



From the WWW topology to ranking superstability

Albert-László Barabási

Center for Complex Networks Research

Northeastern University

Department of Medicine and CCSB

Harvard Medical School

www.BarabasiLab.com

World Wide Web

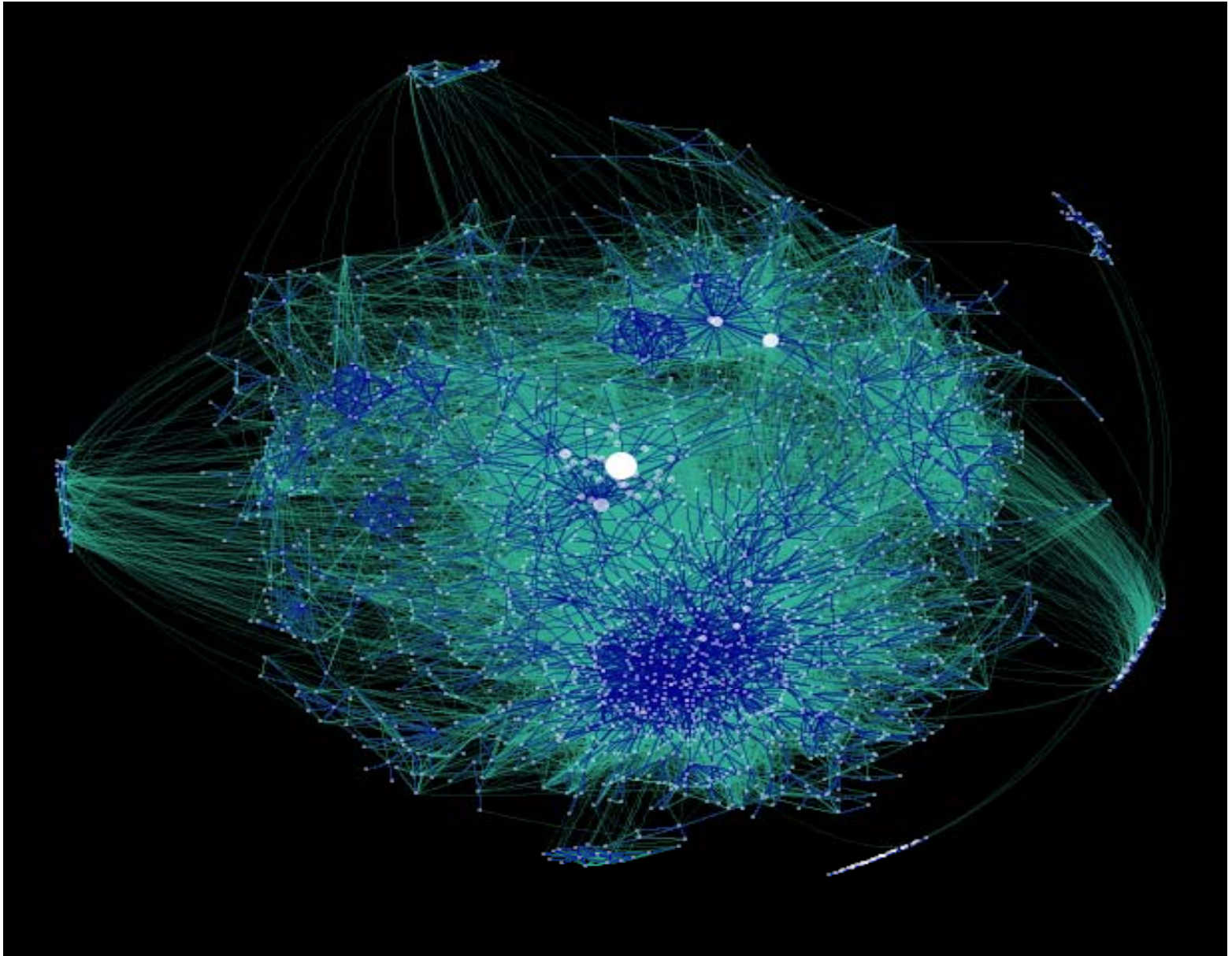
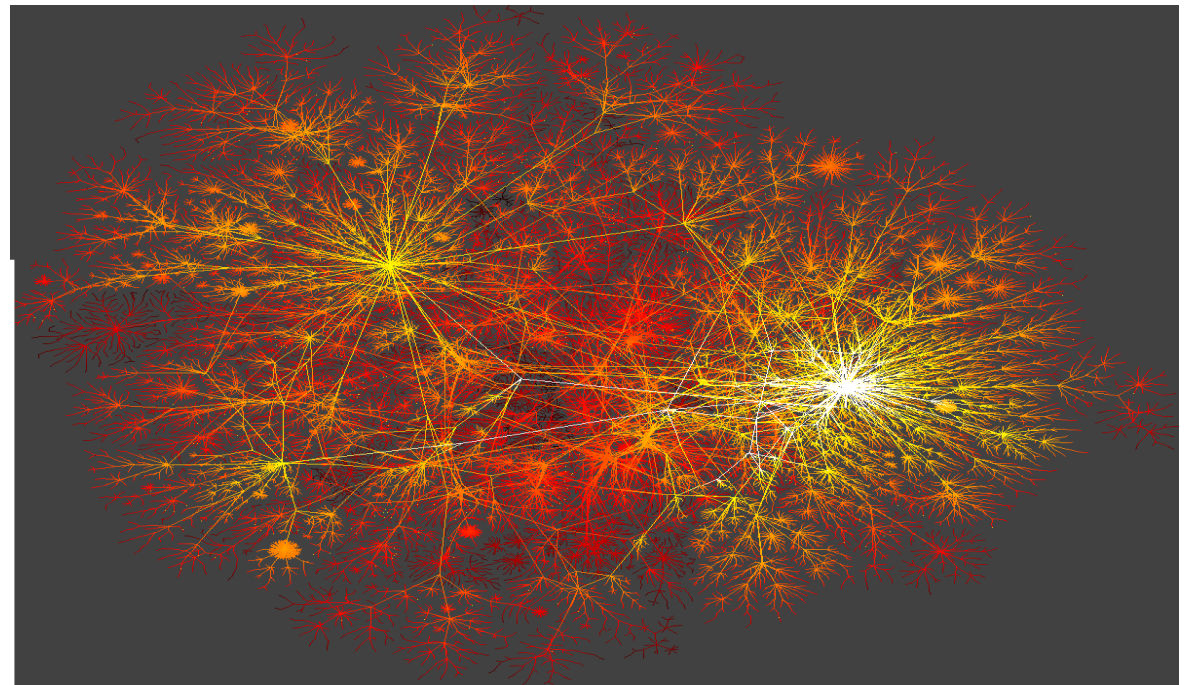
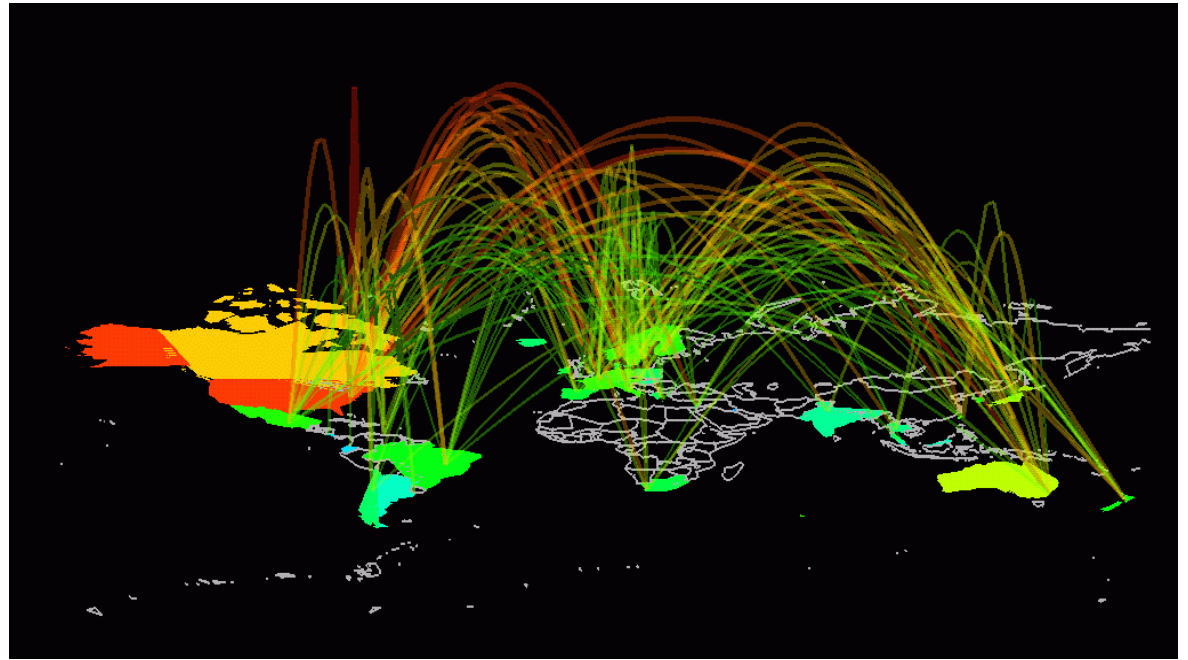
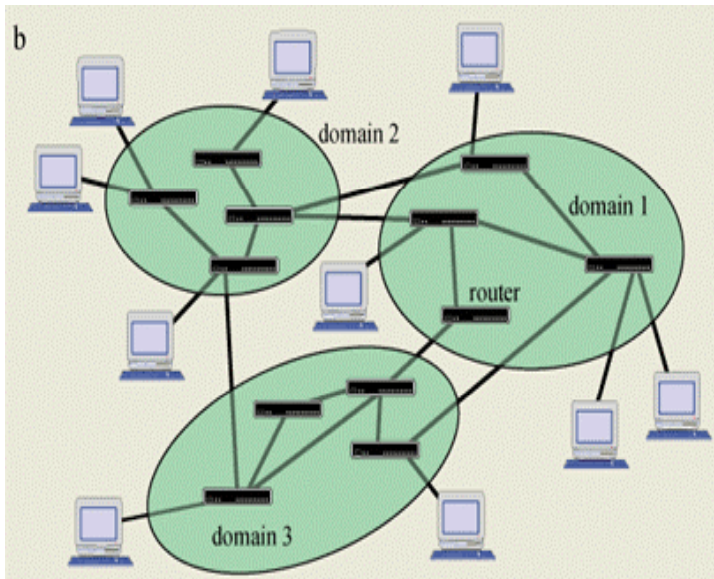
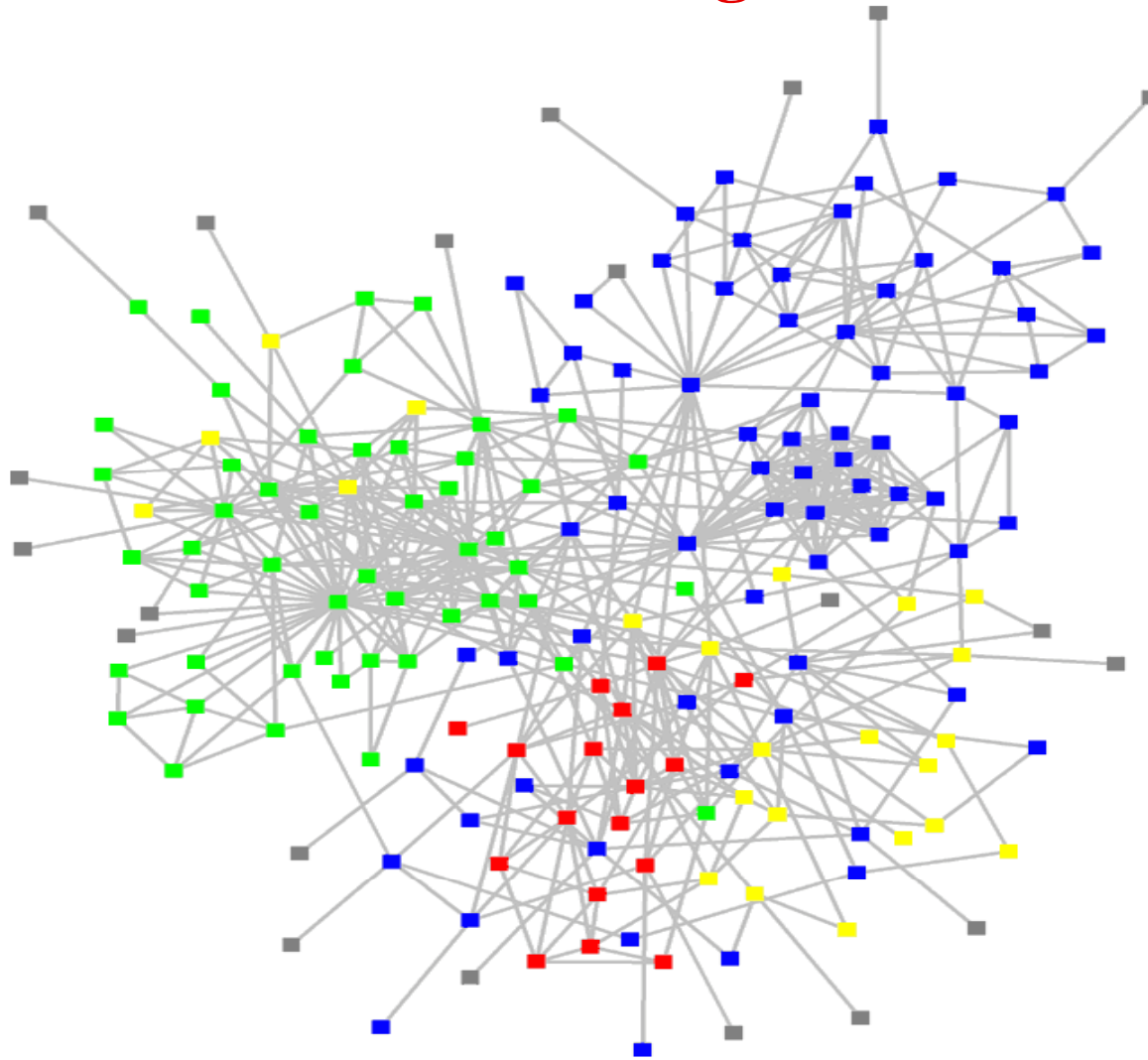


