Statistical Physics of Citations

Sidney Redner, Boston University
collaborators: P. Krapivsky, F. Leyvraz, P. Chen

Fermilab, September 28, 2005

Observations about scientific citations:

- amusing facts/idle gossip
- analysis of citation data

Preferential attachment network model

Master equation approach:

- degree distributions
- redirection & copying

Google page rank analysis

Summary & Outlook

Phys. Rev Citation Data

353,268 papers, 3,110,839 cites

$\langle \#\ \text{cites} \rangle = 8.81,\ \langle \text{cite age} \rangle = 6.20$

11 papers with > 1000 citations

79 papers with > 500 citations

237 papers with > 300 citations

2340 papers with > 100 citations

8073 papers with > 50 citations

245459 papers with < 10 citations

84144 papers with 1 citation

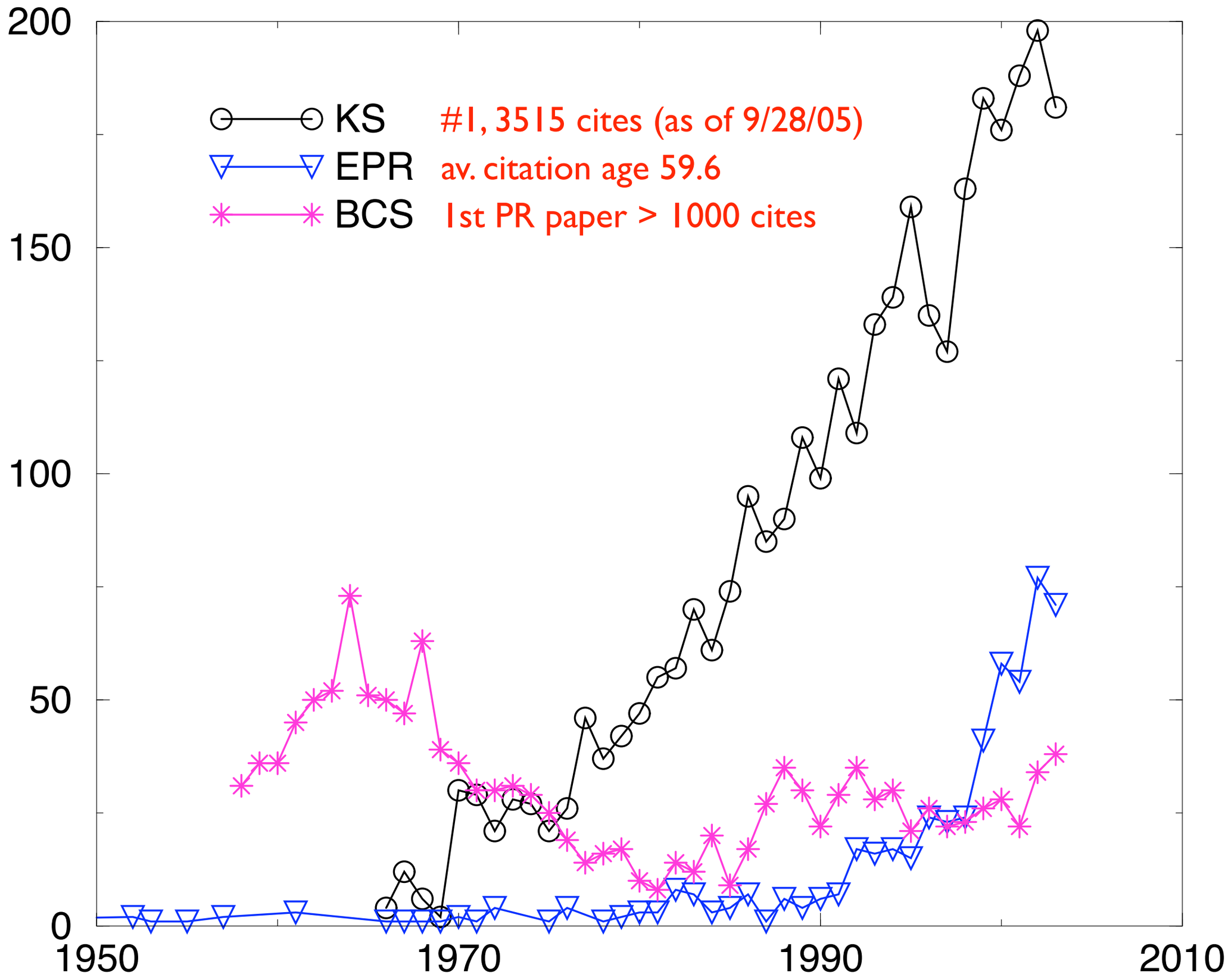
23421 papers with 0 citations

N.B.: ***Internal*** citations only; undercount by factor of 3-5

PR papers with >1000 cites

Cite Rank	Publication				# cites	Av. Age	Impact	Title	Author(s)
1	PR	140	A1133	1965	3227*	26.64	85972	Self-Consistent Equations...	W. Kohn & L. J. Sham
2	PR	136	B864	1964	2460*	28.70	70604	Inhomogeneous Electron Gas	P. Hohenberg & W. Kohn
3	PRB	23	5048	1981	2079	14.38	29896	Self-Interaction Correction to...	J. P. Perdew & A. Zunger
4	PRL	45	566	1980	1781	15.42	27463	Ground State of the Electron ...	D. M. Ceperley & B. J. Alder
5	PR	108	1175	1957	1364	20.18	27526	Theory of Superconductivity	J. Bardeen, L. N. Cooper, & J. R. Schrieffer
6	PRL	19	1264	1967	1306	15.46	20191	A Model of Leptons	S. Weinberg
7	PRB	12	3060	1975	1259	18.35	23103	Linear Methods in Band Theory	O. K. Andersen
8	PR	124	1866	1961	1178	27.97	32949	Effects of Configuration...	U. Fano
9	RMP	57	287	1985	1055	9.17	9674	Disordered Electronic Systems	P. A. Lee & T. V. Ramakrishnan
10	RMP	54	437	1982	1045	10.82	11307	Electronic Properties of...	T. Ando, A. B. Fowler, & F. Stern
11	PRB	13	5188	1976	1023	20.75	21227	Special Points for Brillouin-...	H. J. Monkhorst & J. D. Pack

