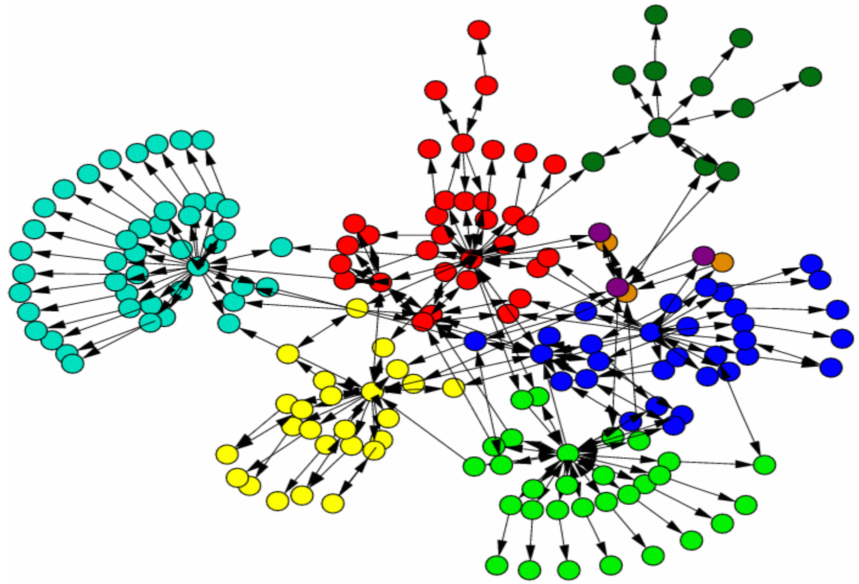
Optimal ranking in networks with community structure

Huafeng Xie
Dept of Physics, Brookhaven National Laboratory
New Media Lab, The Grad. Center, CUNY

NetSci 2006, Bloomington IN USA

World Wide Web

- Nodes (Vertices): Web pages in the WWW
- Links: Hyperlinks on the web pages
- Large size: $N \sim 10^{10}$
- Heterogeneous: community structure



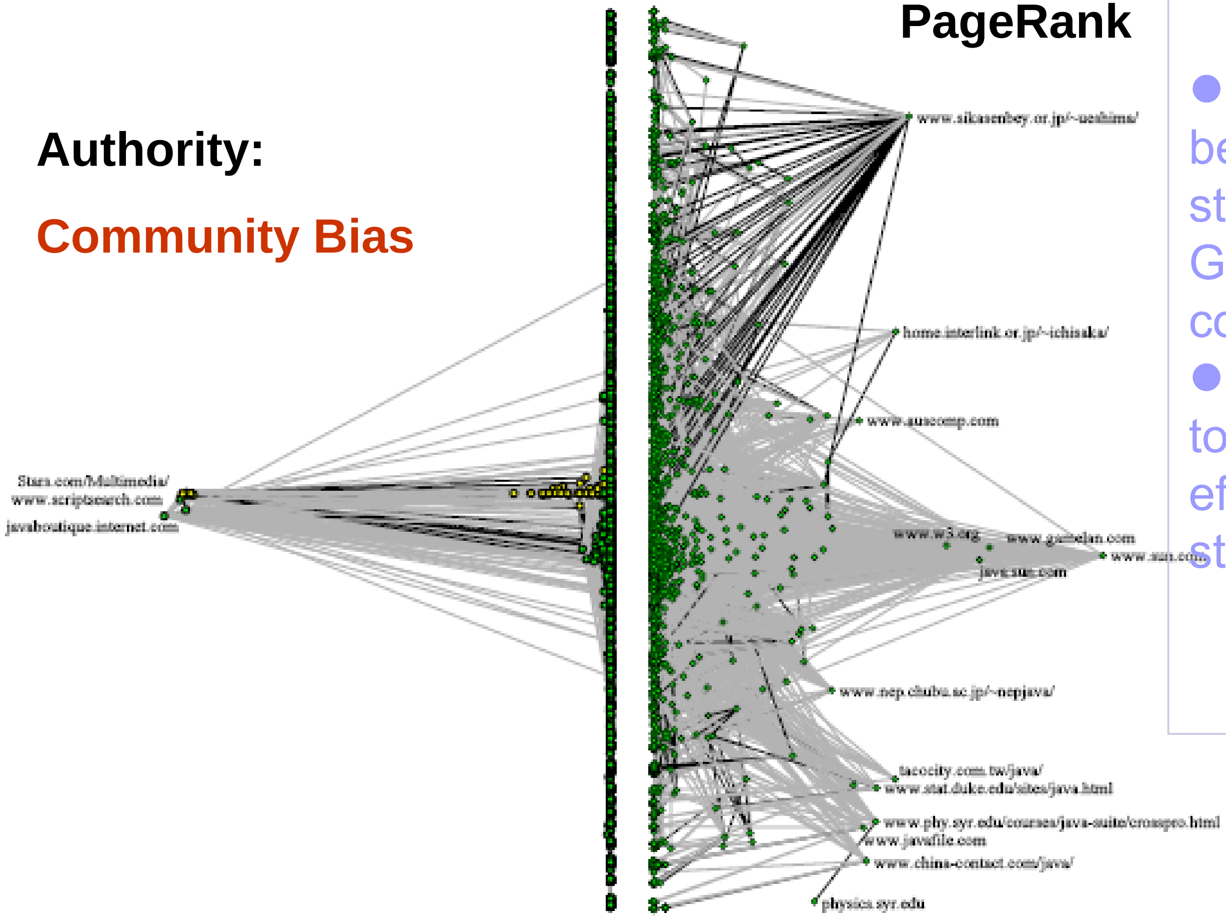
M. E. J. Newman and M. Girvan, Physical Review E **69**, 026113 (2004)