Image by Matthew Hurst

Internet



Structure of an organization

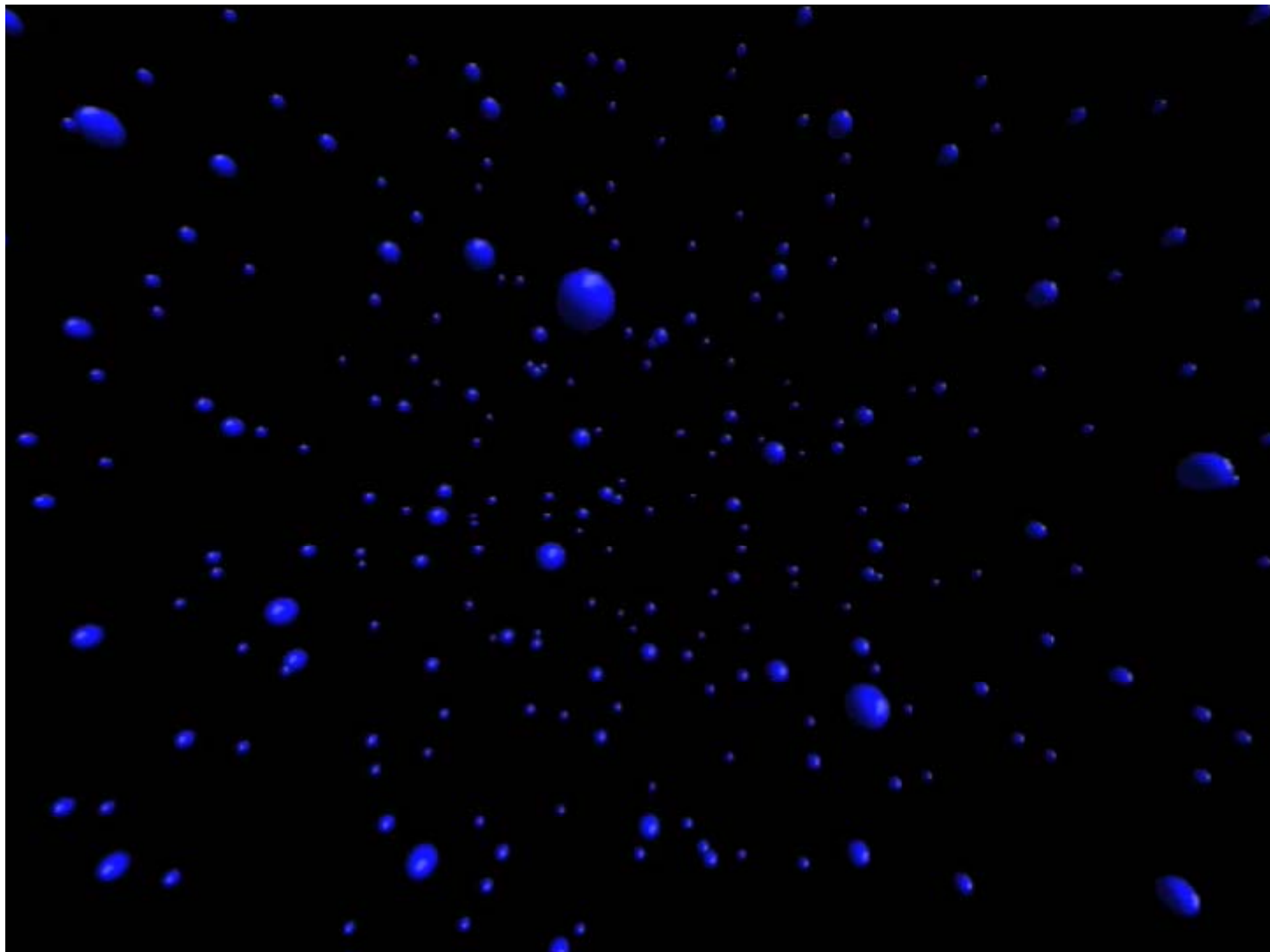


Red, blue, or green: departments

Yellow: consultants

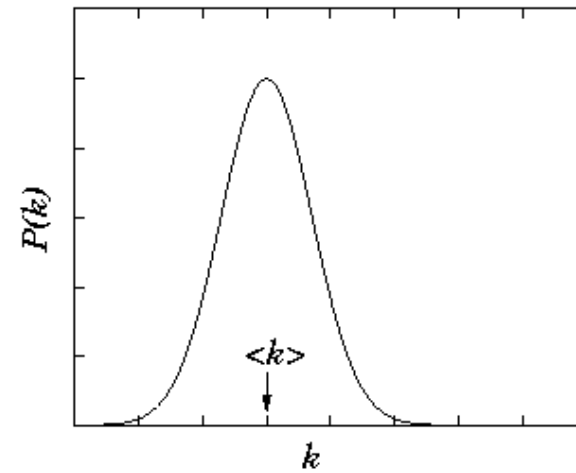
Grey: external experts

www.orgnet.com



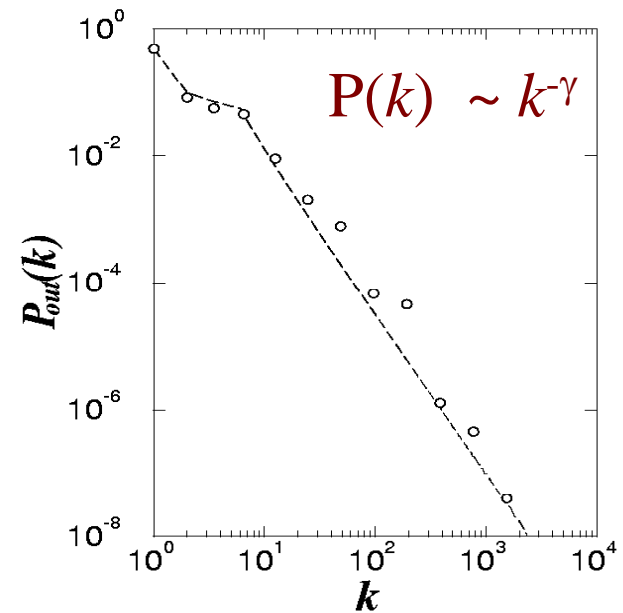
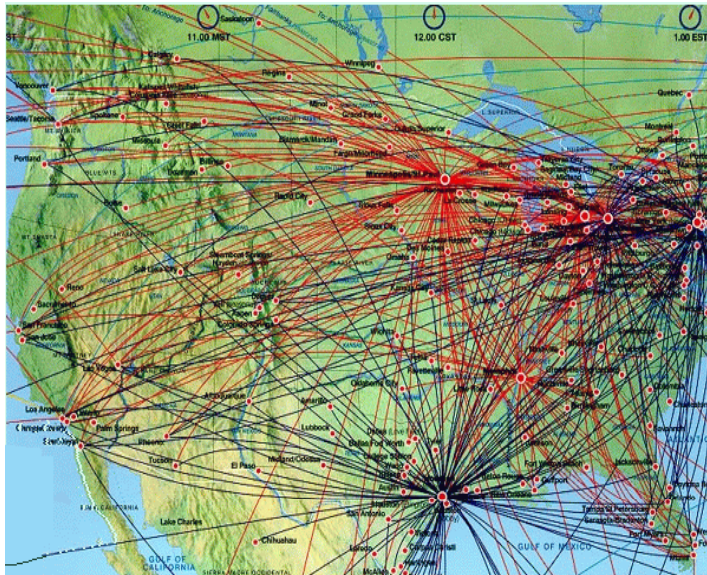
World Wide Web

Exponential Network



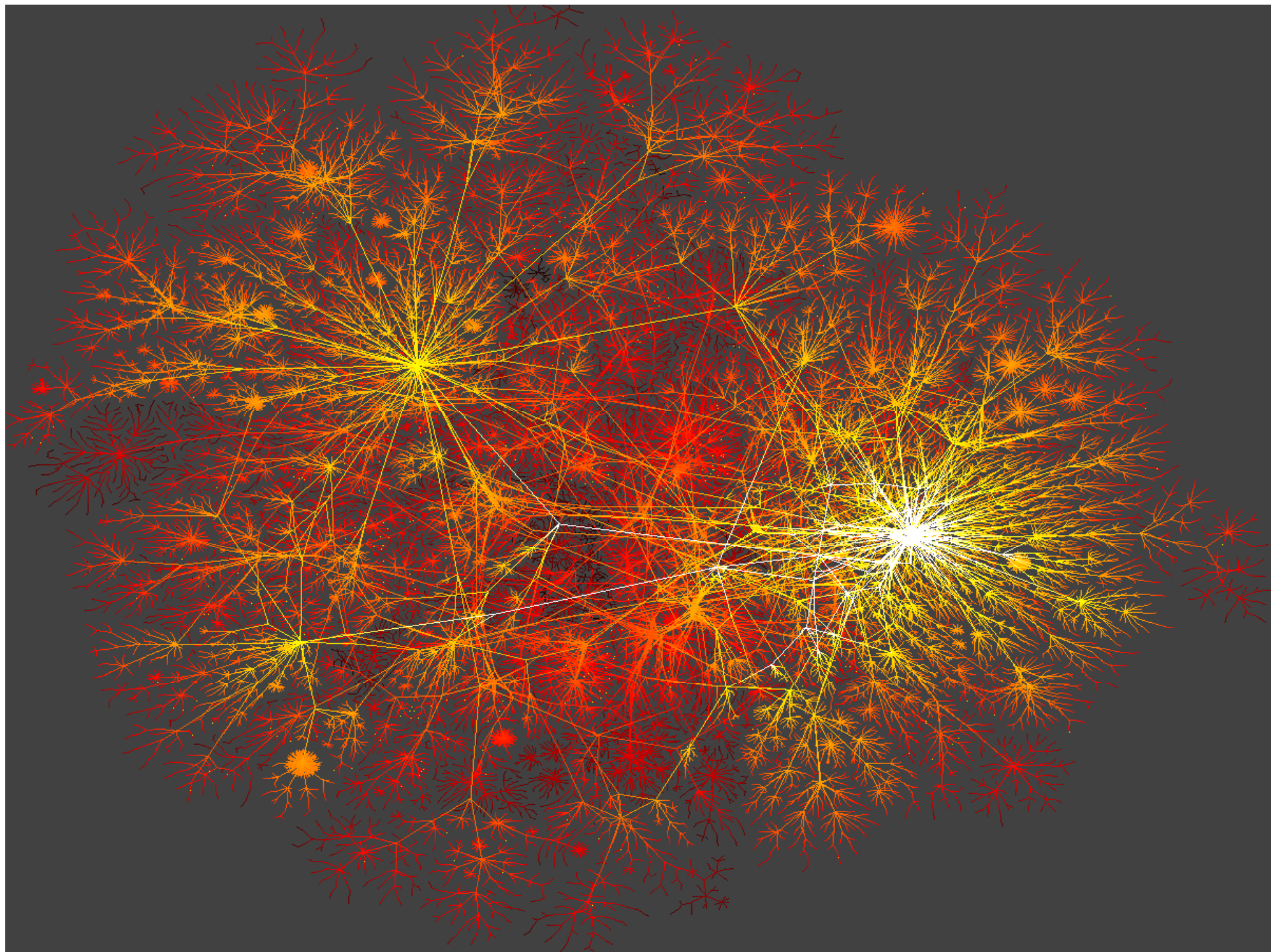
Expected

Scale-free Network



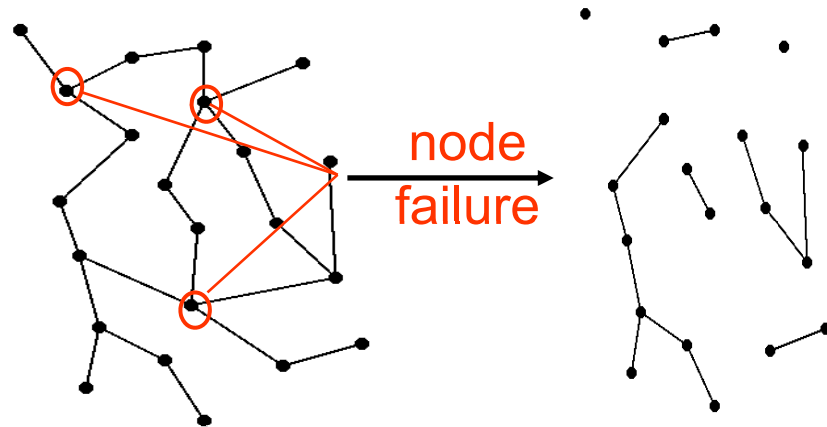
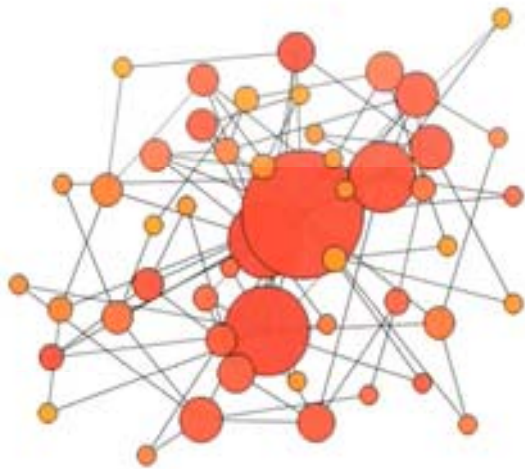
Found

R. Albert, H. Jeong, A-L Barabási, *Nature*, **401** 130 (1999).



Topology matters!

Network Robustness:

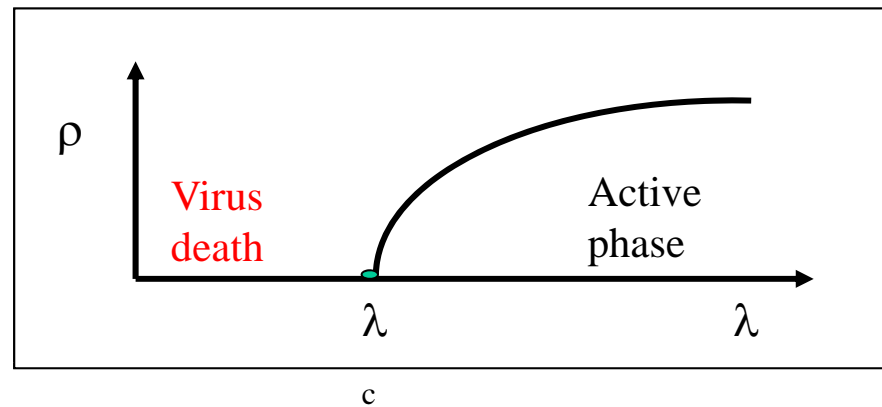


$$f_c = 1 - \frac{1}{\langle k^2 \rangle / \langle k \rangle - 1} \quad \gamma < 3 \quad \langle k^2 \rangle \rightarrow \infty \quad N \rightarrow \infty \quad f_c \rightarrow 1$$

Spreading Phenomena:

$$\lambda_c = \frac{\langle k \rangle}{\langle k^2 \rangle}$$

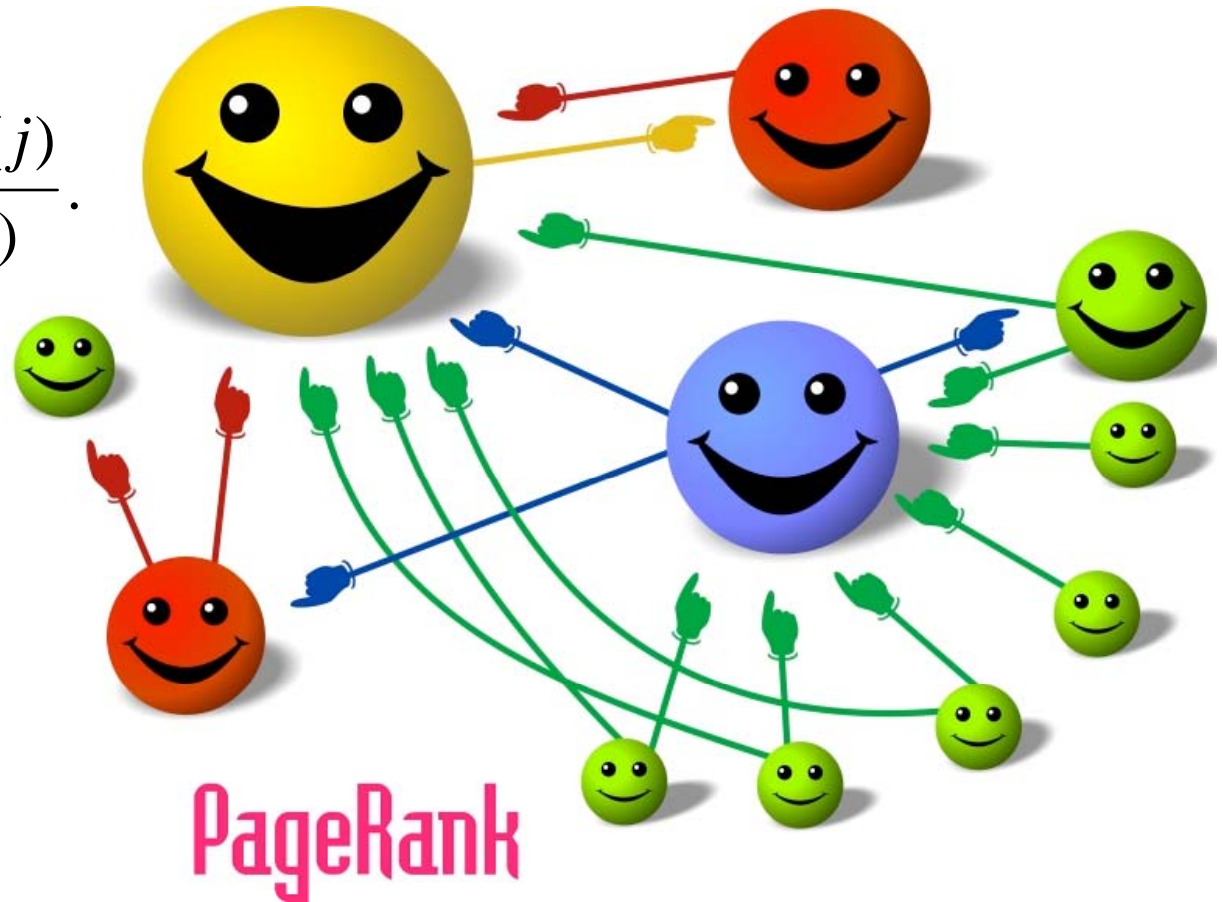
$$\gamma < 3 \quad \langle k^2 \rangle \rightarrow \infty \quad N \rightarrow \infty \quad \lambda_c \rightarrow 0$$



PageRank

$$p_t(i) = \frac{1 - \alpha}{N} + \alpha \sum_j \frac{A_{ij} p_{t-1}(j)}{k_{out}(j)}.$$

$p_t(i)$: pagerank of node i in iteration (time) t



- Surf the network by following edges at random, occasionally (probability α) “jumping” to a randomly chosen node.
- Stationary state ($t \rightarrow \infty$) leads to pagerank $p(i)$ of node i .
- Diffusion on a directed network.

Googling Food Webs: Can an Eigenvector Measure Species' Importance for Coextinctions?

Stefano Allesina^{1*}, Mercedes Pascual^{2,3,4}

1 National Center for Ecological Analysis and Synthesis, Santa Barbara, California, United States of America, **2** Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, Michigan, United States of America, **3** Santa Fe Institute, Santa Fe, New Mexico, United States of America, **4** Howard Hughes Medical Institute

Abstract

A major challenge in ecology is forecasting the effects of species' extinctions, a pressing problem given current human impacts on the planet. Consequences of species losses such as secondary extinctions are difficult to forecast because species are not isolated, but interact instead in a complex network of ecological relationships. Because of their mutual dependence, the loss of a single species can cascade in multiple coextinctions. Here we show that an algorithm adapted from the one Google uses to rank web-pages can order species according to their importance for coextinctions, providing the sequence of losses that results in the fastest collapse of the network. Moreover, we use the algorithm to bridge the gap between qualitative (who eats whom) and quantitative (at what rate) descriptions of food webs. We show that our simple algorithm finds the best possible solution for the problem of assigning importance from the perspective of secondary extinctions in all analyzed networks. Our approach relies on network structure, but applies regardless of the specific dynamical model of species' interactions, because it identifies the subset of coextinctions common to all possible models, those that will happen with certainty given the complete loss of prey of a given predator. Results show that previous measures of importance based on the concept of "hubs" or number of connections, as well as centrality measures, do not identify the most effective extinction sequence. The proposed algorithm provides a basis for further developments in the analysis of extinction risk in ecosystems.

When the Web meets the cell: using personalized PageRank for analyzing protein interaction networks

Gábor Iván^{1,2} and Vince Grolmusz^{1,2,*}

¹Protein Information Technology Group, Eötvös University, Pázmány Péter sétány 1/C and ²Uratim Ltd., InfoPark D, H-1117 Budapest, Hungary

Associate Editor: Trey Ideker

ABSTRACT

Motivation: Enormous and constantly increasing quantity of biological information is represented in metabolic and in protein interaction network databases. Most of these data are freely accessible through large public depositories. The robust analysis of these resources needs novel technologies, being developed today.

Results: Here we demonstrate a technique, originating from the PageRank computation for the World Wide Web, for analyzing large interaction networks. The method is fast, scalable and robust, and its capabilities are demonstrated on metabolic network data of the tuberculosis bacterium and the proteomics analysis of the blood of melanoma patients.

Availability: The Perl script for computing the personalized PageRank in protein networks is available for non-profit research applications (together with sample input files) at the address: <http://uratim.com/pp.zip>.

Contact: grolmusz@cs.elte.hu.

The most successful web page ranking algorithm, the PageRank algorithm, was developed by Brin and Page (1998), and used in the search engine of Google. The algorithm can be described as the following random walk on the graph: the walker starts at a uniformly chosen random vertex of the graph, then with probability $1 - c$ it follows a uniformly selected, random outleading edge from the vertex, and with probability c it teleports to a uniformly selected, random vertex of the graph, where $0 < c < 1$. The PageRank of a node v , corresponding to a certain sense to its importance, is the stationary limit probability distribution, that the walker is at the node v .