Citation histories of 3 classic PR papers

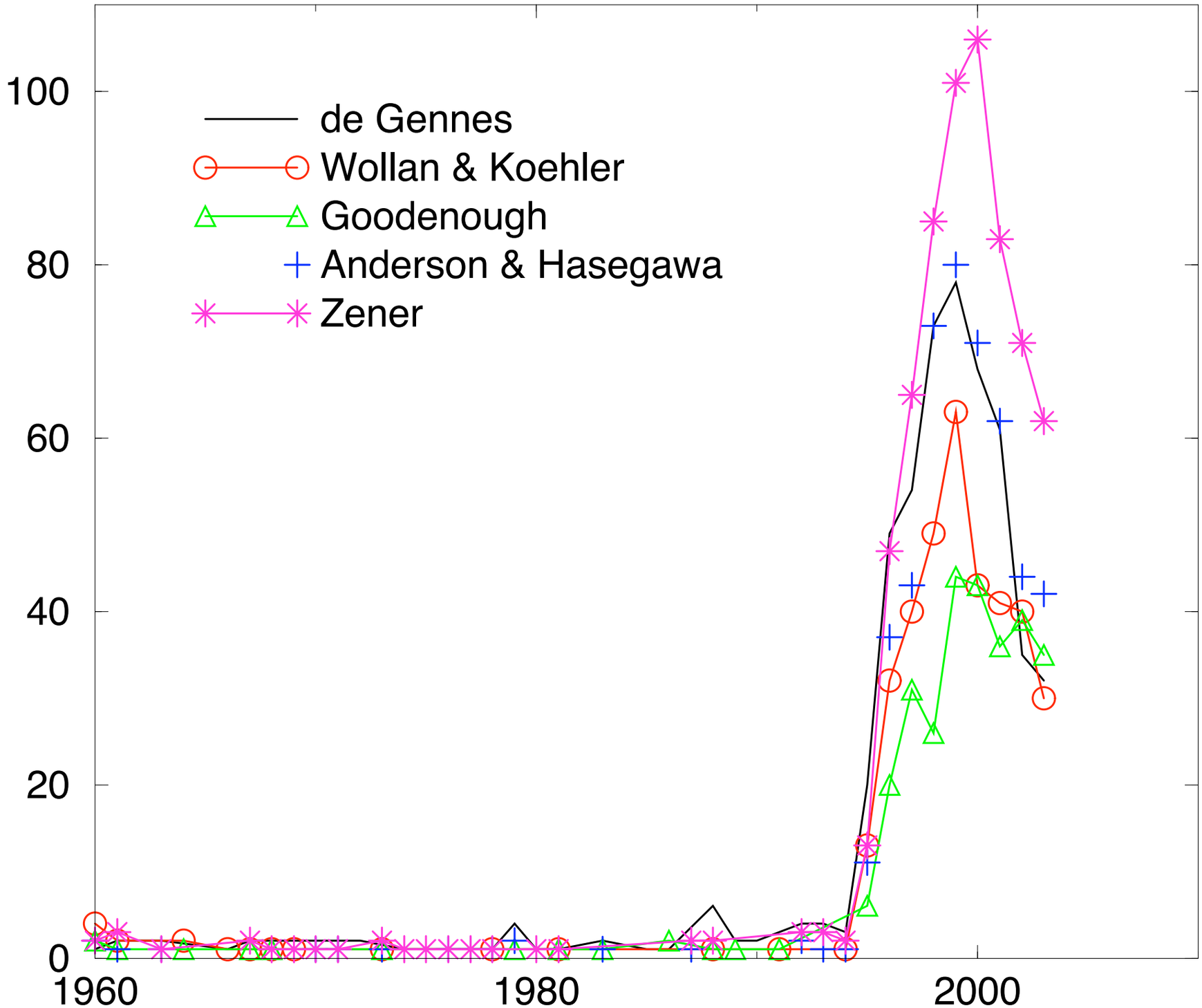


"Sleeping Beauties"