Objective

PageRank

Authority:

Community Bias

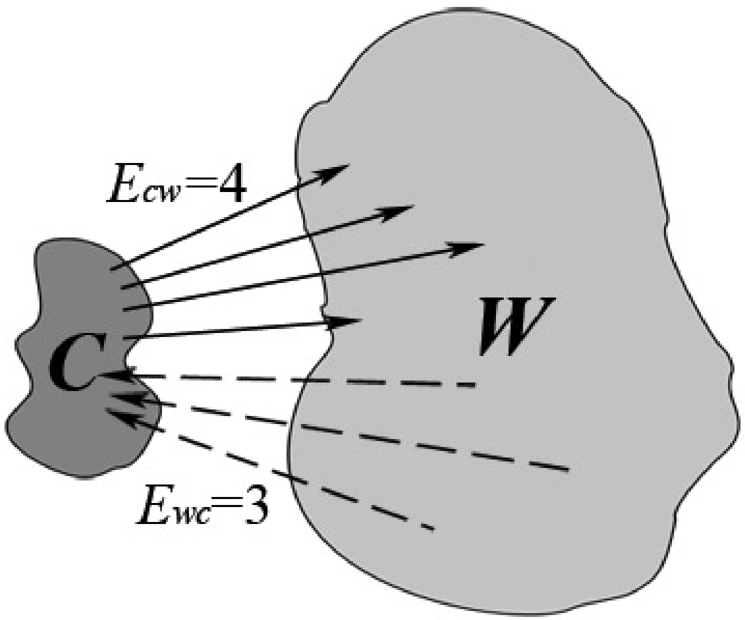


- Understand the interplay between the community structure and the average Google rank inside the community.
- As a search engine, how to reduce the undesired effects of community structure.