In applications for biological networks, the stability of the PageRank is the most attractive property, since the published protein interaction networks contain numerous false positive and false negative interaction edges, even for the highest quality of data gathered for one of the most researched subjects, the yeast interactome (Gavin *et al.*, 2006; Goll and Uetz, 2006; Kroger *et al.*, 2006). Therefore, network-ranking algorithms need to be stable in

Phys. Rev. E 80, 056103 (2009) [10 pages]

Diffusion of scientific credits and the ranking of scientists

Abstract

References

Citing Articles (7)

1973

2004

Rank	Author	NP	WP	BM	DM	PM	Rank	Author	NP	WP	BM	DM	PM
1	GELL-MANN, M	1969	-	-	-	-	1	ANDERSON, PW	1977	-	-	-	-
2	WEINBERG, S	1979	-	-	-	-	2	WITTEN, E	-	-	-	1985	-
3	SCHWINGER, J	1965	-	-	-	-	3	TOKURA, Y	-	-	-	-	-
4	FEYNMAN, RP	1965	-	-	-	-	4	PERDEW, JP	-	-	-	-	-
5	LEE, TD	1957	-	-	-	-	5	KOHN, W	-	-	-	-	-
6	ANDERSON, PW	1977	-	-	-	-	6	KRESSE, G	-	-	-	-	-
7	BJORKEN, JD	-	-	-	2004	-	7	BÜTTIKER, M	-	-	-	-	-
8	YANG, CN	1957	-	-	-	-	8	WEINBERG, S	1979	-	-	-	-
9	SLATER, JC	-	-	-	-	-	9	CIRAC, JI	-	-	-	-	-
10	ADLER, SL	-	-	-	1998	-	10	ZUNGER, A	-	-	-	-	-
11	GLAUBER, RJ	2005	-	-	-	-	11	BARABÁSI, AL	-	-	-	-	-
12	CHEW, GF	-	-	-	-	-	12	LEE, PA	-	-	-	2005	-
13	WIGNER, EP	1963	-	-	-	1961	13	VANDERBILT, D	-	-	-	-	-
14	LOVELACE, C	-	-	-	-	-	14	SACHDEV, S	-	-	-	-	-
15	SATCHLER, GR	-	-	-	-	-	15	NEWMAN, MEJ	-	-	-	-	-
16	MOTT, NF	1977	-	-	1985	-	16	AFFLECK, I	-	-	-	-	-
17	FISHER, ME	-	1980	1983	-	-	17	MACDONALD, AH	-	-	-	-	-
18	MANDELSTAM, S	-	-	-	1991	-	18	HIRSCH, JE	-	-	-	-	-
19	BETHE, HA	1967	-	-	-	1955	19	ZOLLER, P	-	-	-	2006	2005
20	PHILLIPS, JC	-	-	-	-	-	20	PARISI, G	-	-	1992	1999	-

PageRank has transcended Google

Citations

D. Walker et al, *J.Stat. Mech.* (2007)
F. Radicchi et al, *Phys. Rev. E.* (2009).

Ecological species

Allesina and Pascual, *Plos Comput. Biol.* (2009).

Genes

J. Chen et al, *BMC Bioinformatics* (2009)
Ivan and Grolmusz, *Bioinformatics* (2011).

PhD Programs

B.M. Schmidt and M.M. Chingos, *PS: Political Science and Politics* (2007).

Traffic Flow

B. Jiang et al, *J. Stat. Mech* (2008).

Word Disambiguation

Navigli and Lapata, *IEEE-TPAMI* (2010).

Major differences between the topological properties of these systems!

→ How does the ranking quality depend on the network topology?

→ Is pagerank intrinsically better at ranking some networks than other networks?

Ranking quality is a major issue in computer science

Pagerank is score-stable (L_1) to edge additions and deletions

S. Chien et al, Link Evolution: Analysis and Algorithms, *Internet Math.* **1**, 3 277(2003).

Stability governed by reset parameter α

A.Y. Ng et al, Link Analysis, Eigenvectors and Stability, *ICJAI-01*, 903 (2001).

Pagerank is not rank-stable to edge perturbations

R. Lempel and S. Moran, Rank-stability and rank-similarity of link-based web ranking algorithms in authority-connected graphs, *Inf. Retr.* **8** (2005).

Local topology of network can be manipulated to maximize effective search and visitation times using pagerank

C. de Kerchove et al, Maximizing PageRank via outlinks, *Linear Algebra and its Applications*, **429** 1254 (2008).

Spectral properties of Google matrix

O. Giraud et al *Phys. Rev E* **80** (2009)

Mean Field Approximation

Average pagerank of a node with degree $\mathbf{k} = (k_{in}, k_{out})$:

$$\bar{p}(\mathbf{k}) = \frac{(1 - \alpha)}{N} + \frac{\alpha}{N} \times \frac{k_{in}}{\langle k_{in} \rangle}.$$

Fluctuation of pagerank around mean value:

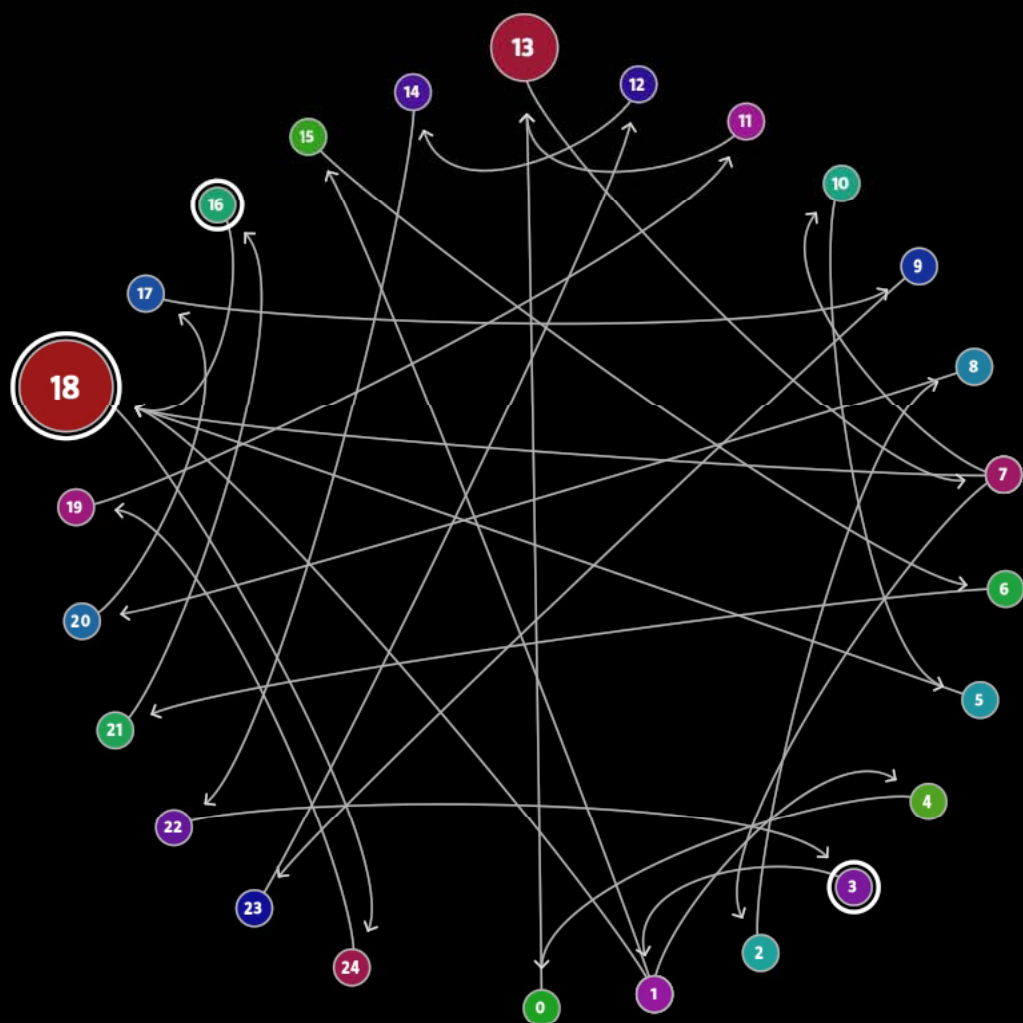
$$\sigma(\mathbf{k}) \approx \frac{\alpha^2}{N} \left\langle \frac{k_{in}^2}{k_{out}} \right\rangle^{1/2} \left\langle k_{in}^{-3/2} \right\rangle \times k_{in}^{1/2}.$$

If a node's in-degree changes, it will obviously alter its PageRank.

Degree preserving perturbations maintain the role of the network topology.

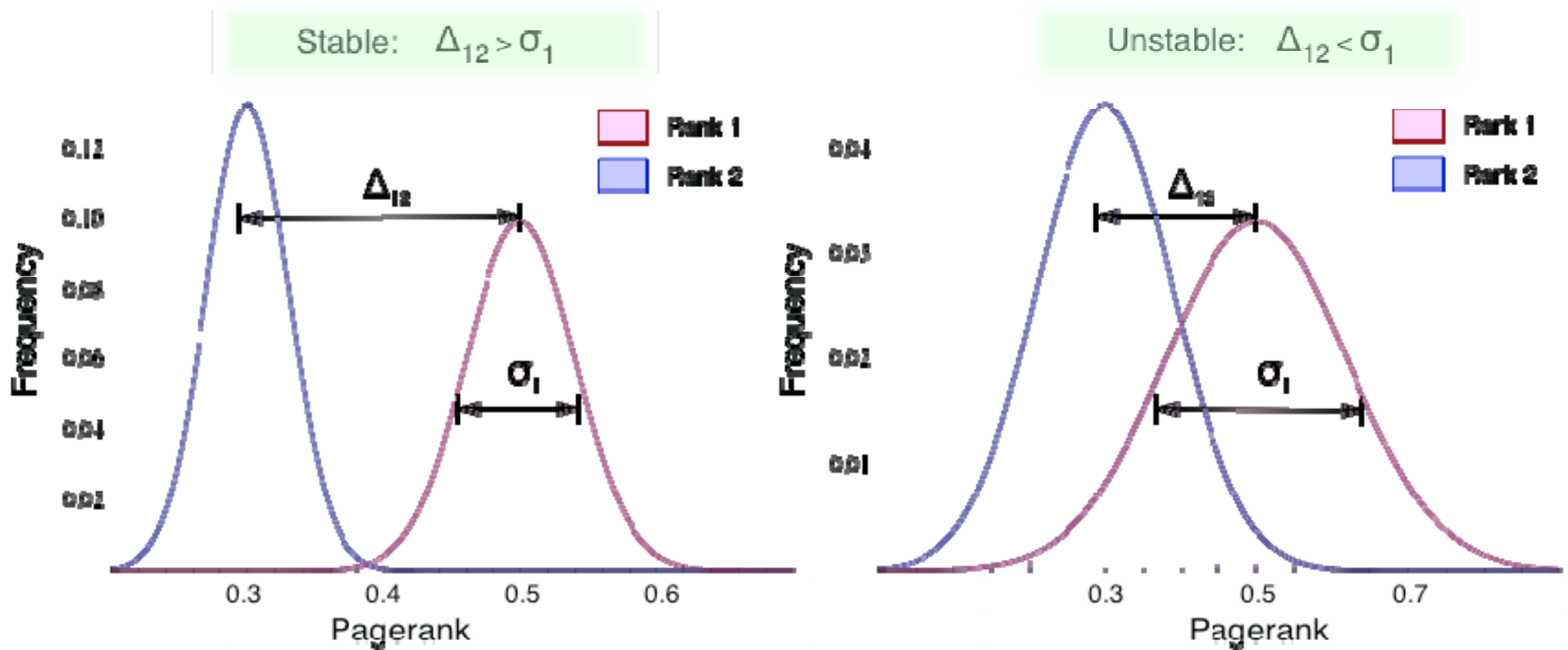
¹S. Fortunato *et al*, Lecture Notes in Computer Science **4936**,59 (2008).

Network: 0



- 1 **18**
- 2 **13**
- 3 **24**
- 4 **7**
- 5 **19**
- 6 **11**
- 7 **1**
- 8 **3**
- 9 **22**
- 10 **14**
- 11 **23**
- 12 **9**
- 13 **17**
- 14 **20**
- 15 **8**
- 16 **5**
- 17 **2**
- 18 **10**
- 19 **21**
- 20 **6**
- 21 **0**
- 22 **15**
- 23 **4**
- 24 **16**
- 25 **12**

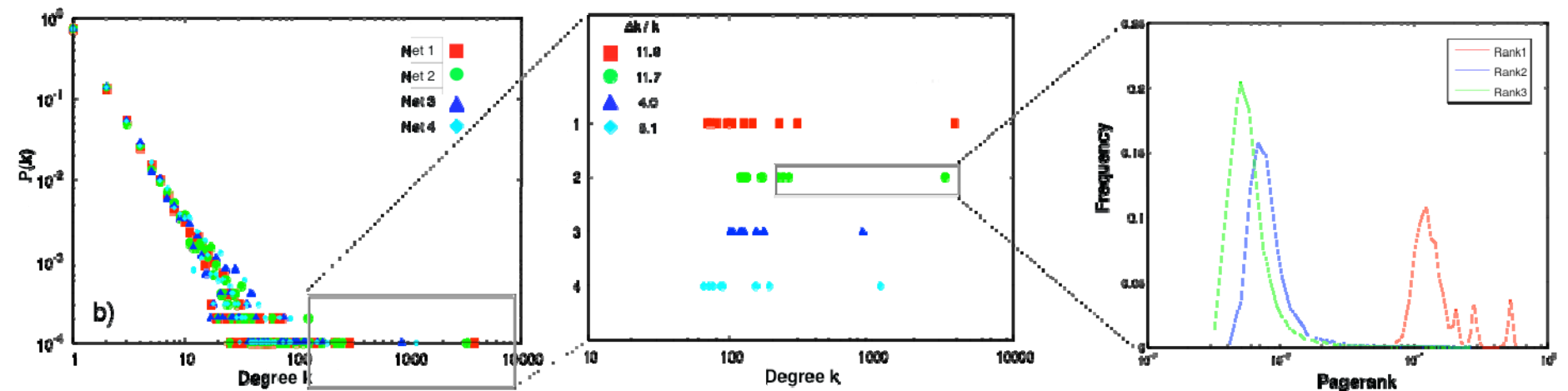
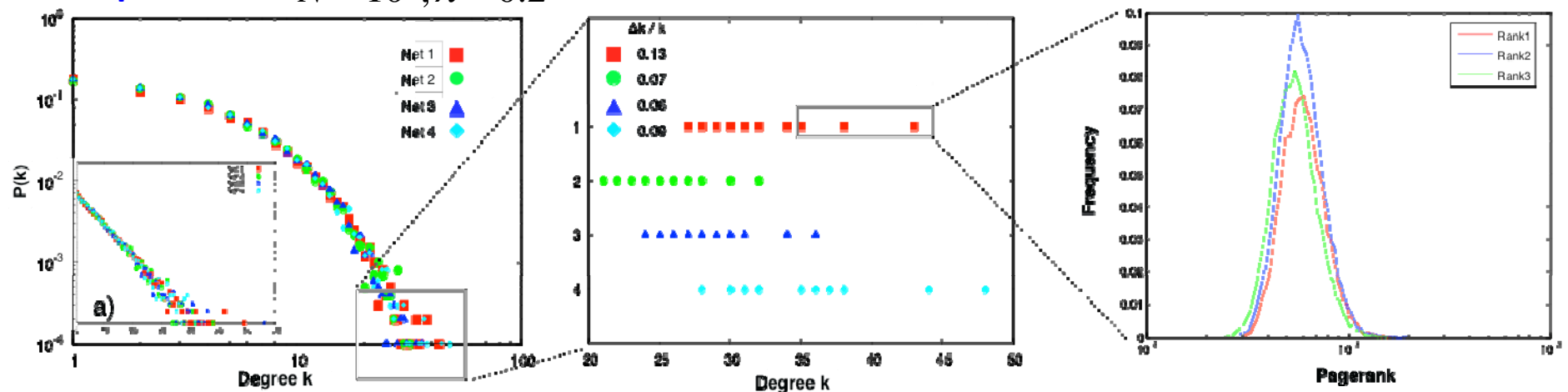
Stability Criteria



Stability Criteria: $\sigma(p_m) \leq \Delta(p_m).$ $\frac{\Delta(p_m)}{\sigma(p_m)} \geq 1.$

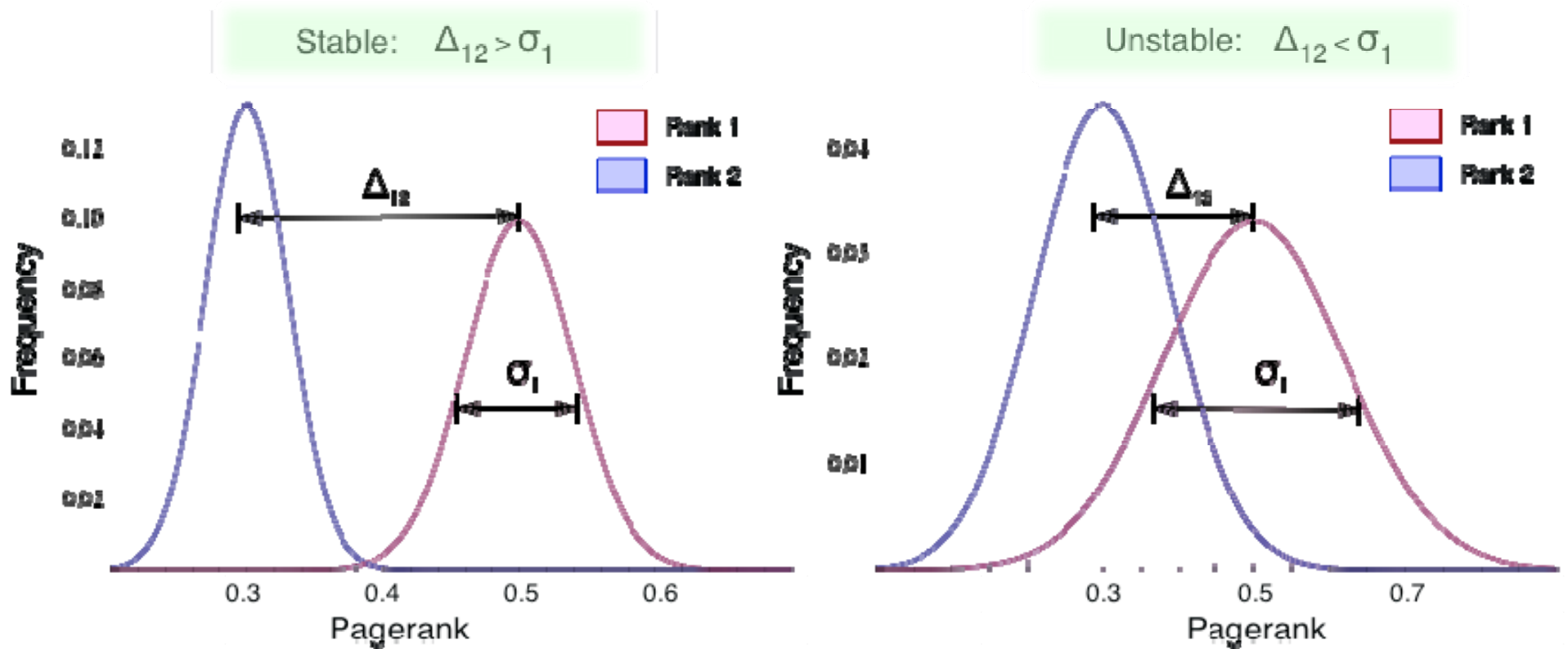
Difference Between Scale-Free and Exponential Distributions

Exponential: $N = 10^4, \lambda = 0.2$



Scale-free: $N = 10^4, \gamma = 2.3$

Stability Criteria



Stability Criteria:

$$\sigma(p_m) \leq \Delta(p_m).$$

$$\frac{\Delta(p_m)}{\sigma(p_m)} \geq 1.$$

The gap $\Delta(p_m) = (p_m - p_{m+1})$ and fluctuation $\sigma(p_m)$ for different topologies.