 # cites > 300
<cite age>/paper age > 3/4

8 papers total, 5 on double exchange

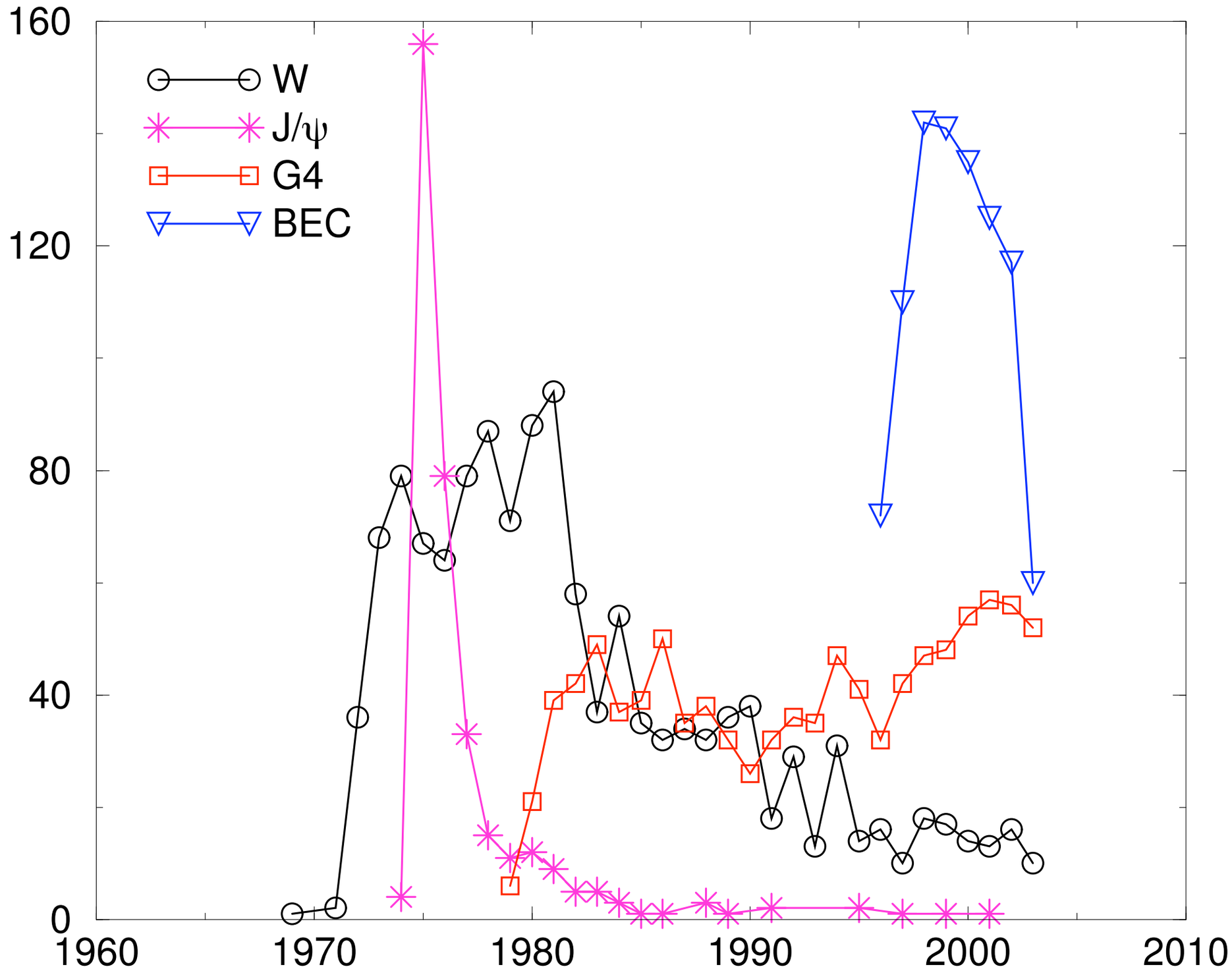
⇒ *colossal magnetoresistance*



Discovery Publications

 # cites > 300
<cite age>/paper age < 0.4

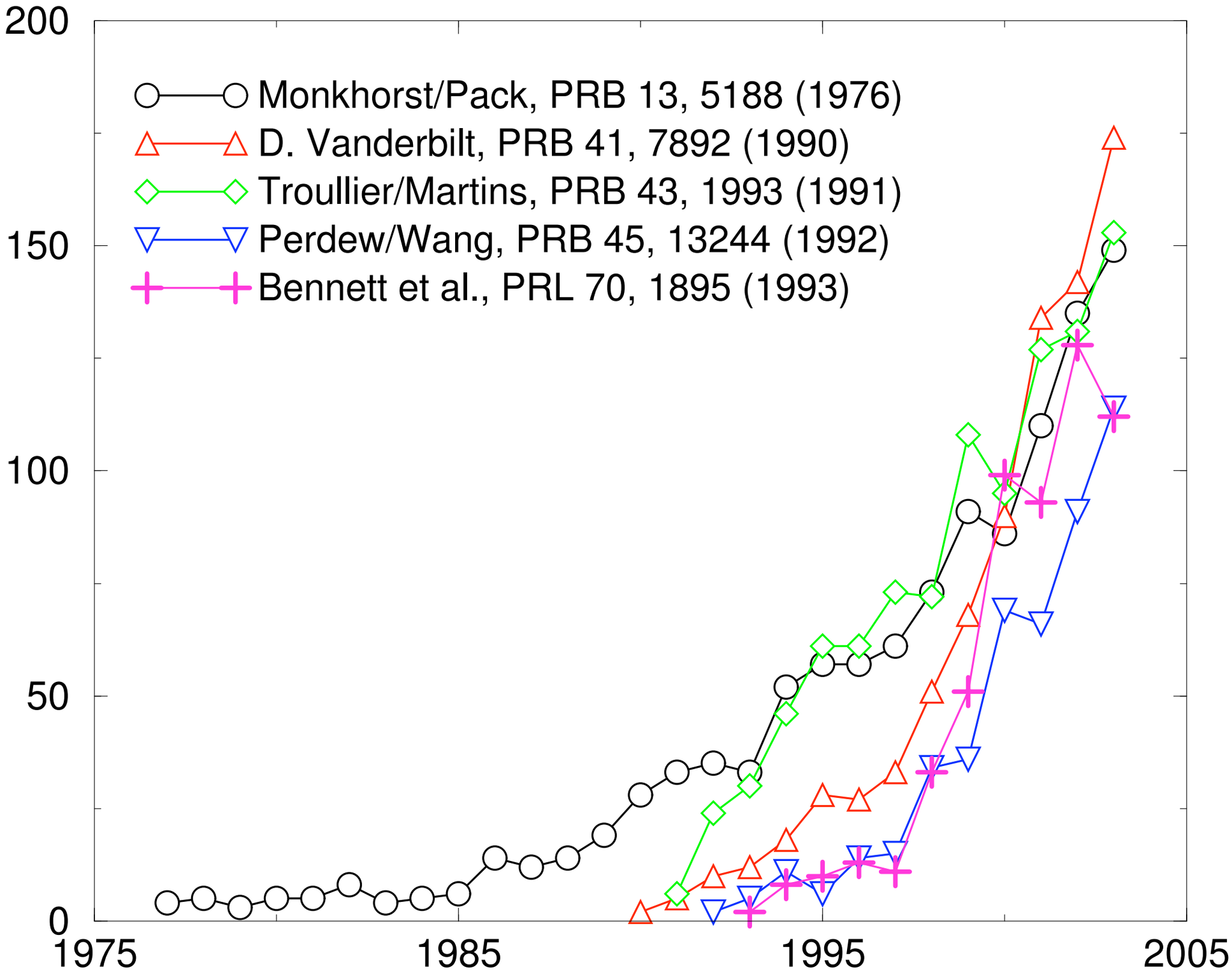
39 papers, 22/25 HEP <1975, all 14 CMP >1975



Hot Publications

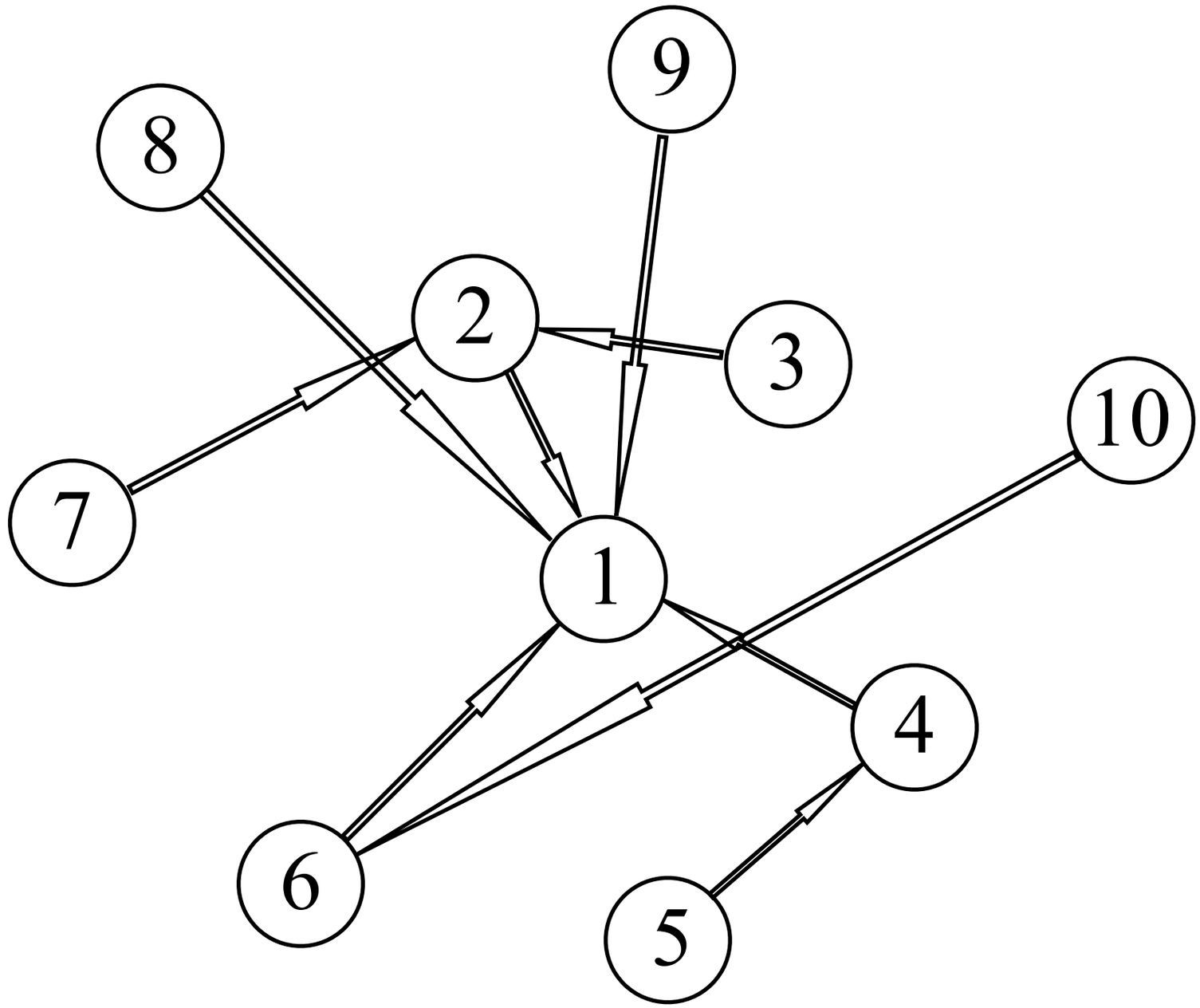
 # cites > 350
citation rate ↑

10 papers total



Growing network model for citations

Simon (1955)
Barabasi & Albert (1999)



