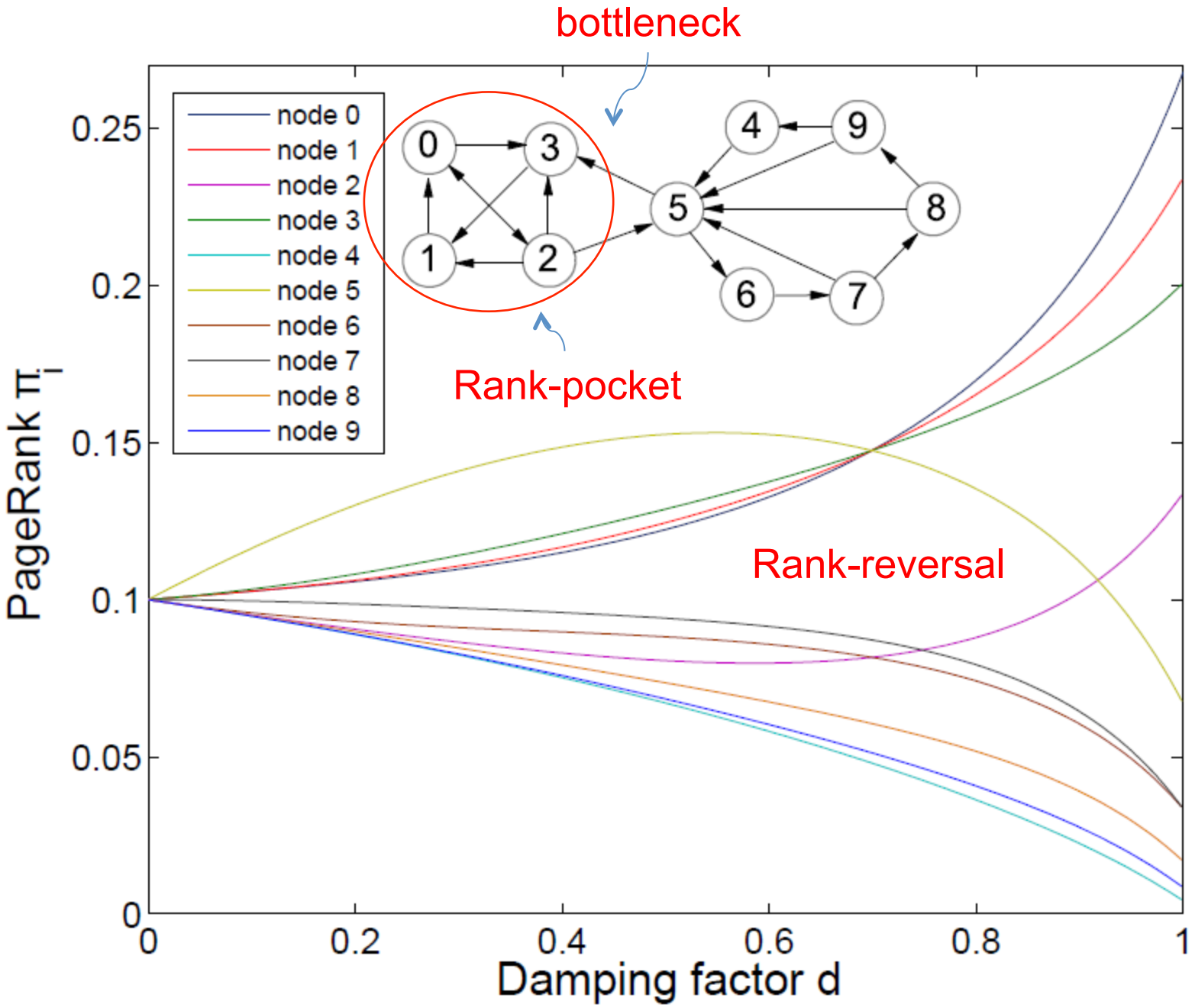
$$\pi_0 = d\left(\pi_1 + \frac{\pi_2}{4}\right) + \frac{1-d}{10}, \qquad \pi_1 = d\left(\pi_3 + \frac{\pi_2}{4}\right) + \frac{1-d}{10},$$

$$\pi_2 = d\,\frac{\pi_0}{2} + \frac{1-d}{10}, \qquad \pi_3 = d\left(\frac{\pi_0}{2} + \frac{\pi_2}{4} + \frac{\pi_5}{2}\right) + \frac{1-d}{10},$$