Exponential $[P(k) \sim e^{-\lambda k}]$:

$$\Delta^{\text{exp}}(p_m) = \frac{\alpha}{N} \times f_{\Delta}^{\text{exp}}(m)$$

$$\sigma^{\text{exp}}(p_m) = \frac{\alpha^2}{N} \times g_{\sigma}^{\text{exp}}(m, \lambda)$$

Scale-Free $[P(k) \sim k^{-\gamma}]$:

$$\Delta^{SF}(p_m) = \frac{\alpha}{N^{(\gamma+2)/(\gamma-1)}} \times f_{\Delta}^{SF}(m, \gamma)$$

$$\sigma^{SF}(p_m) = \frac{\alpha^2}{N^{(2\gamma-3)/2(\gamma-1)}} \times g_{\sigma}^{SF}(m, \gamma)$$

**Stability
Ratio:**

$$\frac{\Delta^{\text{exp}}(p_m)}{\sigma^{\text{exp}}(p_m)} = \alpha^{-1} \times F^{\text{exp}}(m, \lambda).$$

$$\frac{\Delta^{SF}(p_m)}{\sigma^{SF}(p_m)} = \frac{N^{1/2(\gamma-1)}}{\alpha} \times F^{SF}(m, \gamma).$$

**Stability
Criteria:**

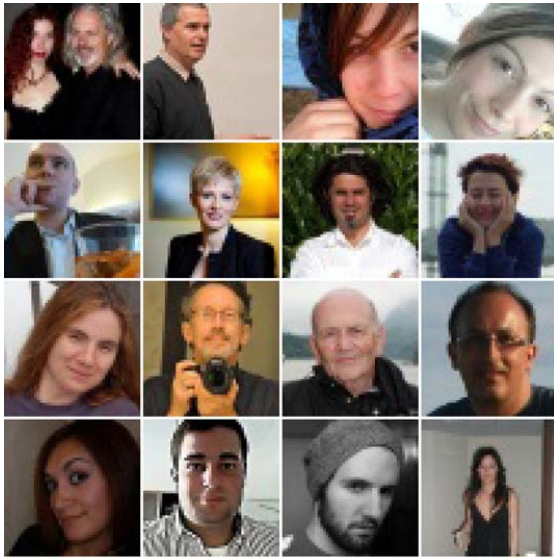
$$\frac{\Delta(p_m)}{\sigma(p_m)} \geq 1.$$

Does not depend on N!

The larger N, the more likely that the stability criteria is satisfied!

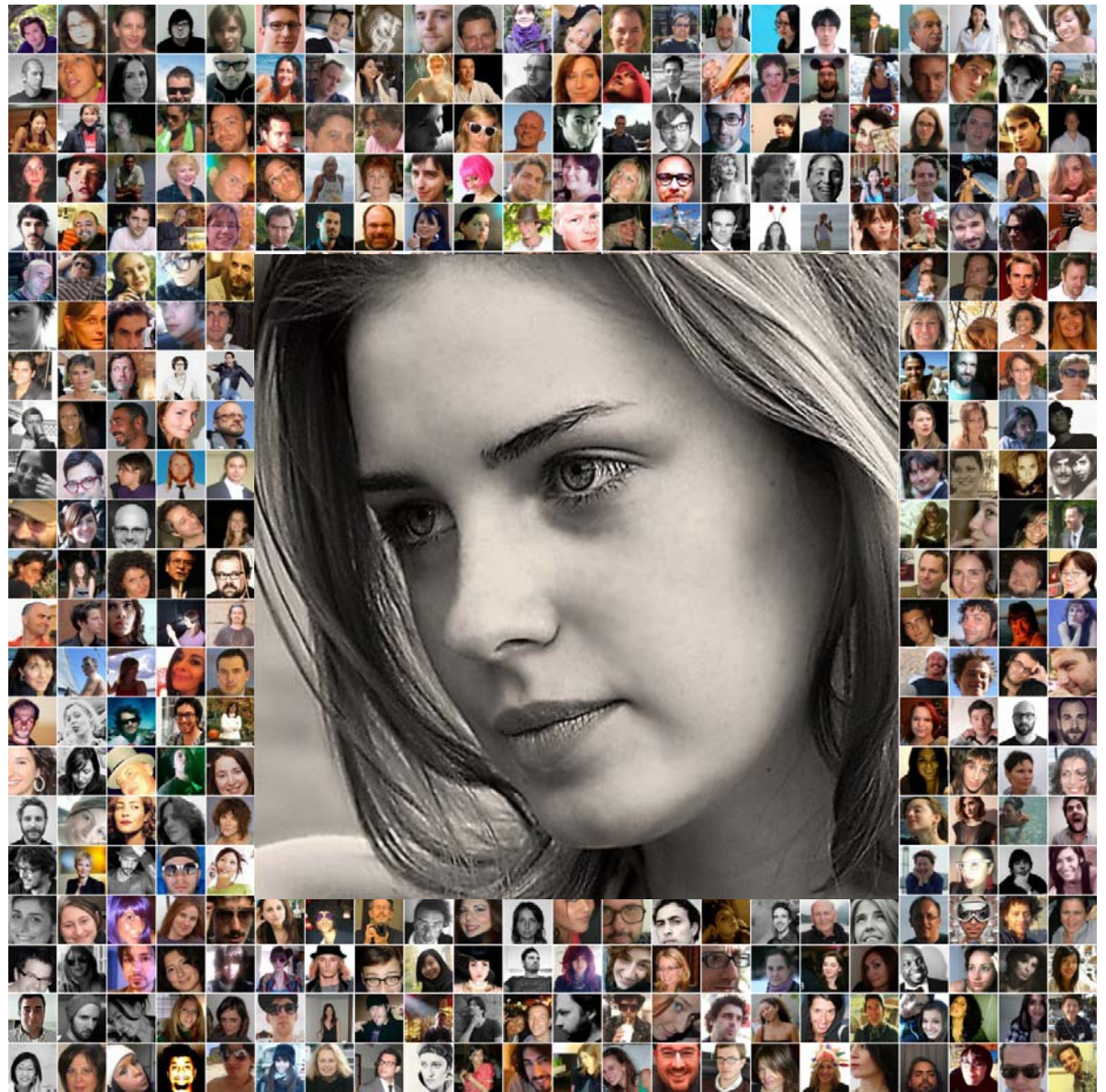
m : the rank of a node

SIZE MATTERS!

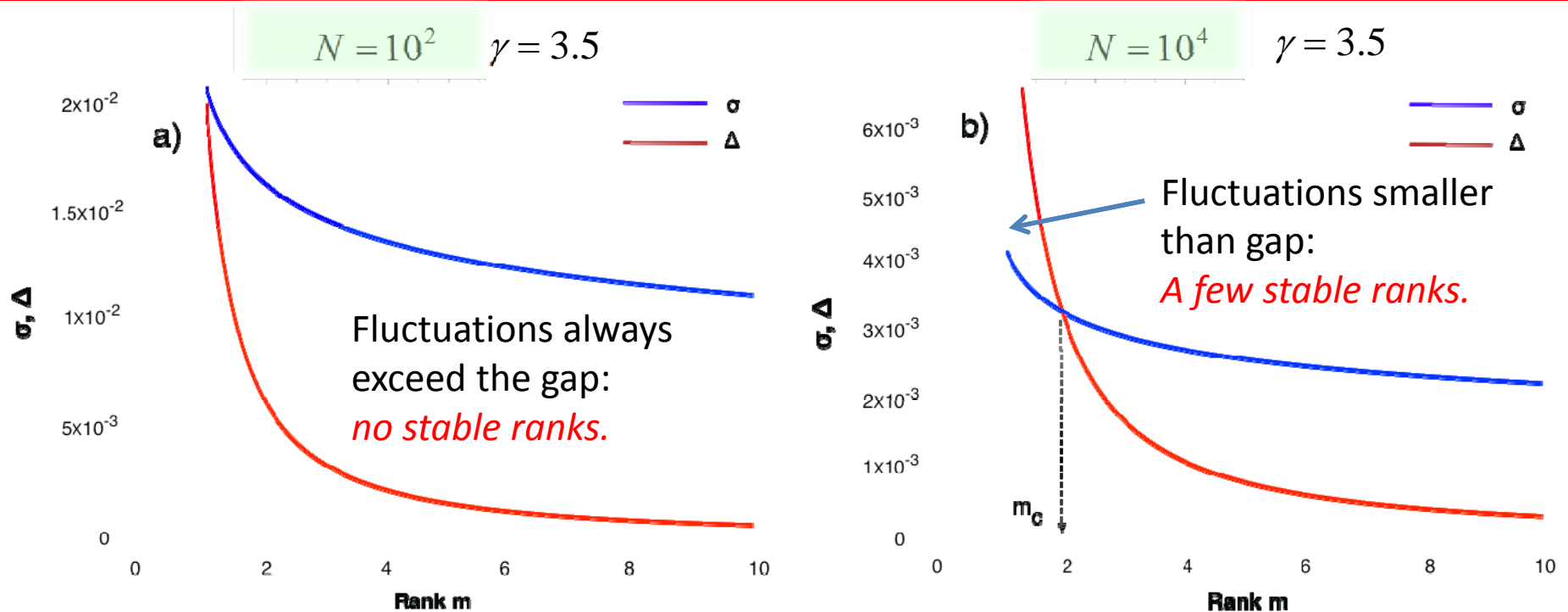


In a larger system, there is a higher chance that we have an *outlier*: significantly prettier, taller, smarter, or richer than everyone else.

Scale-free systems *have outliers*, bounded systems *do not*.



Size matters when we rank in a scale-free environment



→ critical system size N_c below which the stability criteria is never satisfied.

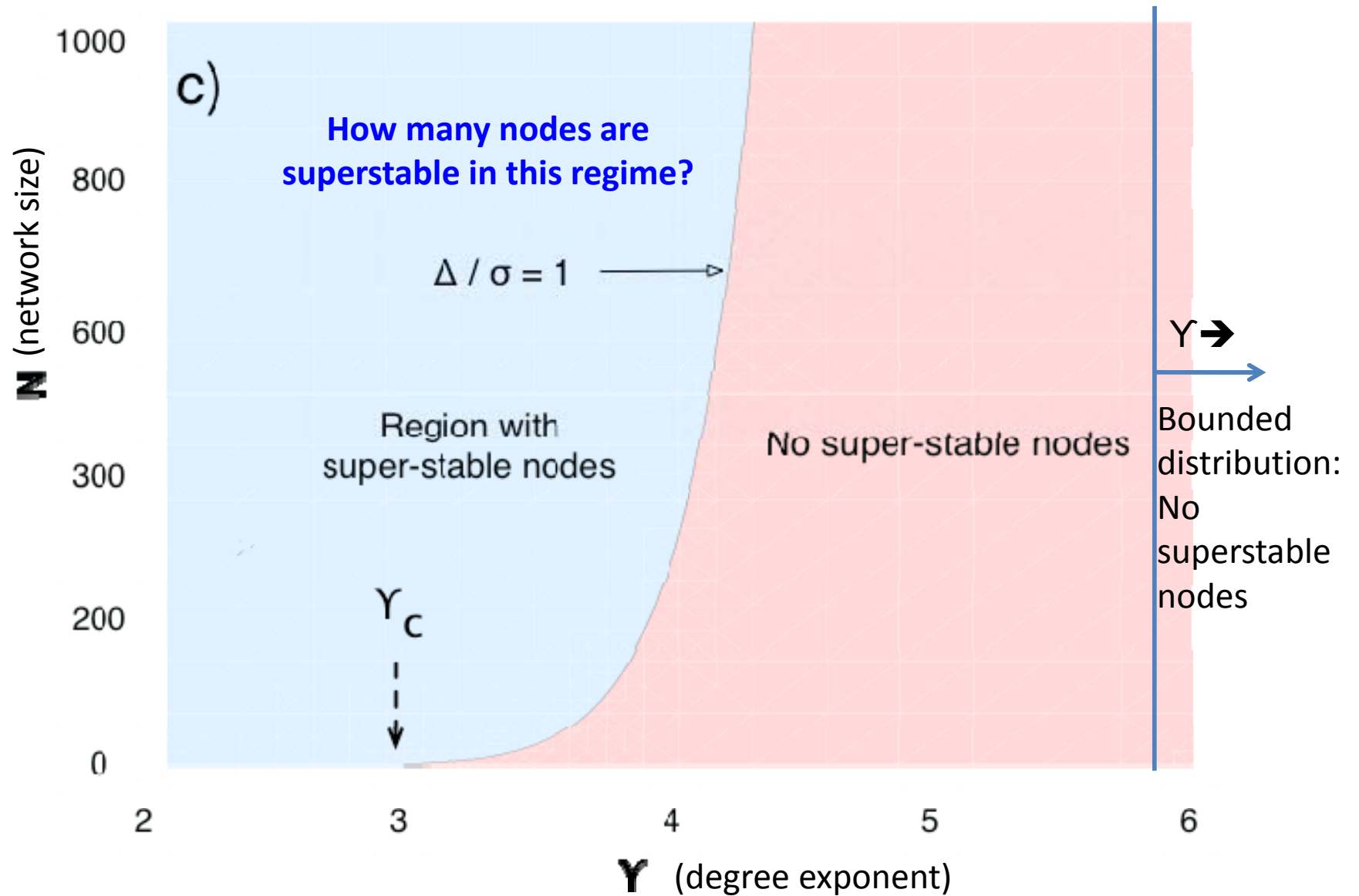
To have at least one stable node, the gap must exceed the fluctuation for the top-ranked node $m = 1$.

$$\frac{\Delta(p_1)}{\sigma(p_1)} = 1,$$

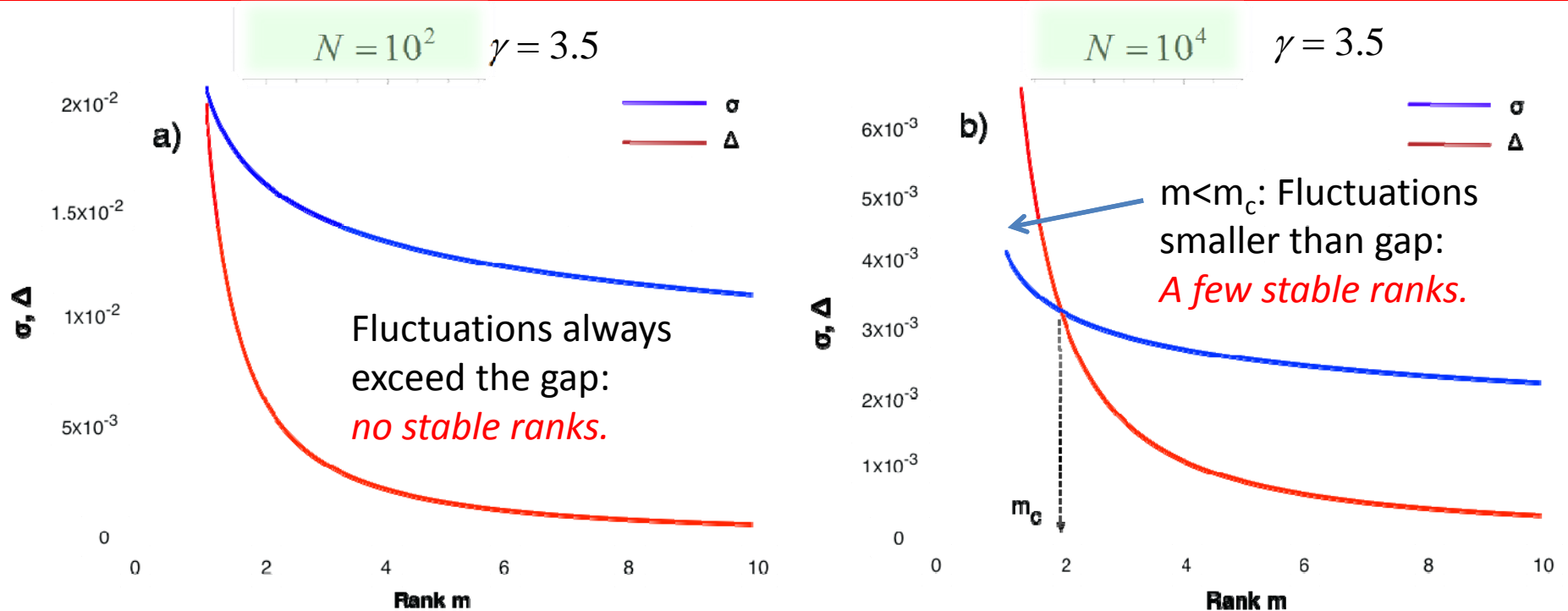
$$\pm \frac{N^{1/2(\gamma-1)}}{[1 - N^{(3-\gamma)/(\gamma-1)} \Gamma(1 - \frac{1}{\gamma-1})^{3-\gamma}]} = \frac{\alpha^2(\gamma-1)^2}{\Gamma(1 - \frac{1}{\gamma-1})} \times \left[\frac{(\gamma-1)(\gamma-2)}{\pm \gamma(\gamma-3)} \right].$$

N_c is a solution to this transcendental equation.

Critical system size N_c



Size matters in scale-free networks



→ critical system size N_c below which stability criteria is *never* satisfied.

→ critical rank m_c below which stability criteria *is* satisfied.