1. Introduce nodes one at a time
2. Attach to one earlier node with k links at rate A_k

Master equation approach

KRL (2000)
Dorgovtsev & Mendes (2000)

Basic observable: N_k, the number of nodes with k links
the degree distribution

Master Equation:

$$\frac{dN_k}{dN} = \frac{A_{k-1}N_{k-1} - A_k N_k}{A} + \delta_{k,1}$$

$$A = \sum_j A_j N_j$$

$= \text{total rate}$

For attachment rate: $A_k \sim k^\gamma$

Total Rate: $A = \sum_j A_j N_j = \sum_j j^\gamma N_j \equiv M_\gamma$

Moment equations:

$$\dot{M}_0 \equiv \sum_j \dot{N}_j = 1; \quad \dot{M}_1 \equiv \sum_j j\dot{N}_j = 2$$

These suggest: $A(N) = \sum_j j^\gamma \, N_j \;\; \propto \mu(\gamma) N \quad \text{for } 0 \le \gamma \le 1$

$$N_k(N) \equiv N \; n_k$$

Converts the rate eqns. to linear recursions

$$\frac{dN_k}{dN} = \frac{A_{k-1}N_{k-1} - A_k N_k}{A} + \delta_{k,1}$$

$$\Rightarrow \quad n_k = \frac{A_{k-1}n_{k-1} - A_k n_k}{\mu} + \delta_{k,1}$$

Formal solution: $n_k = \dfrac{\mu}{A_k} \prod_{j=1}^{k} \left(1 + \dfrac{\mu}{A_j}\right)^{-1}$

Asymptotics for $A_k \sim k^\gamma$

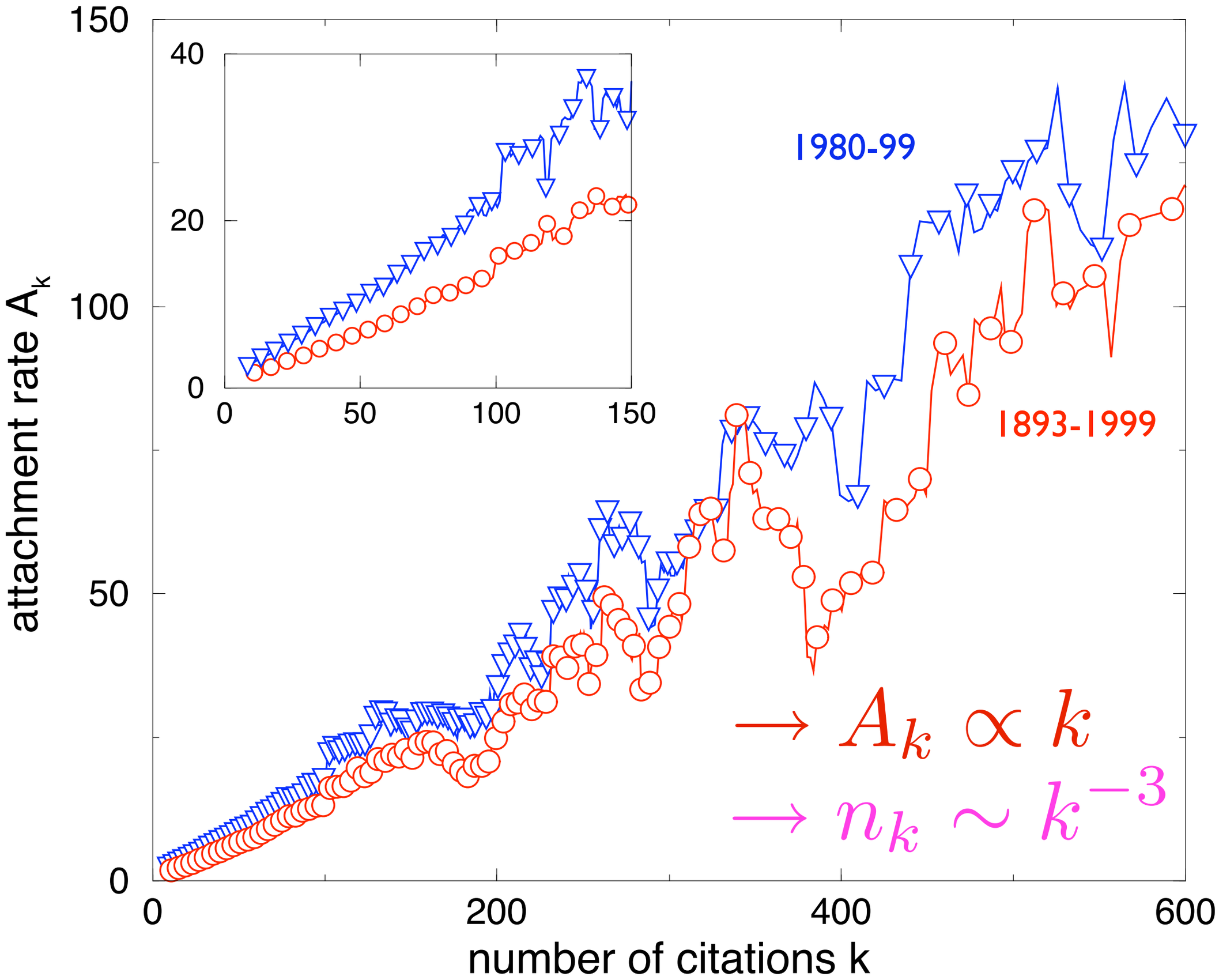
$$n_k \sim \begin{cases} k^{-\gamma} \exp\left[-\mu\left(\frac{k^{1-\gamma} - 2^{1-\gamma}}{1-\gamma}\right)\right] & 0 \le \gamma < 1 \\ k^{-\nu},\ \nu > 2 & \gamma = 1 \\ \text{best seller} & 1 < \gamma \le 2 \\ \text{bible} & \gamma > 2 \end{cases}$$