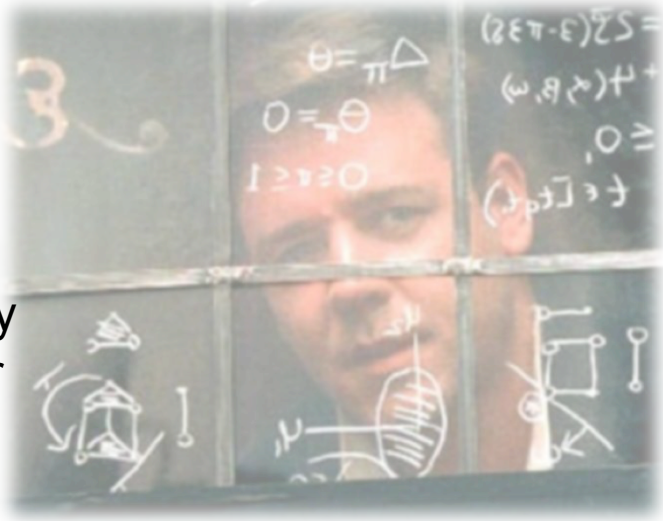
$$\pi_4 = d\,\frac{\pi_9}{2} + \frac{1-d}{10}, \qquad \pi_5 = d\left(\pi_4 + \frac{\pi_2}{4} + \frac{\pi_7}{2} + \frac{\pi_8}{2} + \frac{\pi_9}{2}\right) + \frac{1-d}{10},$$

$$\pi_6 = d\,\frac{\pi_5}{2} + \frac{1-d}{10}, \qquad \pi_7 = d\,\pi_6 + \frac{1-d}{10},$$

$$\pi_8 = d\,\frac{\pi_7}{2} + \frac{1-d}{10}, \qquad \pi_9 = d\,\frac{\pi_8}{2} + \frac{1-d}{10}.$$


bottleneck
Rank-pocket
Rank-reversal
node 0
node 1
node 2
node 3
node 4
node 5
node 6
node 7
node 8
node 9
0
1
2
3
4
5
6
7
8
9
PageRank π_i
Damping factor d
0
0.05
0.1
0.15
0.2
0.25
0
0.2
0.4
0.6
0.8
1

Summary



- ✓ These days Google's PageRank (PR) is deeply related to the success of modern businesses or the ranking of athletes, scientists, their paper, and even scientific journals.
- ✓ We have investigated PR as a function of its damping factor on a subset of pages from a single domain in WWW.
- ✓ Rank-reversal occurs frequently and over a broad range of PR.
- ✓ Rank-reversal happens not only in directed networks containing rank-sinks but also in a single strongly connected component. This is due to the presence of rank-pockets and bottlenecks.
- ✓ A better understanding rank-reversals may be essential to optimizing the stability of PR, to thwarting attempts to cheat such as spam farms, and ultimately to determining which scientists will be cited, which products will sell , and which businesses or other ventures will prosper.

To get more eigenfactor…

- S.-W. Son, B. J. Kim, H. Hong, and H. Jeong, "**Dynamics and Directionality in Complex Networks**", Phys. Rev. Lett. **103**, 228702 (2009).
- Y. Kim, S.-W. Son, and H. Jeong, "**Link Rank: Finding Communities in Directed Networks**", Phys. Rev. E **81**, 016103 (2010).
- A. Zeng, S.-W. Son, C. H. Yeung, Y. Fan, and Z. Di, "**Enhancing synchronization by directionality in complex networks**", Phys. Rev. E **83**, 045101(R) (2011).
- S.-W. Son, C. Christensen, G. Bizhani, D. V. Foster, P. Grassberger, and M. Paczuski, "**Sampling properties of directed networks**", Phys. Rev. E **86**, 046104 (2012).
- S.-W. Son, C. Christensen, P. Grassberger, and M. Paczuski, "**PageRank and rank-reversal dependence on the damping factor**", Phys. Rev. E **86**, 066104 (2012).