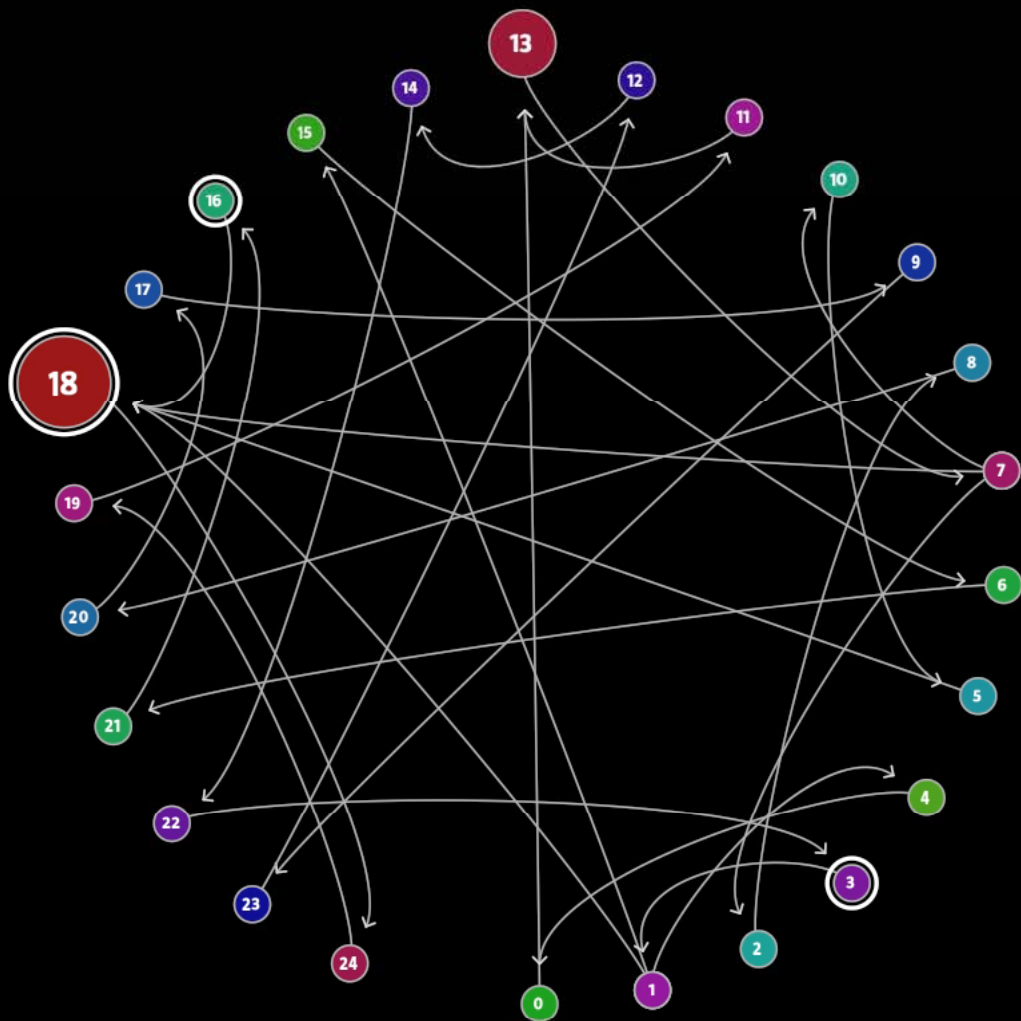
Determine m_c from the stability ratio:

$$\frac{\Delta(p_m)}{\sigma(p_m)} = 1 \quad \rightarrow \quad m_c \approx N^{1/(2\gamma-1)} \times F(\alpha, \gamma).$$

$m < m_c$: superstable nodes.

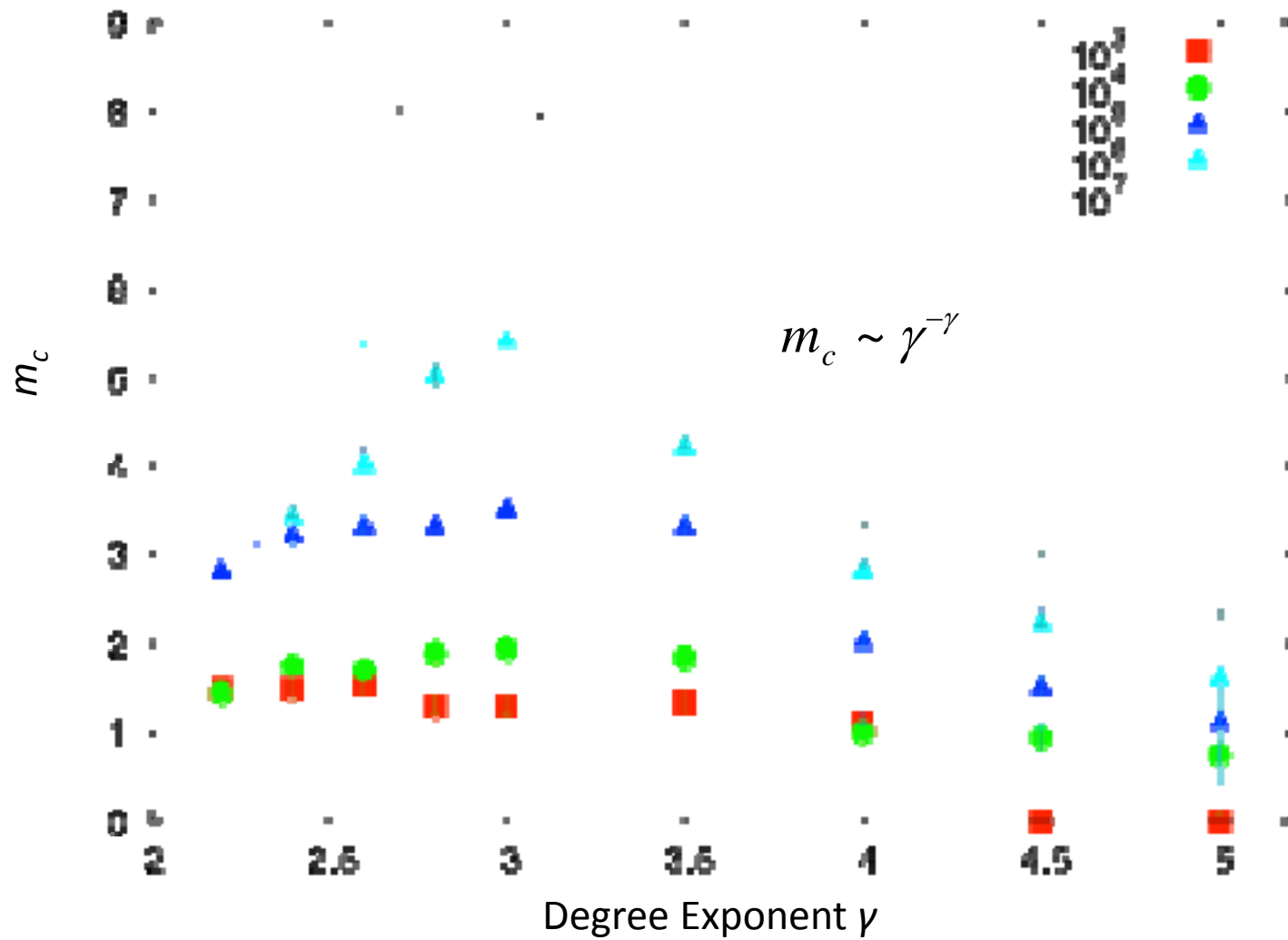
By the virtue of the many links they have, their ranking is independent of who points at them!!!

Network: 0



- 1 18
- 2 13
- 3 24
- 4 7
- 5 19
- 6 11
- 7 1
- 8 3
- 9 22
- 10 14
- 11 23
- 12 9
- 13 17
- 14 20
- 15 8
- 16 5
- 17 2
- 18 10
- 19 21
- 20 6
- 21 0
- 22 15
- 23 4
- 24
- 25

Critical rank m_c

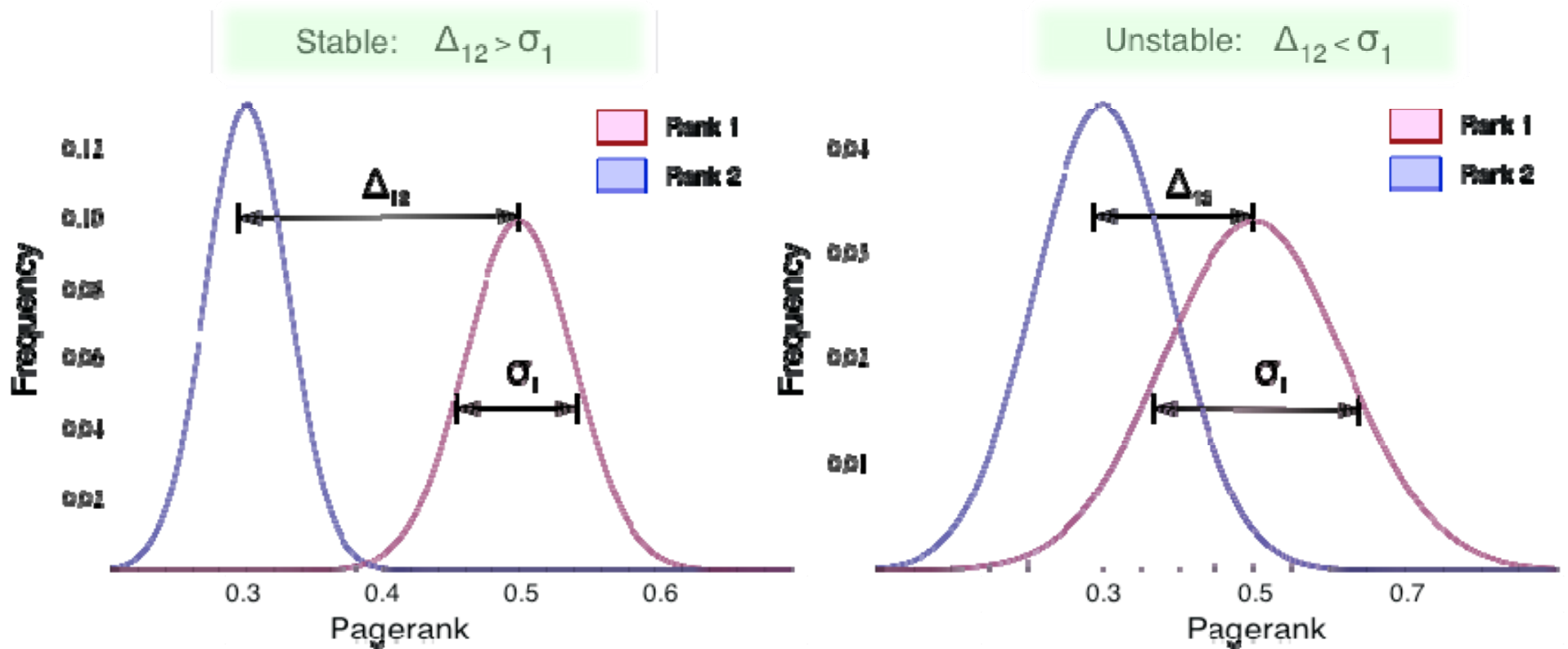


**Do we have superstable nodes
in real systems?**

Super-stable nodes in real networks

	Network	N	γ	N_c	m_c	m_c
					predicted	measured
Scale-Free	Google web sample	875,713	2.5	0	4	1
	Notre Dame web sample	325,729	2.1	0	3	3
	Stanford web sample	281,903	2.1	0	2	2
	Berkeley-Stanford web sample	685,230	2.1	0	3	3
	Hep-th citations	27,770	2.8	0	3	2
	Amazon co-purchase	223,431	4.3	2,730	2	2
	Wikipedia admin. voting	7115	3.3	50	2	1
	Mobile Call Graph	4,562,263	5.2	150,000	2	3
Exponential	<i>C.Elegans</i> neural net.	307	-	∞	0	1
	Food Web (Little Rock Lake)	183	-	∞	0	0
	Food Web (Silwood Park)	154	-	∞	0	0

Stability Criteria



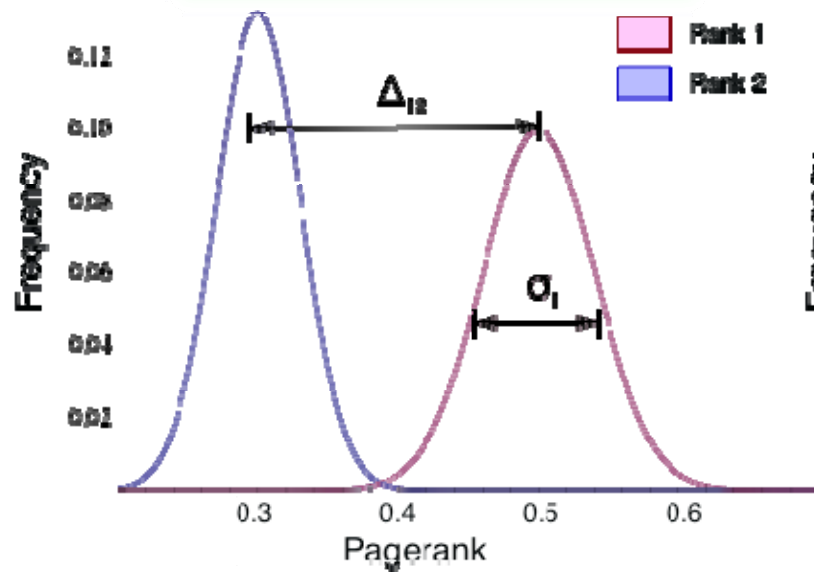
Stability Criteria:

$$\sigma(p_m) \leq \Delta(p_m).$$

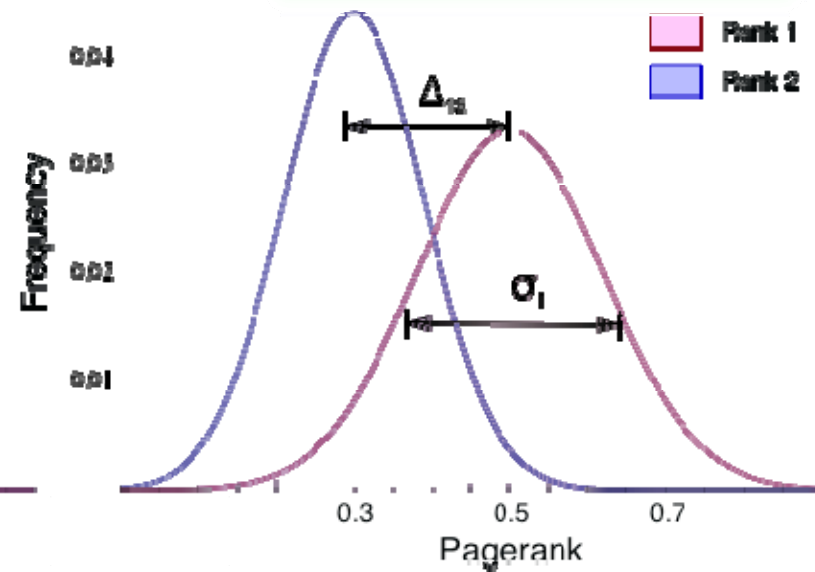
$$\frac{\Delta(p_m)}{\sigma(p_m)} \geq 1.$$