Important: $n_k \sim k^{-3}$ only for $A_k = k$

If $A_k = k + \lambda$, then $n_k \sim k^{-(3+\lambda)}$ $(\lambda > -1)$

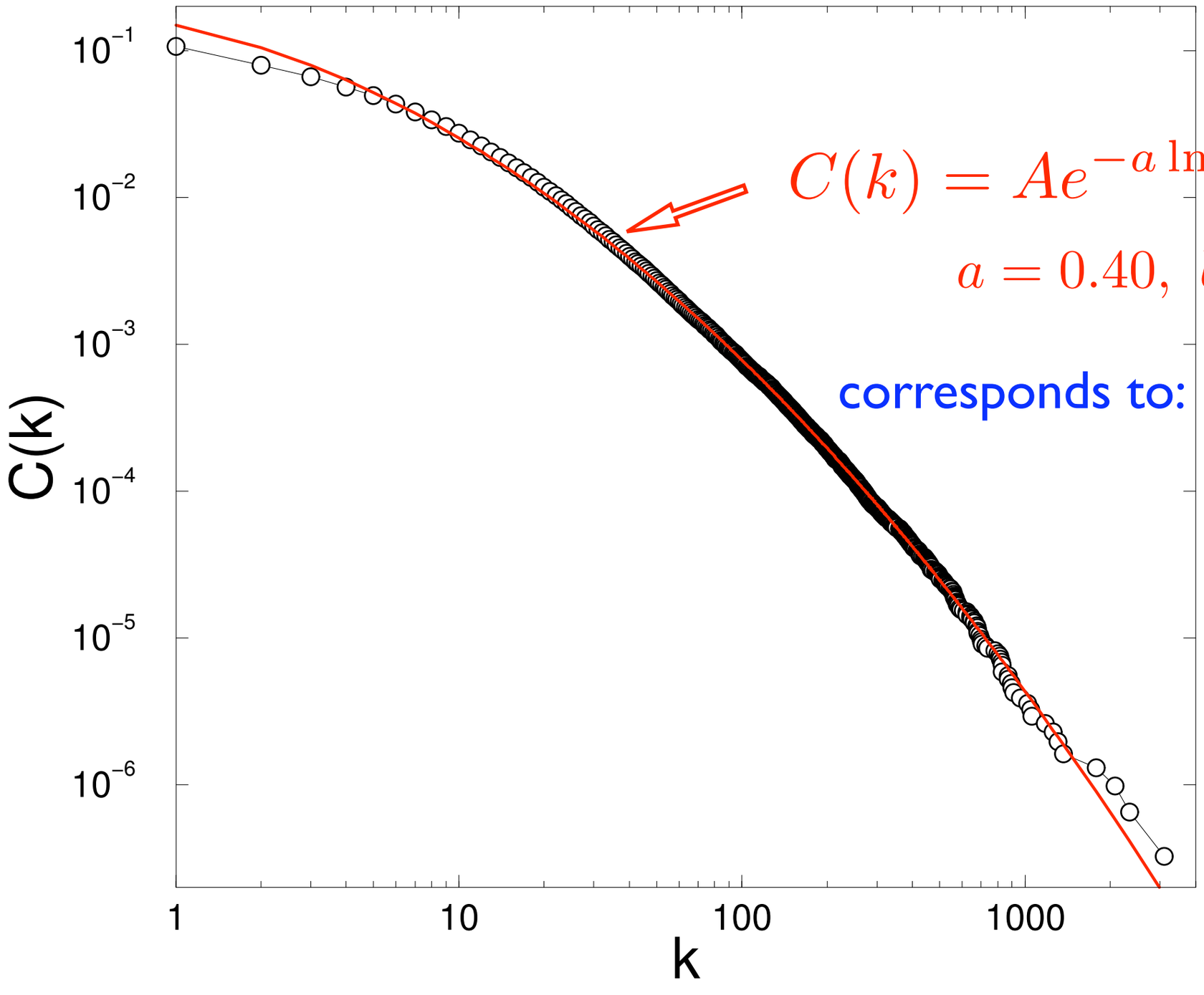
Attachment rate for PR publications

Jeong et al (2003)
SR (2004)