Authority and PageRank visualization of “java” query result

U. Brandes and S. Cornelsen, *Visual Ranking*, JGAA, 7(2) 181–201 (2003) 194

Definitions



- E_{cw} , E_{wc},
- N_c , N_w
- $<K_{in}>_c$, $<K_{out}>_c$
- $<K_{in}>_w$, $<K_{out}>_w$
- $1 << N_c << N_w$

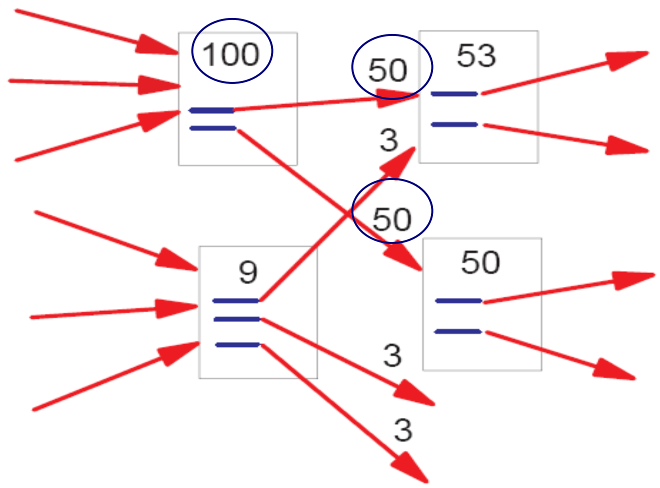
- G_c : the average Google rank value of the community member nodes
- G_w : the average Google rank value of the outside world
- Effects of community on G_c is only determined by E_{cw}, E_{wc} and a parameter in PageRank α.
- What's the optimal value for α

Google PageRank Algorithm

•***PageRank***: simulates random walks on the web

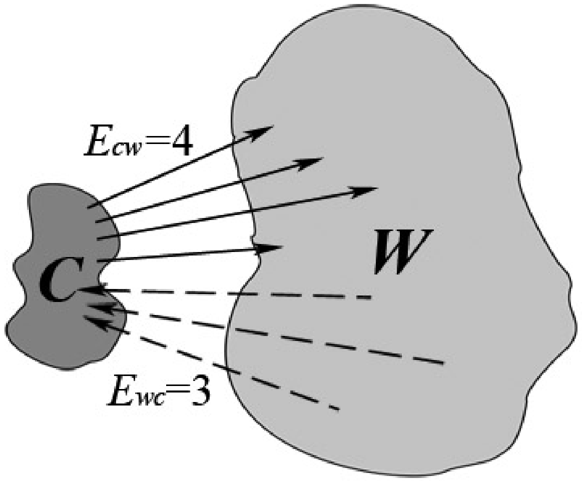
•***Rank Value*** of a node *i* is proportional to the number of random walkers on this node at stationary state

$$G(i) = \alpha + \sum_{j \to i} (1-\alpha) \frac{G(j)}{K_{out}(j)}.$$



L. Page, S. Brin, R. Motwani and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web" Stanford Digital Library Technologies Project (1998).

Effects of Community Structure



Mean-field assumption:

the average Google ranks and out-degrees of community nodes sending links to the outside world are equal to the overall average values inside the community G_c. Assume the same for node sending links from the outside world to the community.

$J_{cw}=(1-\alpha)G_c E_{cw}/\langle K_{out}\rangle_c+\alpha G_c N_c$
$J_{wc}=(1-\alpha)G_w E_{wc}/\langle K_{out}\rangle_w+\alpha G_w N_c$
$J_{cw}=J_{wc}$

$$\frac{G_c}{G_w}=\frac{(1-\alpha)\dfrac{E_{wc}}{\langle K_{out}\rangle_w N_c}+\alpha}{(1-\alpha)\dfrac{E_{cw}}{\langle K_{out}\rangle_c N_c}+\alpha}$$

$$G_w \approx 1$$

Main Equation

In **random networks** with the same degree sequences,

Expected number of links from the outside world to the community:

$E_{wc}^{(r)} = <K_{out}>_w N_c$

Expected number of links from the community to outside world:

$E_{cw}^{(r)} = <K_{out}>_c N_c N_w / (N_c + N_w) = <K_{out}>_c N_c$

$$G_c = \frac{(1-\alpha)\frac{E_{wc}}{E_{wc}^{(r)}} + \alpha}{(1-\alpha)\frac{E_{cw}}{E_{cw}^{(r)}} + \alpha}.$$

$$R_{wc} = \frac{E_{wc}}{E_{wc}^{(r)}}$$

$$R_{cw} = \frac{E_{cw}}{E_{cw}^{(r)}}$$

Provided that our mean-field assumption is valid

$$G_c = \frac{(1-\alpha)R_{wc} + \alpha}{(1-\alpha)R_{cw} + \alpha}$$

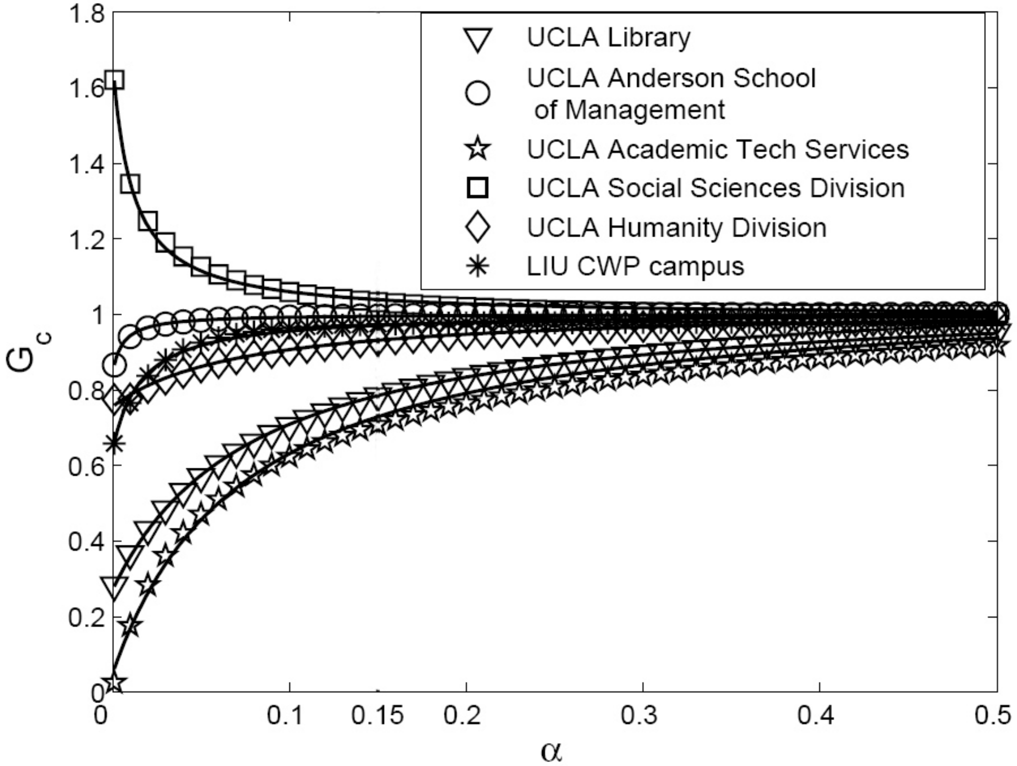
Empirical Study

UCLA: 31621 nodes, 353370 edges

LIU (Long Island University): 15471 nodes, 90111 edges

Community	N_c	E_{cc}	$E_{cc}^{(\tau)}$	E_{wc}	E_{cw}
UCLA Library	2028	23062	1699	755	2141
UCLA School of Management	1340	15983	739	175	169
UCLA Academic Tech. Services	1907	26597	2248	139	3113
UCLA Social Science Division	626	3986	50	258	142
UCLA Humanity Division	864	4846	79	397	445
LIU CWP Campus	2756	18376	4105	336	1393