Real Data

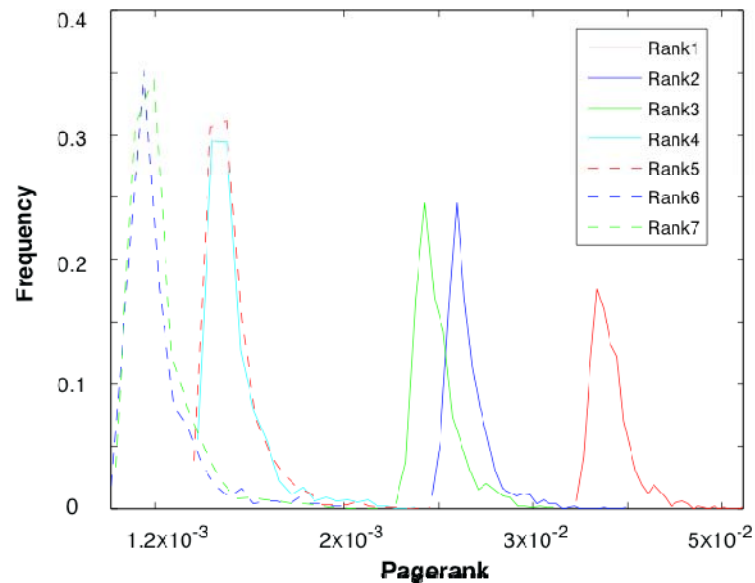
Stable: $\Delta_{12} > \sigma_1$



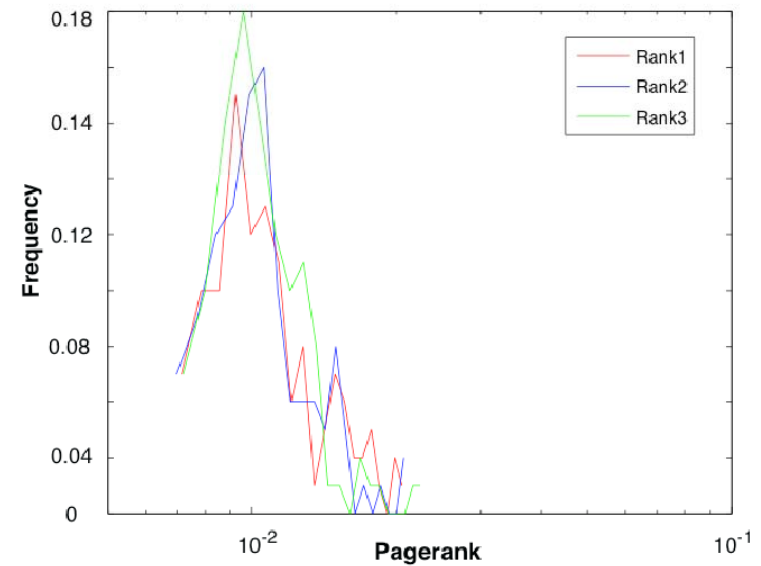
Unstable: $\Delta_{12} < \sigma_1$



Notre Dame website



Food Web Silwood Park

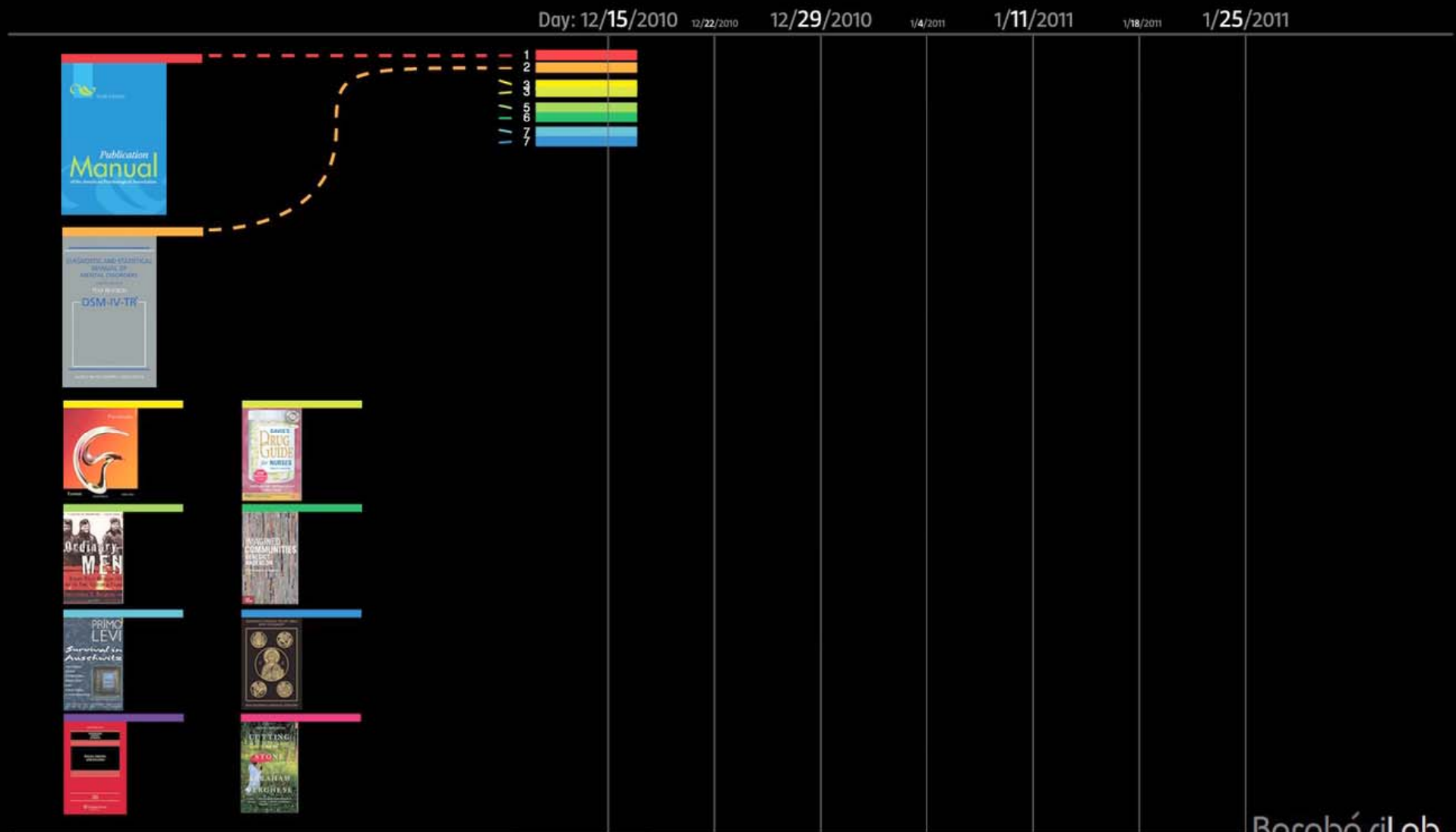


An Unexpected Bonus: Temporal Stability

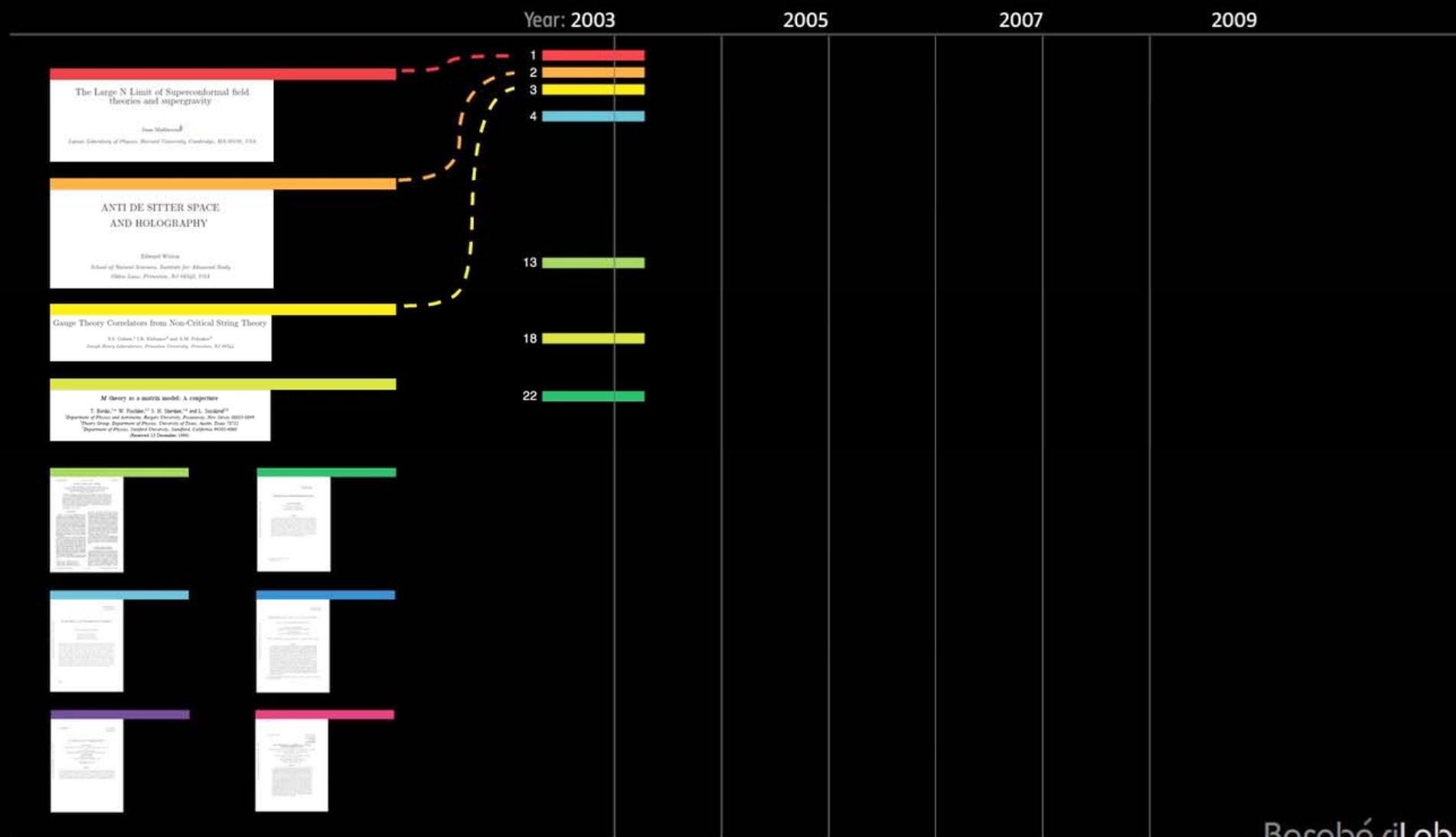
Super-stable nodes in real networks

	Network	N	γ	N_c	m_e	m_e
					predicted	measured
Scale-Free	Google web sample	875,713	2.5	0	4	1
	Notre Dame web sample	325,729	2.1	0	3	3
	Stanford web sample	281,903	2.1	0	2	2
	Berkeley-Stanford web sample	685,230	2.1	0	3	3
	Hep-th citations	27,770	2.8	0	3	2
	Amazon co-purchase	223,431	4.3	2,730	2	2
	Wikipedia admin. voting	7115	3.3	50	2	1
	Mobile Call Graph	4,562,263	5.2	150,000	2	3
Exponential	<i>C.Elegans</i> neural net.	307	-	∞	0	1
	Food Web (Little Rock Lake)	183	-	∞	0	0
	Food Web (Silwood Park)	154	-	∞	0	0

Temporal stability of super-stable nodes



Temporal stability of super-stable nodes



- *Topology does matter*: scale-free networks (e.g. www) contain a few superstable nodes, whose pagerank is remarkably resistant to rewiring perturbations.
- *Size matters too*: the larger the network, the more stable are the top nodes.
- *Super-stable nodes are real*: present in real networks, and their number agrees with the analytical predictions.
- *Super-stability matters*: these nodes demonstrate remarkable temporal stability (not predicted by our analysis, but nice :)



Maybe not, unless you are an outlier...

Thanks to two outliers:
Gourab Ghoshal,
Mauro Martino
CCNR-NEU.

[Journal home](#) > [Archive](#) > [Commentary](#) > [Full Text](#)

Journal content

- [Journal home](#)
- [Advance online publication](#)
- [Current issue](#)
- [Nature News](#)
- [Archive](#)
- [Supplements](#)
- [Web focuses](#)
- [Podcasts](#)
- [Videos](#)
- [News Specials](#)

Journal information

- [About the journal](#)

Commentary

Nature **400**, 107 (8 July 2008)

Accessibility

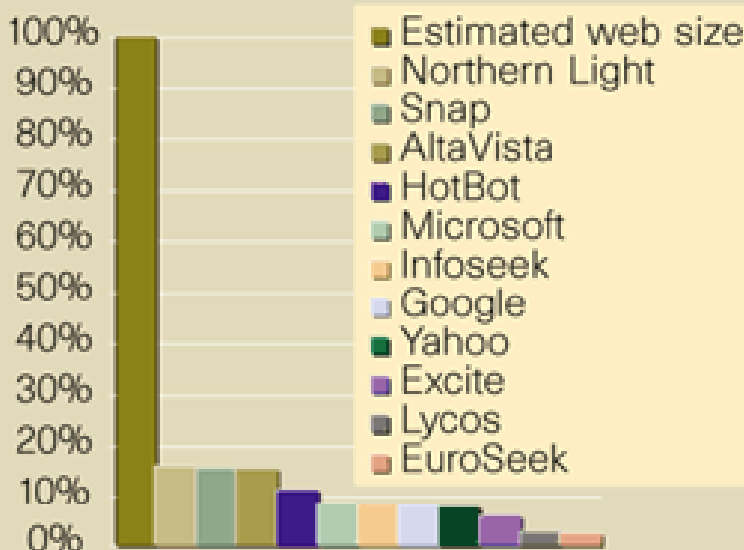
Steve Lawrence¹ and

1. Steve Lawrence
New Jersey 085
e-mail: Email: l

**Search engines c
months, and no c
web becomes a r
made more acce**

The publicly indexable World-Wide Web now contains about 800 million pages, encompassing about 6 terabytes of text data on about 3 million servers. The web is increasingly being used in all aspects of society; for example, consumers use search engines to locate and buy goods, or to research many decisions (such as choosing a holiday destination, medical treatment or election vote). Scientists are increasingly using search engines to locate research of interest: some rarely use libraries, locating research articles primarily online; scientific editors use search

b



subscribe to
nature

FULL TEXT

→ [Previous](#) | [Next](#) →

→ [Table of contents](#)

[Download PDF](#)

[Send to a friend](#)

[CrossRef lists 182 articles citing this article](#)

[Scopus lists 586 articles citing this article](#)

[Export citation](#)

[Export references](#)



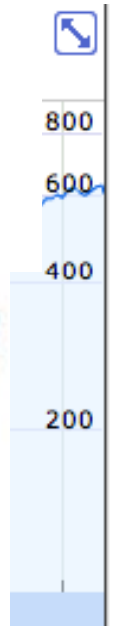
G.Ghoshal and A.-L. Barabasi

Google's dirty little secret revealed...



G.Ghoshal and A.-L. Barabasi
CCNR-NEU.

Network	N	γ	N_c	m_c (predicted)	m_c (measured)
Google Web sample	875,713	2.5	0	4	1
Notre Dame Web sample	325,729	2.1	0	3	3
Stanford Web Sample	281,903	2.1	0	2	2
Berkeley-Stanford Web	685,230	2.1	0	3	3
Hep-th citations	27,770	2.8	0	3	2
Amazon co-purchase	223,431	4.3	2,730	2	2
Wikipedia voting	7115	3.3	50	2	1
Mobile Call Graph	4,562,263	5.2	150,000	2	3
C. Elegans neural	307	-	∞	0	1
Food Web (Little Rock Lake)	183	-	∞	0	0
Food Web (Silwood Park)	154	-	∞	0	0



$\$1.4 \times 10^{10}$

Each!



pagerank



Search

Instant is on ▼

SafeSearch off ▼

About 128,000,000 results (0.15 seconds)

[Advanced search](#)

Everything

Images

Videos

News

Shopping

Blogs

More

Boston, MA

[Change location](#)

Any time

[Latest](#)

[Past 24 hours](#)

[Past week](#)

[Past month](#)

[Past year](#)

[Custom range...](#)

All results

[Wonder wheel](#)

[Related searches](#)

[More search tools](#)

[Something different](#)

► [PageRank - Wikipedia, the free encyclopedia](#) ☆ 🔍 - 6 visits - 9:34am

PageRank is a link analysis algorithm, named after Larry Page and used by the Google Internet search engine, that assigns a numerical weighting to each ...

[Google bomb](#) - [Toolbar](#) - [Google matrix](#) - [EigenTrust](#)
[en.wikipedia.org/wiki/PageRank](#) - [Cached](#) - [Similar](#)

[Google PageRank Checker - Check Google page rank instantly](#) ☆ 🔍

Page Rank Checker is a completely free service to check Google **pagerank** instantly using our online **page rank** check tool or a small **pagerank** button.

[www.prchecker.info/check_page_rank.php](#) - [Cached](#) - [Similar](#)

[Google PageRank Checker - Check Google page rank of any web pages](#) ☆ 🔍

Page Rank Checker is a completely Free tool to check Google PR, **page rank** of ...

[www.prchecker.info/](#) - [Cached](#) - [Similar](#)

[Show more results from prchecker.info](#)

[Pagerank Explained. Google's PageRank and how to make the most of it.](#) ☆ 🔍

Pagerank explained, and what you can do with it. **PageRank** calculator.

[www.webworkshop.net/pagerank.html](#) - [Similar](#)

[PageRank - Google](#) ☆ 🔍

When Google was founded, one key innovation was **PageRank**, a technology that determined the "importance" of a webpage by looking at what other pages link to ...

[www.google.com/corporate/tech.html](#) - [Cached](#) - [Similar](#)

[Check Page Rank!](#) ☆ 🔍

Google **Page Rank** (Google PR) is one of the methods Google uses to determine a page's relevance or importance. Important pages receive a higher **PageRank** and ...

[www.checkpagerank.net/](#) - [Cached](#) - [Similar](#)

[PageRank Checker - Check Your Google Page Rank - iWEBTOOL.com](#) ☆ 🔍

View Google **PageRank** on different Google servers. ... Check a website's Google **PageRank**

Googling Food Webs: Can an Eigenvector Measure Species' Importance for Coextinctions?

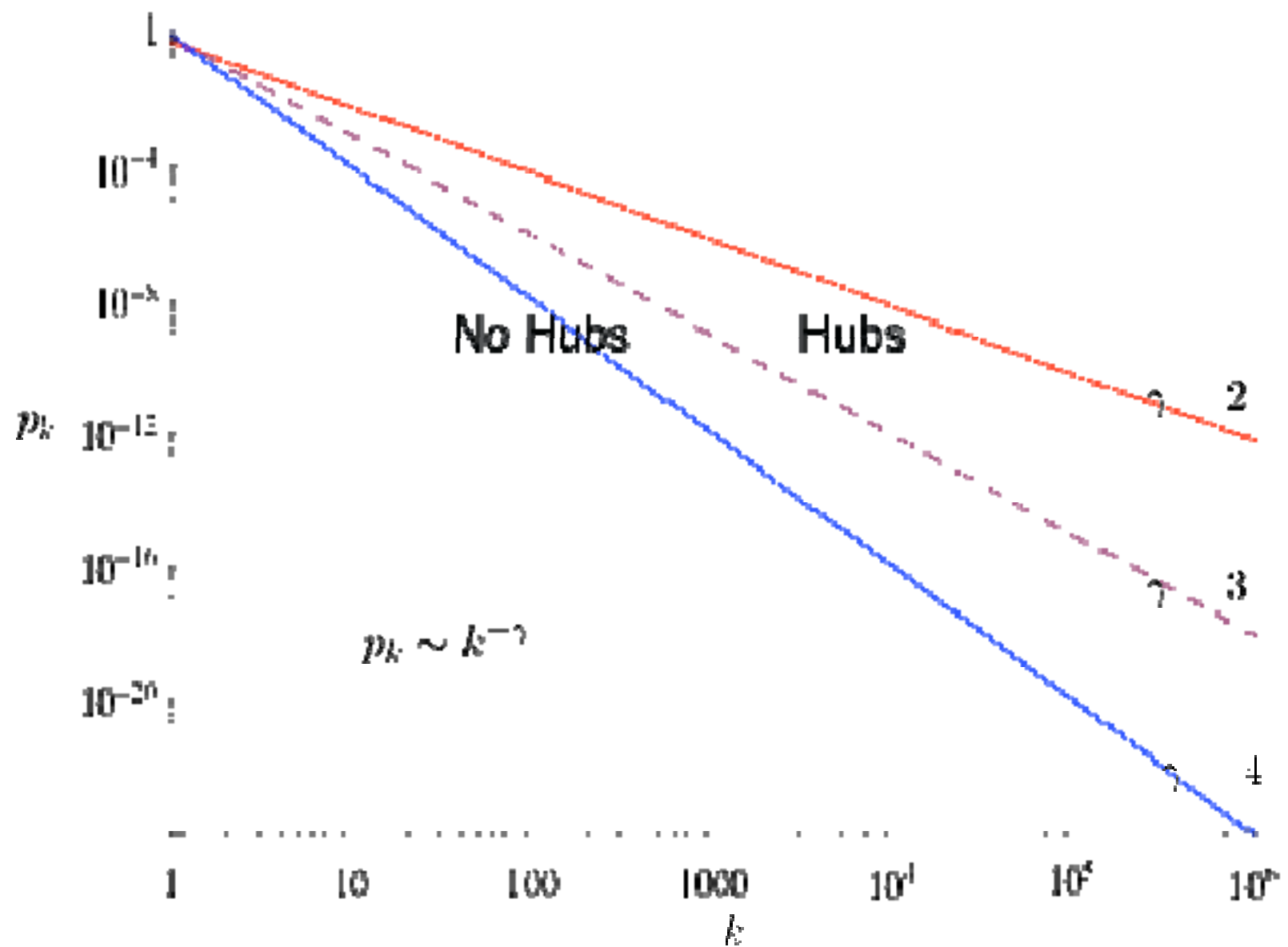
Stefano Allesina^{1*}, Mercedes Pascual^{2,3,4}

1 National Center for Ecological Analysis and Synthesis, Santa Barbara, California, United States of America, **2** Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, Michigan, United States of America, **3** Santa Fe Institute, Santa Fe, New Mexico, United States of America, **4** Howard Hughes Medical Institute

Abstract

A major challenge in ecology is forecasting the effects of species' extinctions, a pressing problem given current human impacts on the planet. Consequences of species losses such as secondary extinctions are difficult to forecast because species are not isolated, but interact instead in a complex network of ecological relationships. Because of their mutual dependence, the loss of a single species can cascade in multiple coextinctions. Here we show that an algorithm adapted from the one Google uses to rank web-pages can order species according to their importance for coextinctions, providing the sequence of losses that results in the fastest collapse of the network. Moreover, we use the algorithm to bridge the gap between qualitative (who eats whom) and quantitative (at what rate) descriptions of food webs. We show that our simple algorithm finds the best possible solution for the problem of assigning importance from the perspective of secondary extinctions in all analyzed networks. Our approach relies on network structure, but applies regardless of the specific dynamical model of species' interactions, because it identifies the subset of coextinctions common to all possible models, those that will happen with certainty given the complete loss of prey of a given predator. Results show that previous measures of importance based on the concept of "hubs" or number of connections, as well as centrality measures, do not identify the most effective extinction sequence. The proposed algorithm provides a basis for further developments in the analysis of extinction risk in ecosystems.