but... a power law is not the whole story

the cumulative citation distribution



$$C(k) = Ae^{-a \ln k - b(\ln k)^2}$$

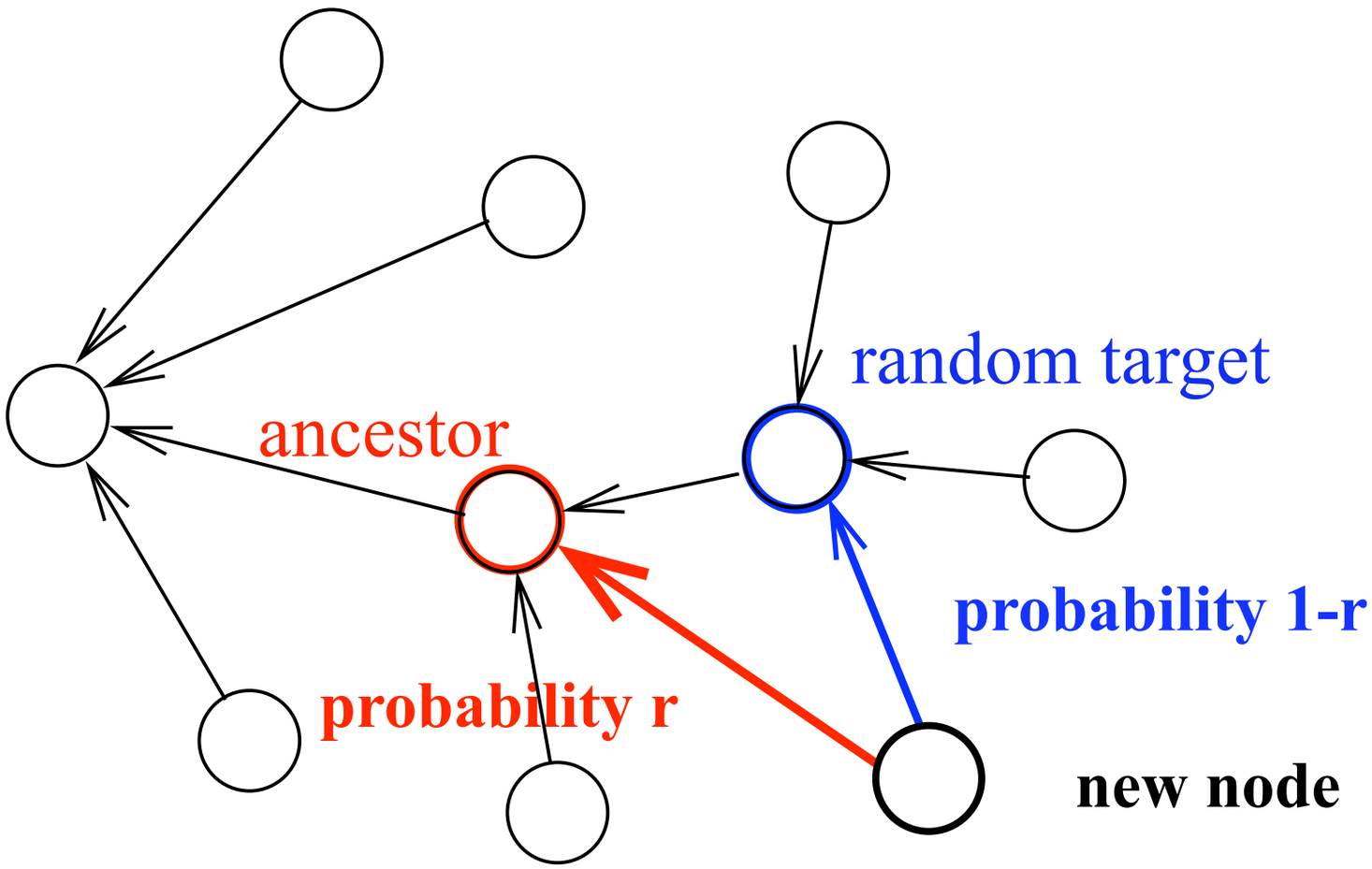
$$a = 0.40, \quad b = 0.16$$

corresponds to: $A_k \propto \dfrac{k}{1 + a \ln k}$

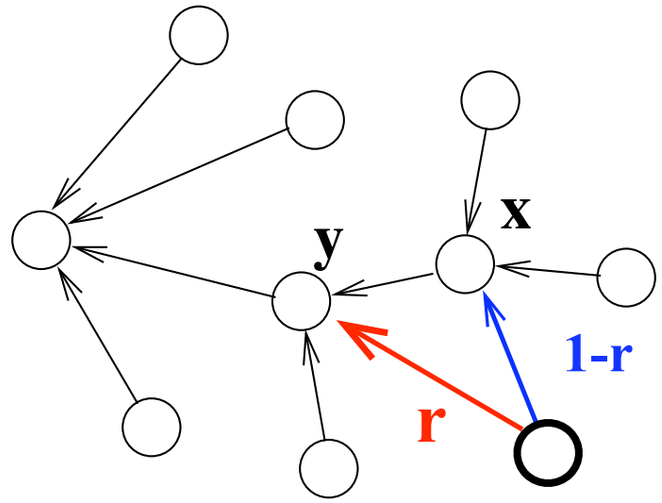
Random attachment + Redirection

Kleinberg et al (1999)
KR (2001)

initial network



Master equations:



$$\frac{dN_k}{dN} = \frac{1-r}{M_0}\left[N_{k-1} - N_k\right]$$

$$+\frac{r}{M_0}\left[(k-2)N_{k-1} - (k-1)N_k\right] \quad + \delta_{k1}$$

$$= \frac{r}{M_0}\left\{\left[(k-1) + \frac{1}{r} - 2\right] N_{k-1} - \left[k + \frac{1}{r} - 2\right] N_k\right\} + \delta_{k1}$$

$\Longrightarrow$ *shifted* linear attachment rate:

$$A_k = k + \left(\frac{1}{r} - 2\right)$$

$$\equiv k + \lambda$$

local rule produces preferential attachment!

substitute into $n_k = \dfrac{\mu}{A_k}\displaystyle\prod_{j=1}^{k}\left(1 + \frac{\mu}{A_j}\right)^{-1} \sim k^{-(3+\lambda)}$ $\quad (-1 < \lambda < \infty)$

Increasingly Entangled Webs

Broder et al (2000)
Broido et al (2002)
Donato et al (2004)

Internet data:

year	1997	1998	1999	2000	2001
# AS	3060	4318	6107	9116	12155
AS links	5302	7874	12037	18196	25179
links/nodes	1.73	1.82	1.97	2.00	2.07

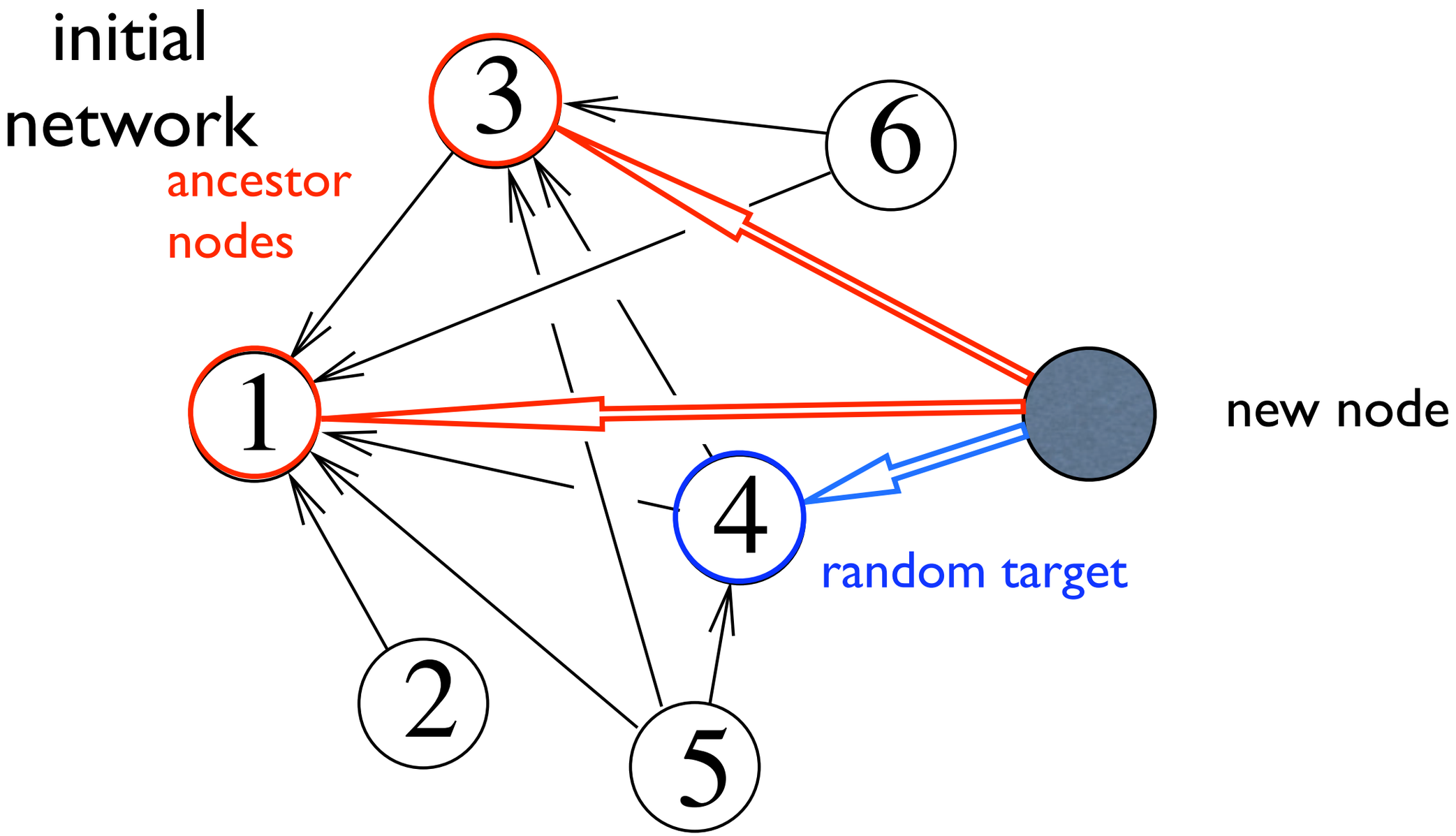
Broido et al (2002)

ratio of links to nodes is growing slowly with time

Random attachment + copying

KR (2004)