Empirical Study



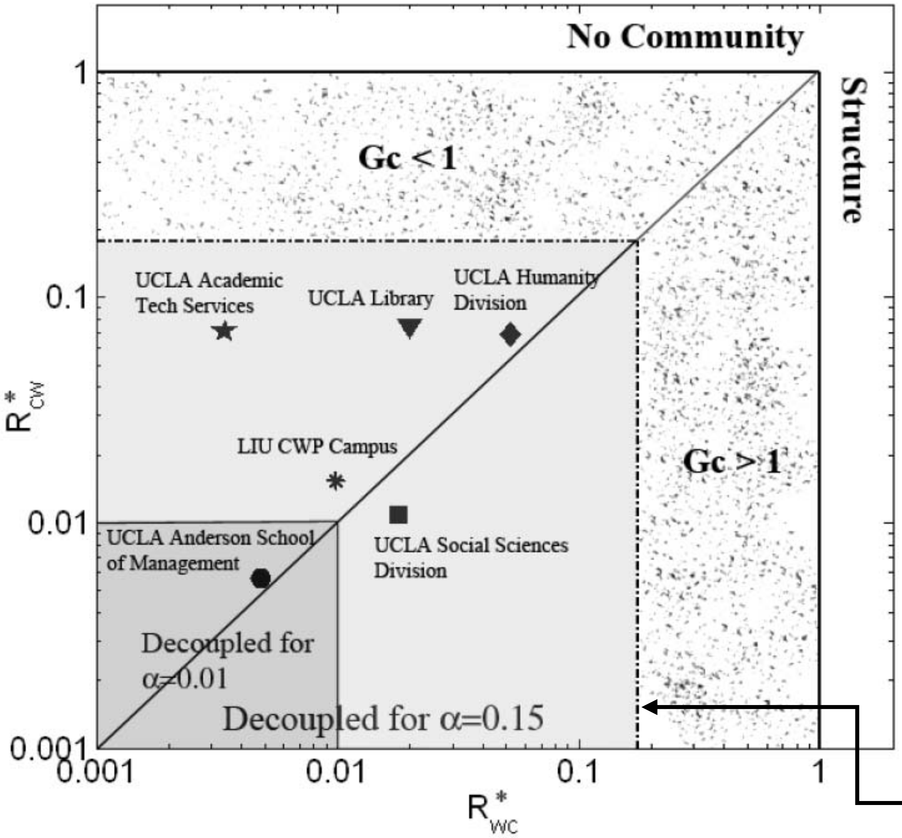
$$G_c = \frac{(1-\alpha)R_{wc} + \alpha}{(1-\alpha)R_{cw} + \alpha}$$

$$R_{wc} = \frac{E_{wc}}{E_{wc}^{(r)}}$$

$$R_{cw} = \frac{E_{cw}}{E_{cw}^{(r)}}$$

Community	R_{wc}	R_{cw}	R^*_{wc}	R^*_{cw}
UCLA Library	0.04	0.09	0.02	0.07
UCLA School of Management	0.01	0.01	0.005	0.006
UCLA Academic Tech. Services	0.007	0.1	0.003	0.07
UCLA Social Science Division	0.04	0.03	0.02	0.01
UCLA Humanity Division	0.04	0.08	0.05	0.07
LIU CWP Campus	0.03	0.09	0.01	0.02

Optimal α for PageRank



$$G_c = \frac{(1-\alpha)R^*_{wc} + \alpha}{(1-\alpha)R^*_{cw} + \alpha}$$

- *α should be as large as possible to avoid manipulations.*
- *α should be small enough to take into account network topology.*

Indeed Google Uses a good value of α, 0.15.

Acknowledgment

Sergei Maslov

- Department of Physics, Brookhaven National Laboratory

Koon-kiu Yan

- Department of Physics and Astronomy, Stony Brook University
- Department of Physics, Brookhaven National Laboratory

- Work at Brookhaven National Laboratory was carried out under Contract No. DE-AC02-98CH10886, Division of Material Science, U.S. Department of Energy.
- Supports from The New Media Lab at the Grad Center of CUNY.