Critical system size N_c



Decoding the tail

Order Statistics

Assume that quantity x distributed according to $p(x)$. The probability to take on a value *greater* than x :

$$P(x) = \int_x^{\infty} p(x') dx'.$$

If we draw N numbers repeatedly from $p(x)$ the probability of a particular draw to be the *largest* number is,

$$\pi(x) = Np(x)[1 - P(x)]^{N-1}.$$

Therefore expectation value (the average) of the largest number is,

$$\langle x_{\max} \rangle = \int_0^{\infty} x \pi(x) dx.$$

Decoding the tail

Order Statistics

We are interested in the m 'th largest number, or m 'th ranked number where, m runs from $1 \dots N$, with $m = 1$ the top rank and $m = N$ the bottom rank.

The probability to be the m 'th ranked number is,

$$\pi_m(x) = \frac{p(x)P(x)^{m-1}[1 - P(x)]^{N-m}}{B(N - m + 1, m)}.$$

As before, the expectation value of the m 'th ranked number is,

$$\langle x_m \rangle = \int_0^{\infty} x \pi_m(x) dx.$$

Order Statistics

Exponential: $p(k) \sim e^{-\lambda k}$.

$$\langle x_m \rangle = \frac{1}{\lambda} (H_N - H_{m+1}) \approx \frac{1}{\lambda} [\log(N) + O(N^{-1})]. \quad \text{where} \quad H_N = \sum_{k=1}^N \frac{1}{k}.$$

Scale-Free: $p(k) \sim k^{-\gamma}$.

$$\langle x_m \rangle \approx N^{1/(\gamma-1)} \frac{\Gamma(m - \frac{1}{\gamma-1})}{\Gamma(m)}.$$

Decoding the tail

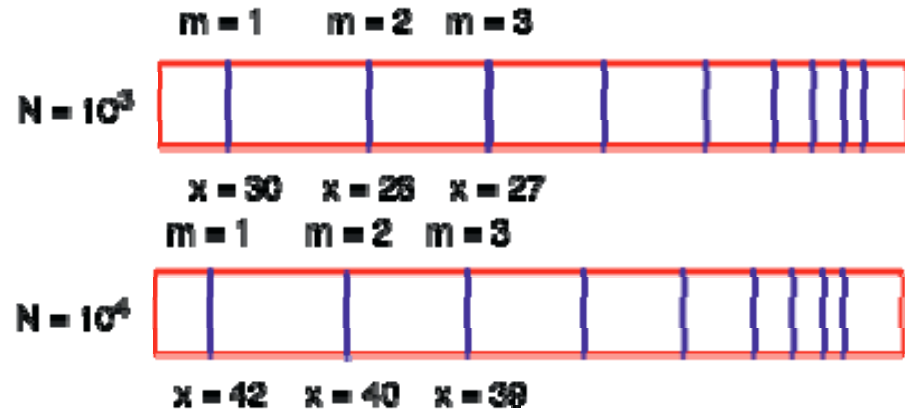
Order Statistics

What about difference between successively ranked values? Define gap as

$$\Delta = \langle x_m \rangle - \langle x_{m+1} \rangle.$$

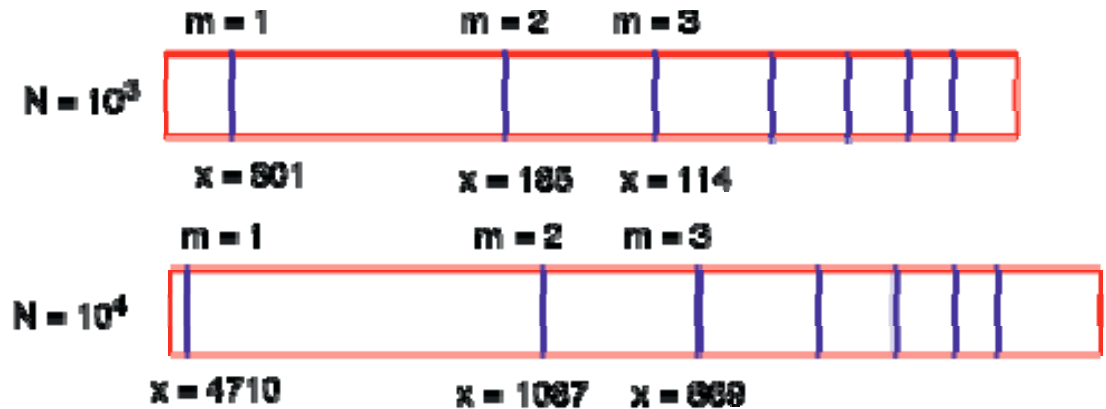
Exponential:

$$\Delta^{\text{exp}} = \frac{1}{\lambda(m+2)}.$$

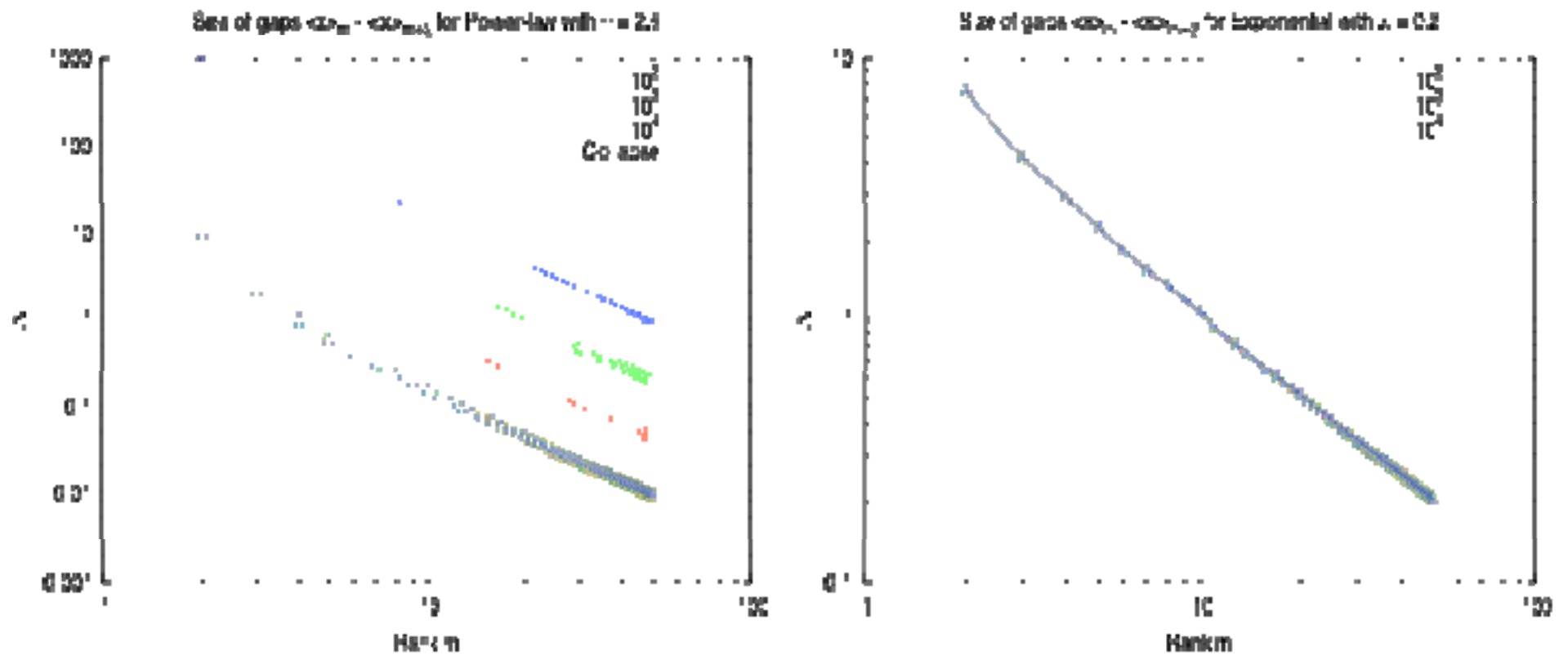


Scale-Free:

$$\Delta^{SF} \approx N^{1/(\gamma-1)} \frac{\Gamma\left(m - \frac{1}{\gamma-1}\right)}{\Gamma(m)} \frac{1}{m(\gamma-1)}.$$

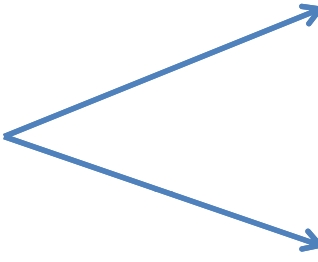


Decoding the tail



Bridging the gap

Connecting the degree distribution and pagerank.

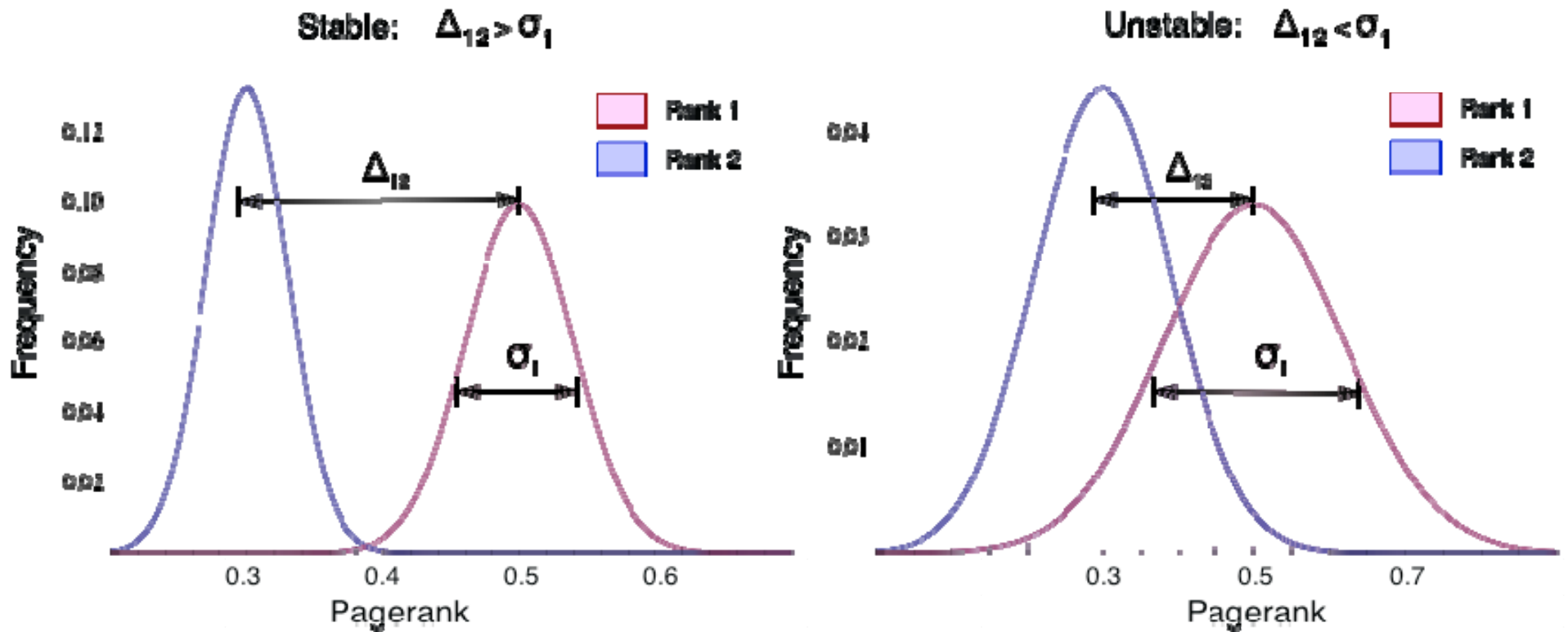
$$\langle x_m \rangle \approx N^{1/(\gamma-1)} \frac{\Gamma(m - \frac{1}{\gamma-1})}{\Gamma(m)}.$$

$$\bar{p}(\mathbf{k}) = \frac{(1-\alpha)}{N} + \frac{\alpha}{N} \times \frac{k_{in}}{\langle k_{in} \rangle}.$$
$$\sigma(\mathbf{k}) \approx \frac{\alpha^2}{N} \left\langle \frac{k_{in}^2}{k_{out}} \right\rangle^{1/2} \langle k_{in}^{-3/2} \rangle \times k_{in}^{1/2}.$$

For example:

$$p_m = \frac{1-\alpha}{N} + \frac{\alpha}{N \langle k_{in} \rangle} \times N^{1/(\gamma-1)} \frac{\Gamma(m - \frac{1}{\gamma-1})}{\Gamma(m)}.$$

The average pagerank of the node ranked m (top-ranked node $m = 1$).

Stability Criteria



$$\sigma(p_m) \leq \Delta(p_m). \quad \text{or} \quad \frac{\Delta(p_m)}{\sigma(p_m)} \geq 1.$$

The importance of $\gamma_c = 3$

Remember the expression for the fluctuations around the average pagerank p_m

$$\sigma(\mathbf{k}) \approx \frac{\alpha^2}{N} \left\langle \frac{k_{in}^2}{k_{out}} \right\rangle^{1/2} \langle k_{in}^{-3/2} \rangle \times k_{in}^{1/2}.$$

Let's calculate it explicitly,

$$\left\langle \frac{k_{in}^2}{k_{out}} \right\rangle = (\gamma_{in} - 1)(\gamma_{out} - 1) \int_1^\infty \int_1^\infty \frac{k_{in}^{2-\gamma_{in}}}{k_{out}^{1-\gamma_{out}}} dk_{in} dk_{out} = \frac{\gamma_{in} - 1}{\gamma_{in} - 3} \times \frac{\gamma_{out} - 1}{\gamma_{out}}.$$

Expression diverges as γ_{in} approaches 3 from above. Consequence of looking at an infinite system.

For a finite system, this must be bounded by the maximum value---degree of the top ranked node...

The importance of $\gamma_c = 3$

By renormalizing with this maximum value, we can calculate precisely *how* the system diverges:

$$\left\langle \frac{k_{in}^2}{k_{out}} \right\rangle = (\gamma_{in} - 1)(\gamma_{out} - 1) \int_1^{K_{in}} \int_1^{K_{out}} \frac{k_{in}^{2-\gamma_{in}}}{k_{out}^{1-\gamma_{out}}} dk_{in} dk_{out} = \pm C_{\pm} \frac{\gamma_{in} - 1}{\gamma_{in} - 3} \times \frac{\gamma_{out} - 1}{\gamma_{out}}$$

Where,

$$C_+(\gamma_{in} > 3) = 1 - N^{(3-\gamma_{in})/(\gamma_{in}-1)} f(\gamma_{in}),$$

$$C_-(\gamma_{in} < 3) = N^{(3-\gamma_{in})/(\gamma_{in}-1)} f(\gamma_{in}) - 1.$$

Most importantly,

$$\lim_{\gamma_{in} \rightarrow 3} \frac{C_{\pm}}{\pm(\gamma_{in} - 3)} = \frac{1}{2} \log(N\pi).$$

The existing network maps are not perfect

Incompletely mapped:

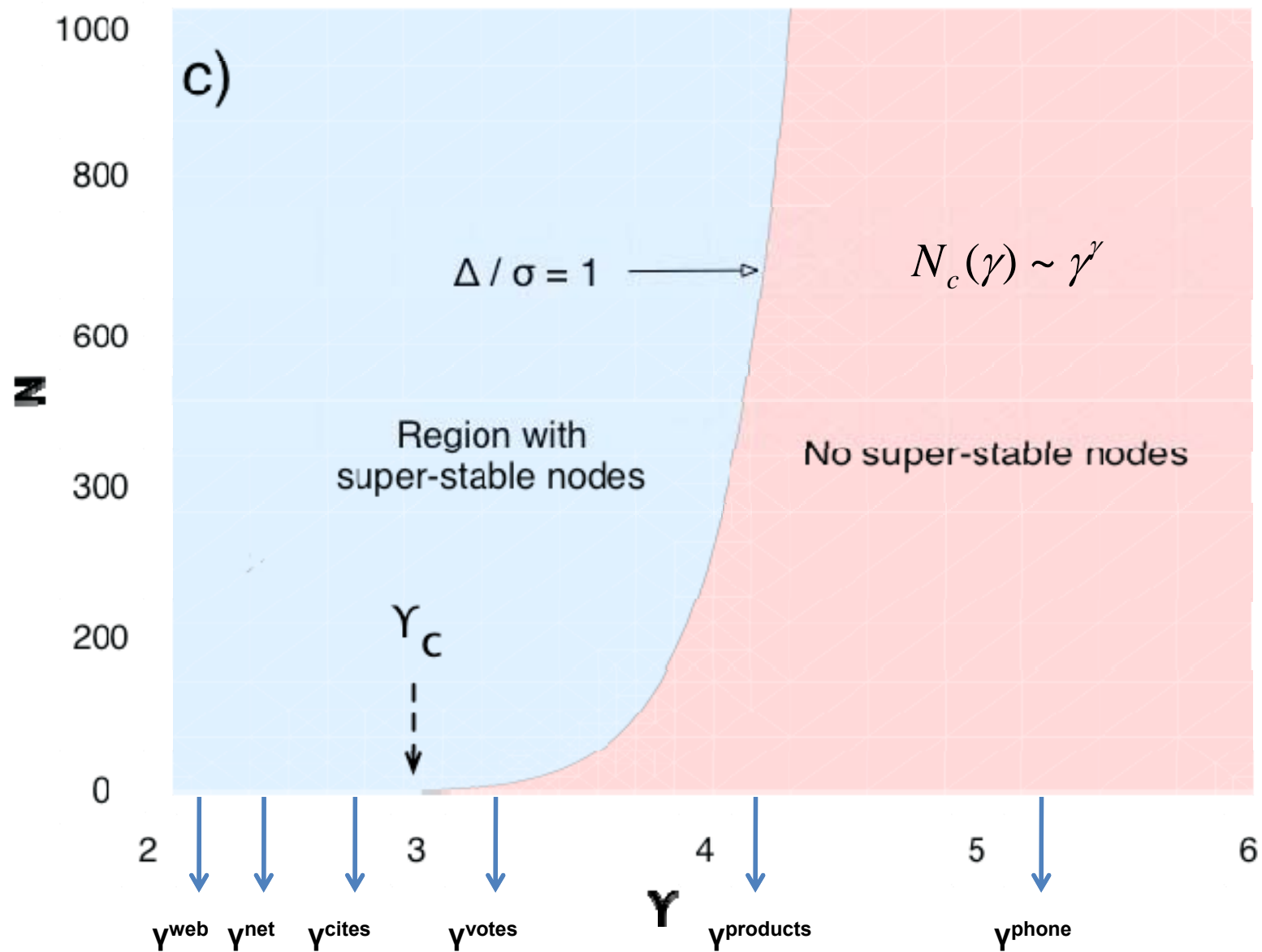
WWW (Lee and Giles' 1999);
Google: maps about XX% of the web
Protein interaction networks (10% coverage)
Food Webs

Noisy:

Protein interaction networks:
high level of false positives and negatives

How stable is pagerank against the most extreme perturbations
that does not change the network's character?