a lazy person's approach to references



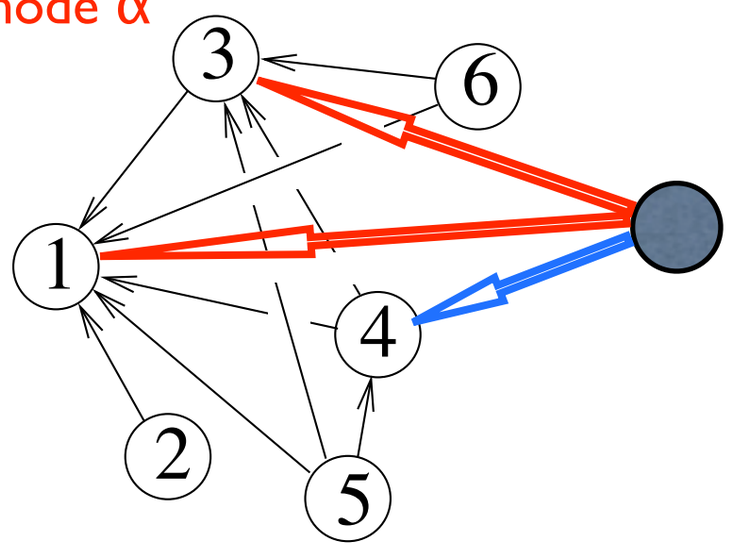
Mean Number of Links L(N)

Evolution equation:

$$L(N+1) = L(N) + \frac{1}{N}\sum_{\alpha}(1+j_\alpha)$$

$$= L(N) + 1 + \frac{L(N)}{N}$$

ancestors of node α



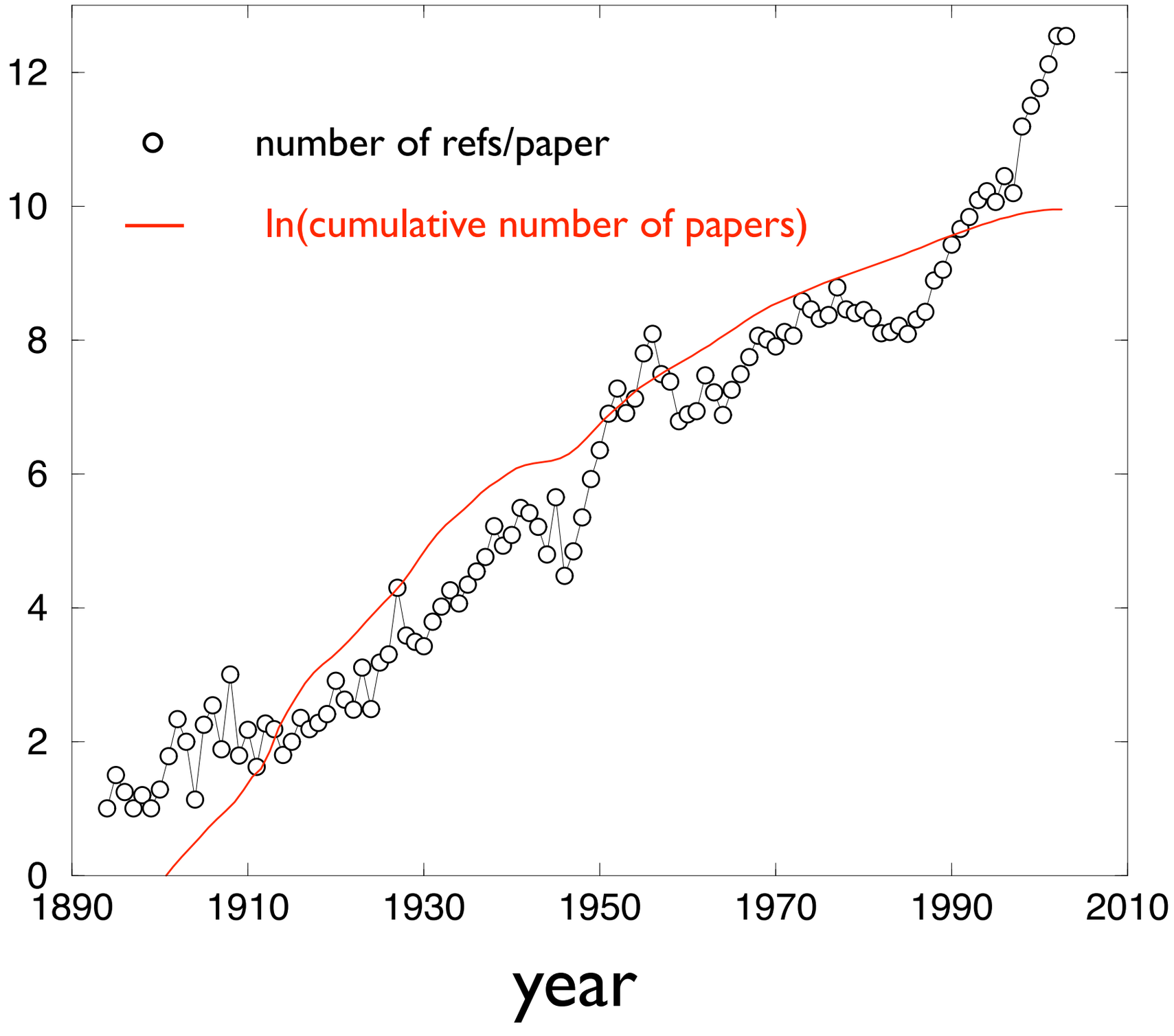
Solution:

$$L(N) = N(H_N - 1)$$

$$= N\ln N - N(1-\gamma) + \frac{1}{2} - \frac{1}{12N} + \cdots$$

$$\rightarrow \text{Degree} = L(N)/N \propto \ln N$$

Comparison with PR citation data



Google Page Rank of Citations Brin & Page (1999)

Basic equation:

$$G_i = \frac{1}{1+N\alpha}\left(\alpha + \sum_j \frac{G_j}{k_j^{\text{out}}}\right)$$

normalization

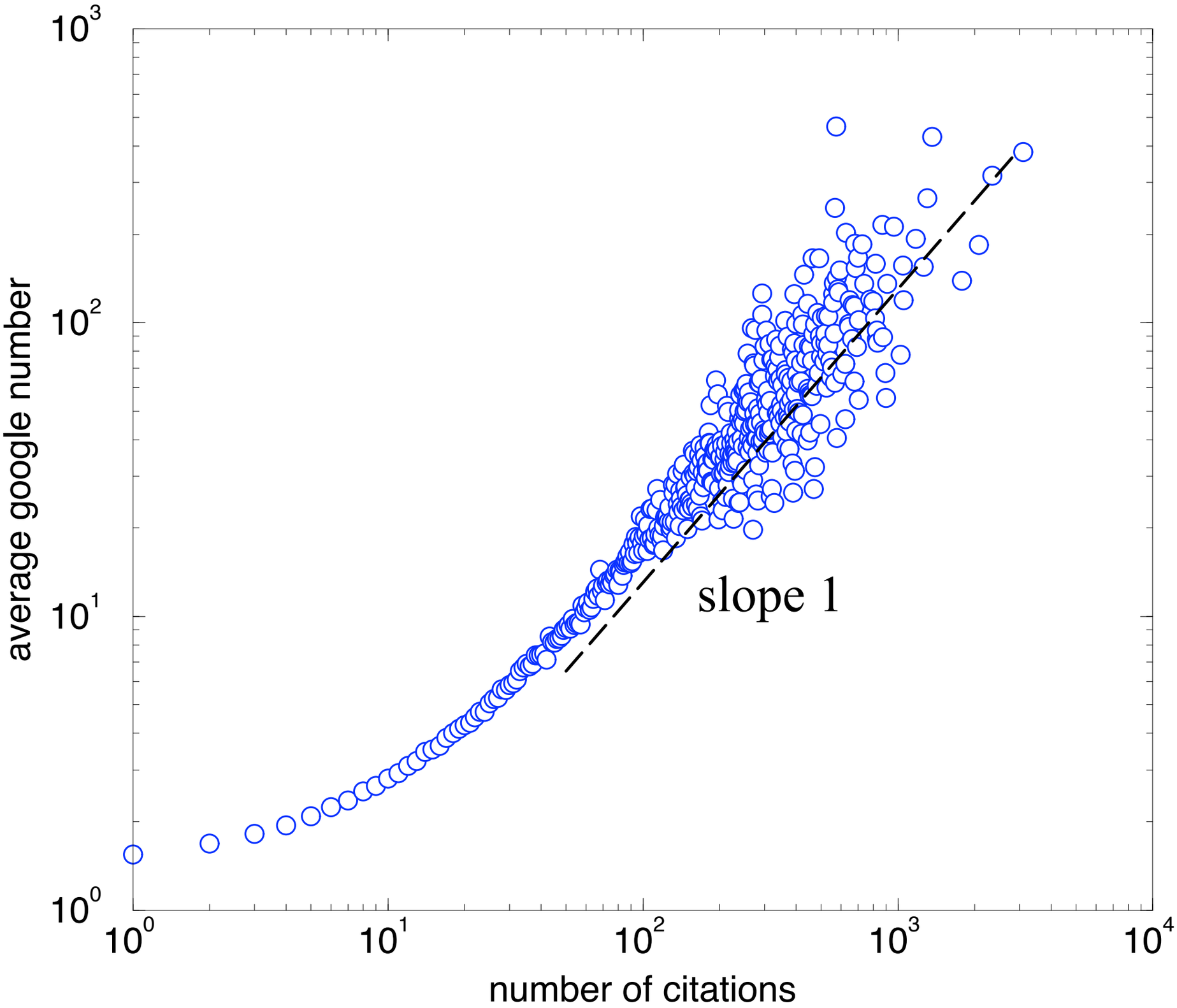
"manna from heaven"

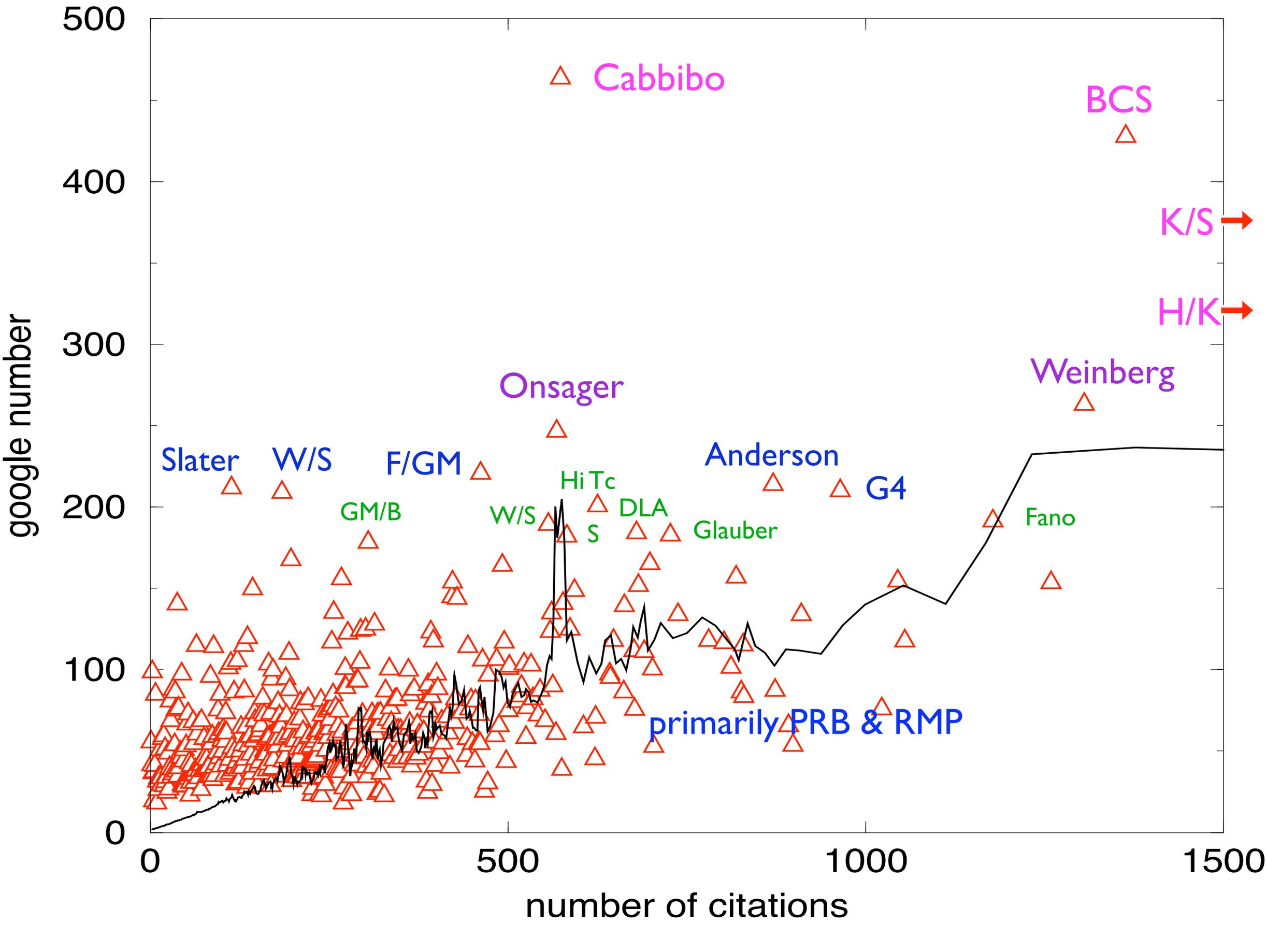
random walk propagation

For $\alpha = 0$ & undirected network: $G_i \propto \text{degree}_i$

For $\alpha > 0$ & directed network: $G_i = f(\text{global topology})$

Correlation between Google & citation counts




500
400
300
200
100
0
0
500
1000
1500
google number
number of citations
Cabbibo
BCS
K/S
H/K
Weinberg
Onsager
Slater
W/S
F/GM
Hi Tc
Anderson
G4
GM/B
W/S
S
DLA
Glauber
Fano
primarily PRB & RMP

Summary & Outlook

Large-scale citation analysis motivates and tests current theories of growing networks

Master equations: an incisive technique to probe geometric properties of networks

Page rank analysis: helps uncover hidden "gems"

For the future: Deeper analysis of citation data:

contextual information, specialization

Larger data sources:

test universality of citation statistics