Critical system size N_c

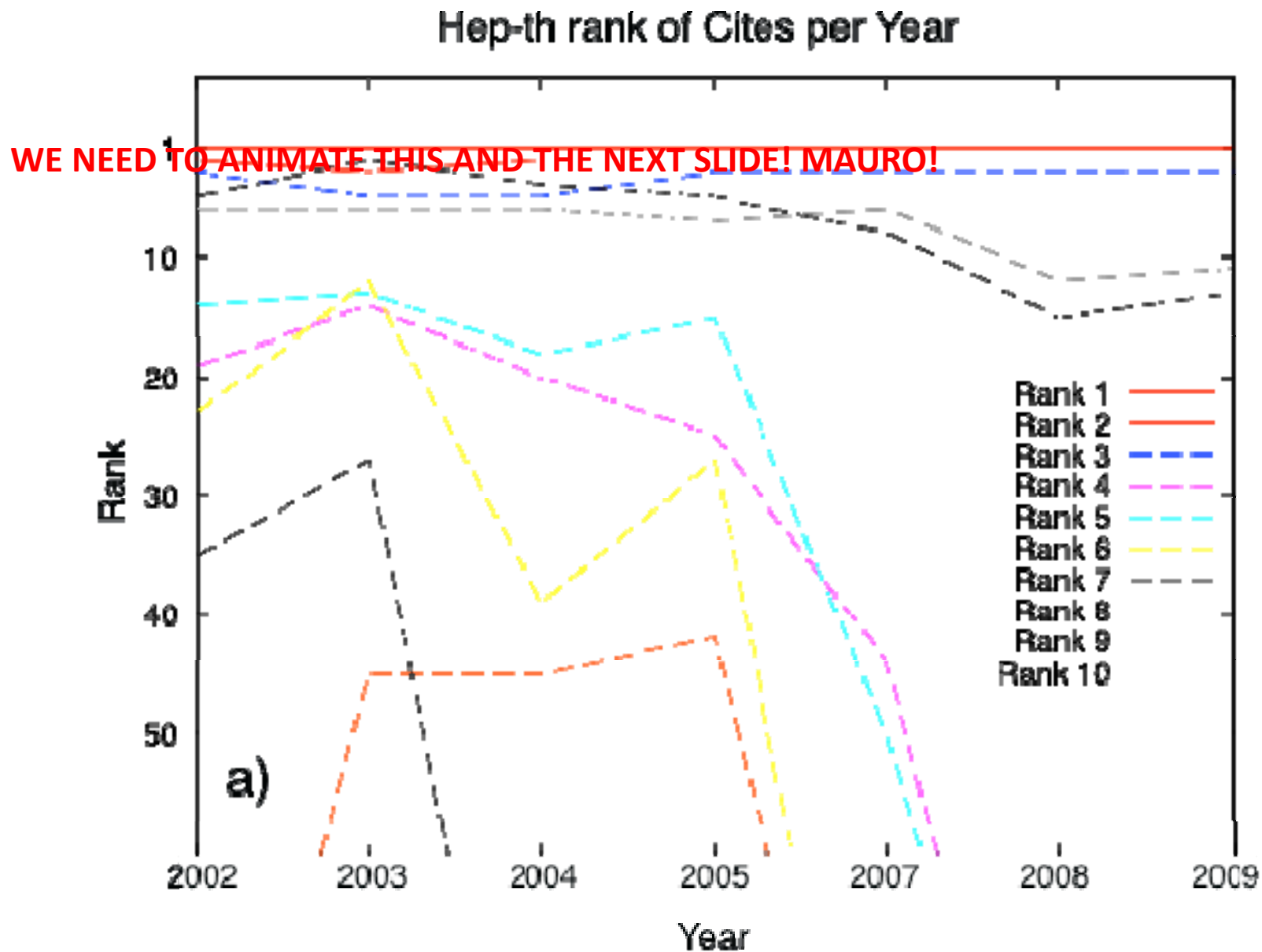


The importance of $\gamma_c = 3$

Three Regimes:

- For $\gamma_{in} > 3$ fluctuations finite, but gap too small.
- For $\gamma_{in} < 3$ gaps are large but fluctuations extensive in system size.
- $\gamma_{in} = 3$ is the “sweet-spot” where gaps are large enough and fluctuations extensive, but only logarithmic in system size.

Temporal stability of super-stable nodes

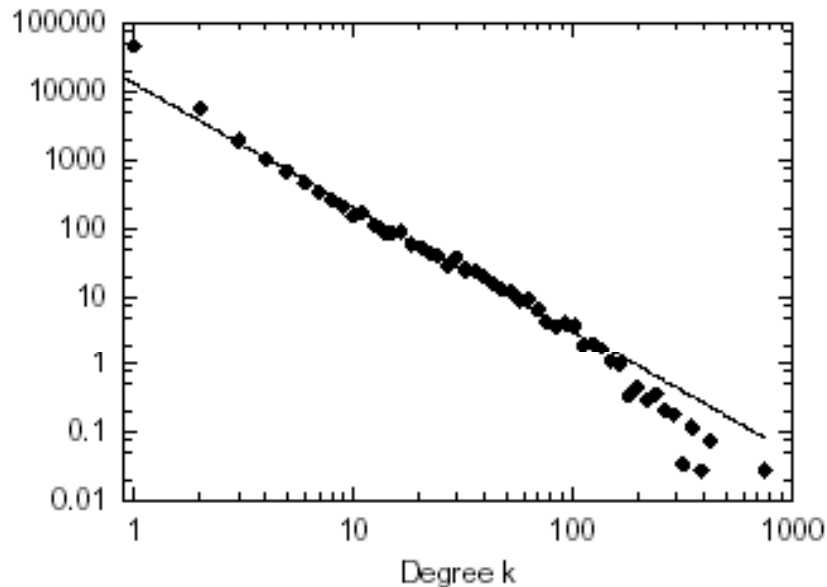


Online communities

Nodes: online user

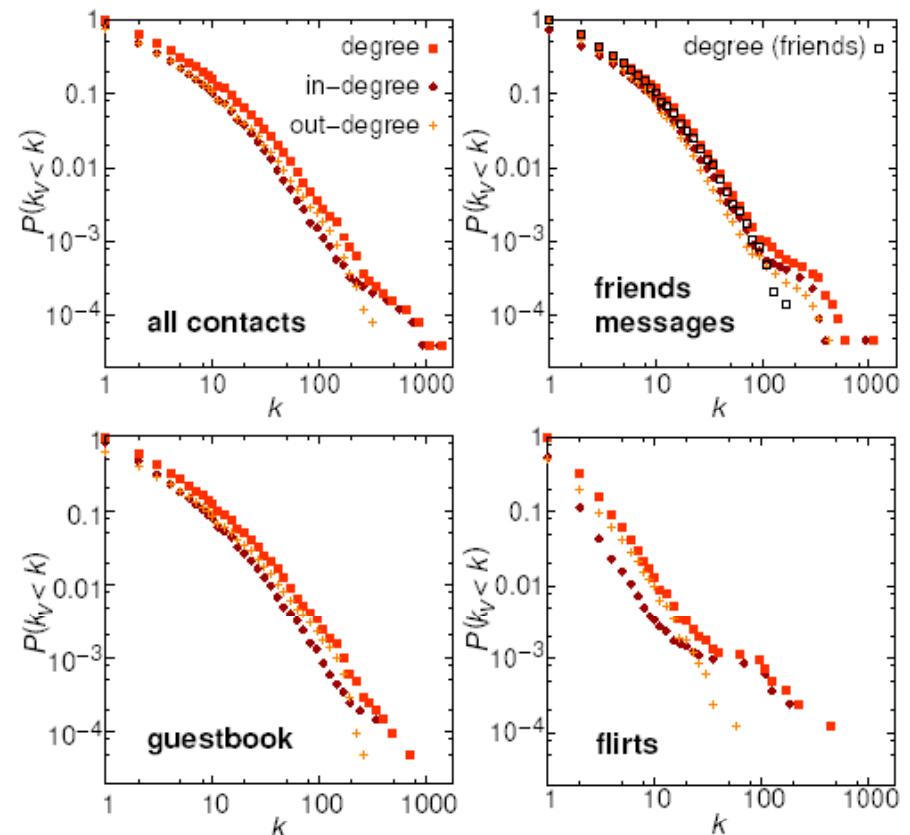
Links: email contact

Kiel University log files
112 days, $N=59,912$ nodes



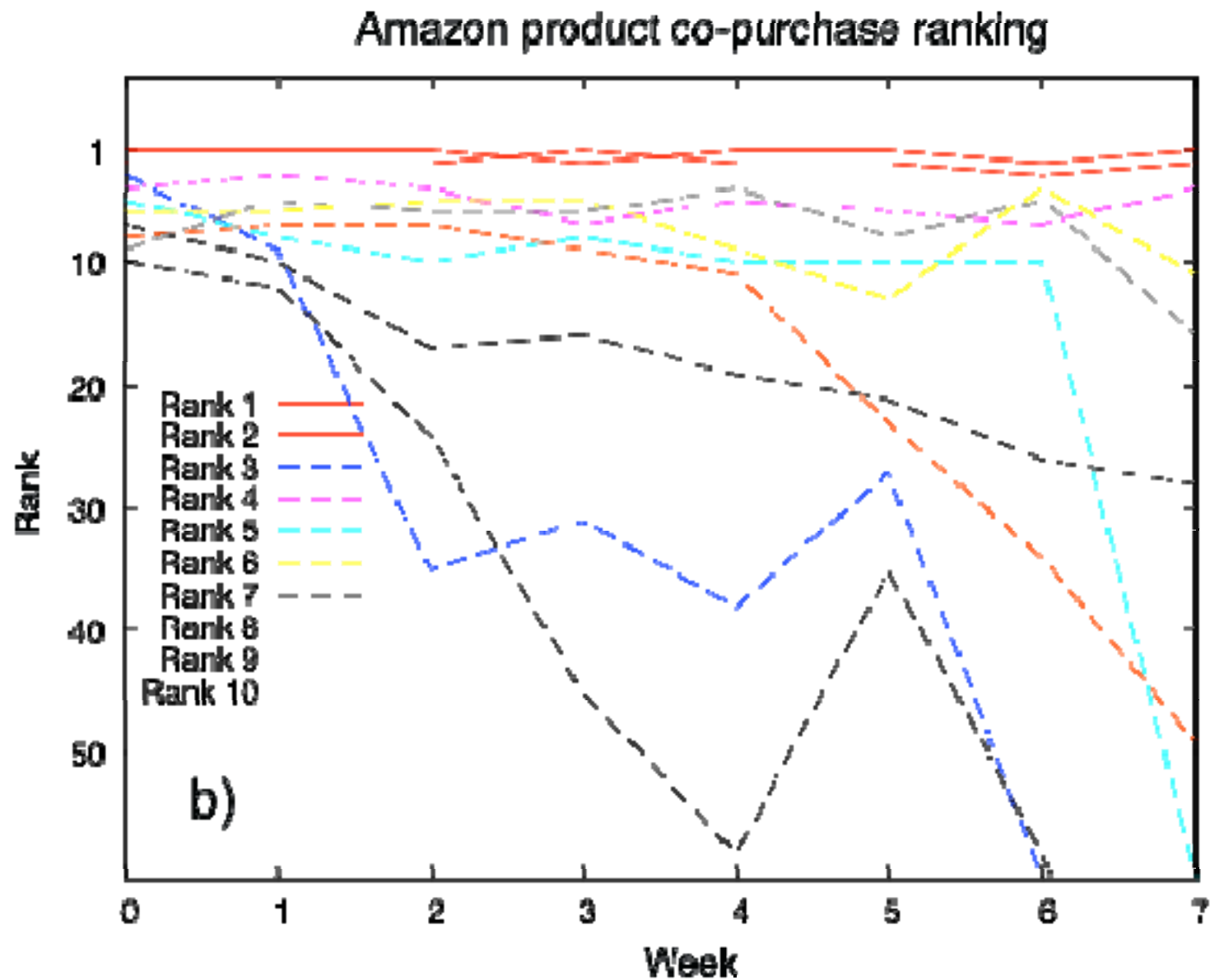
Ebel, Mielsch, Bornholdtz, PRE 2002.

Pussokram.com online
community; 512 days,
25,000 users.



Holme, Edling, Liljeros, 2002.

Temporal stability of super-stable nodes



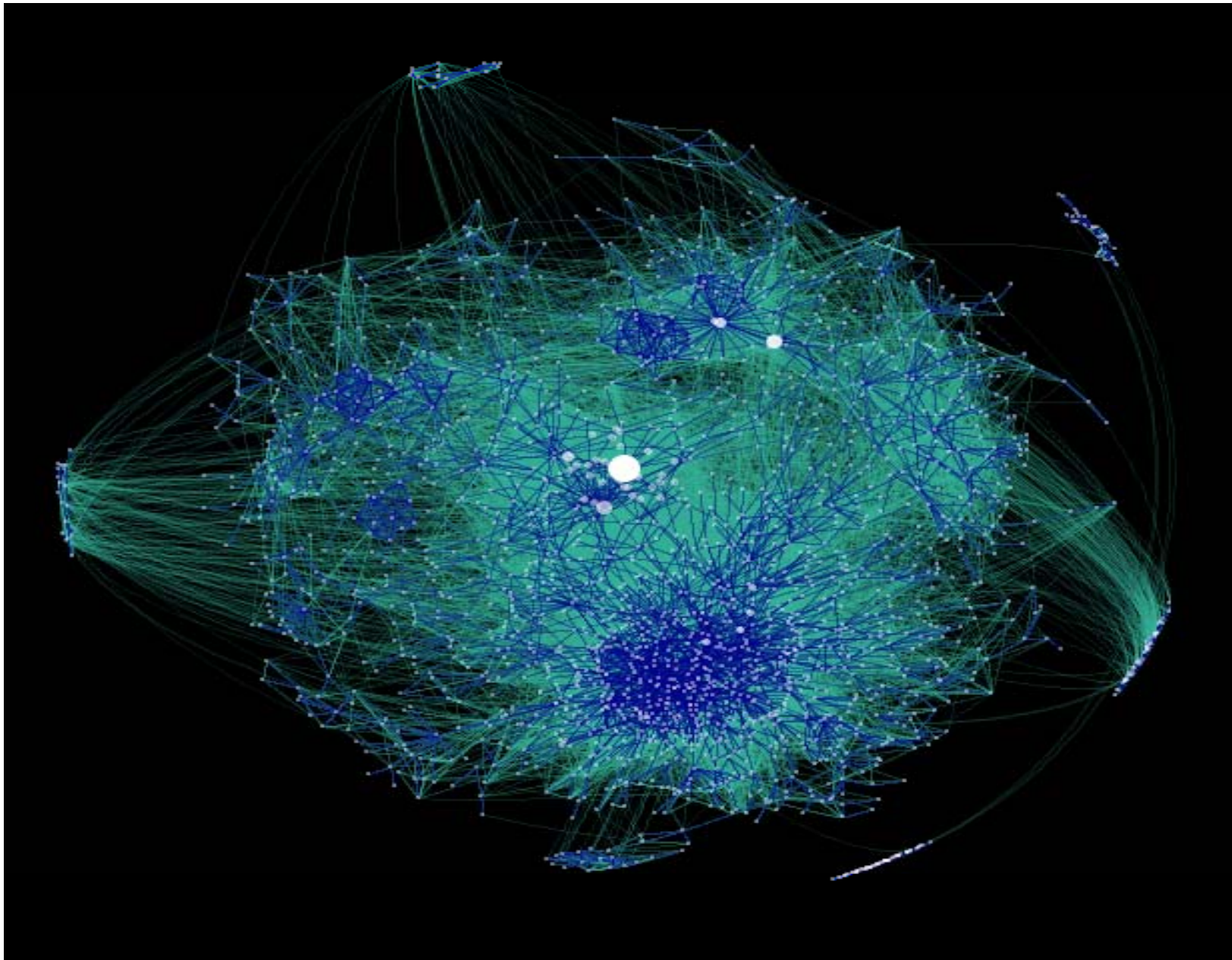


Image by Matthew Hurst

**Many real world networks have a
similar architecture:**

Scale-free networks

WWW, Internet (routers and domains), electronic circuits, computer software, movie actors, coauthorship networks, sexual web, instant messaging, email web, citations, phone calls, metabolic, protein interaction, protein domains, brain function web, linguistic networks, comic book characters, international trade, bank system, encryption trust net, energy landscapes, earthquakes, astrophysical network...

Origin of SF networks: Growth and preferential attachment

- (1) Networks continuously expand by the addition of new nodes

WWW : addition of new documents

- (2) New nodes prefer to link to highly connected nodes.

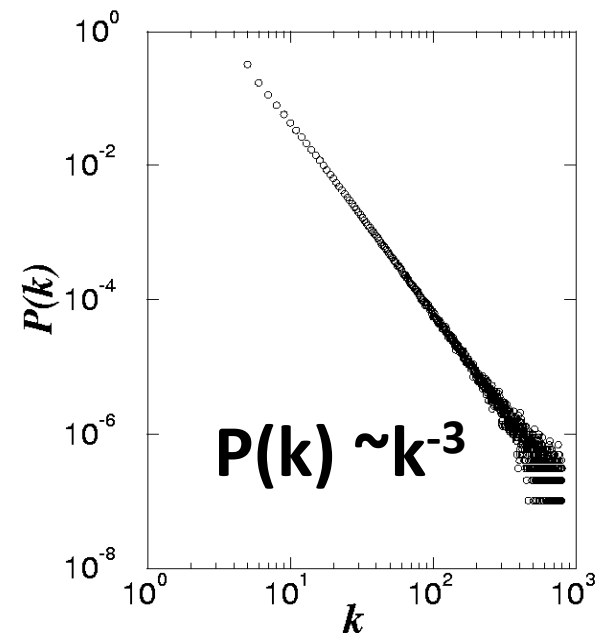
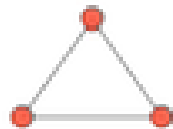
WWW : linking to well known sites

GROWTH:

add a new node with m links

PREFERENTIAL ATTACHMENT: the probability that a node connects to a node with k links is proportional to k .

$$\Pi(k_i) = \frac{k_i}{\sum_j k_j}$$



Barabási & Albert, *Science* **286**, 509 (1999)