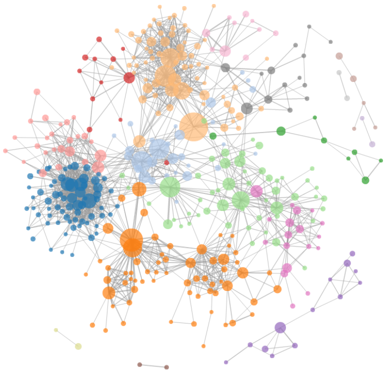


Citation networks as a window to science: a case study

Remco van der Hofstad

Advances in Applied Probability, ICTS, Bengaluru,
August 5–17, 2019

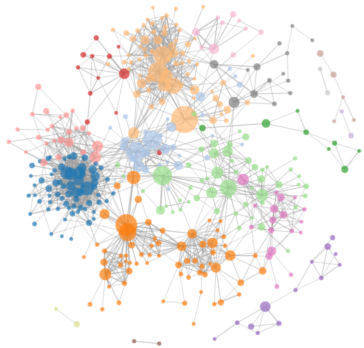
Joint work with
▷ Alessandro Garavaglia (TU/e)
▷ Nelly Litvak (TU/e & U Twente)
▷ Gerhard Woeginger (Aachen)

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY
EURANDOM
NETWORKS

Citations

Citation counts contain important information, yet are hard to interpret:

- ▷ Depend sensitively on age scientists;
- ▷ Highly field dependent (even differences within small subfields);
- ▷ Many good papers with few, and bad papers with many citations;
- ▷ Metrics (such as h-index and journal impact factors) have obvious limitations.



Network of sociology

Neal Claren

http://www.unc.edu/~ncaren/cite_network/cites.html

Let's make science metrics more scientific.
Julia Lane. Nature, **464**:488-489, (2010)

Look at the data!

▷ Investigate citation network dynamics:

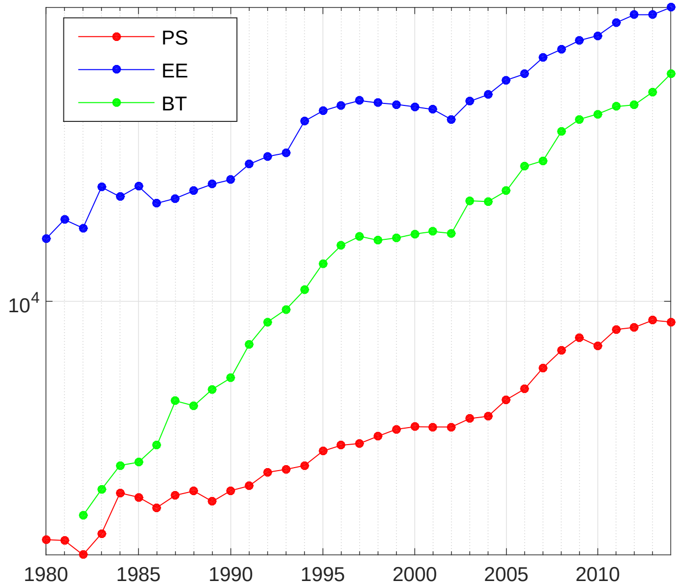
How many citations do papers receive?
What is variability in citation counts?
How long does it take for papers to receive citations?
When is your paper forgotten?

▷ Restrict to homogeneous domains of science:

Probability and statistics
Electrical engineering
Biomedical technology

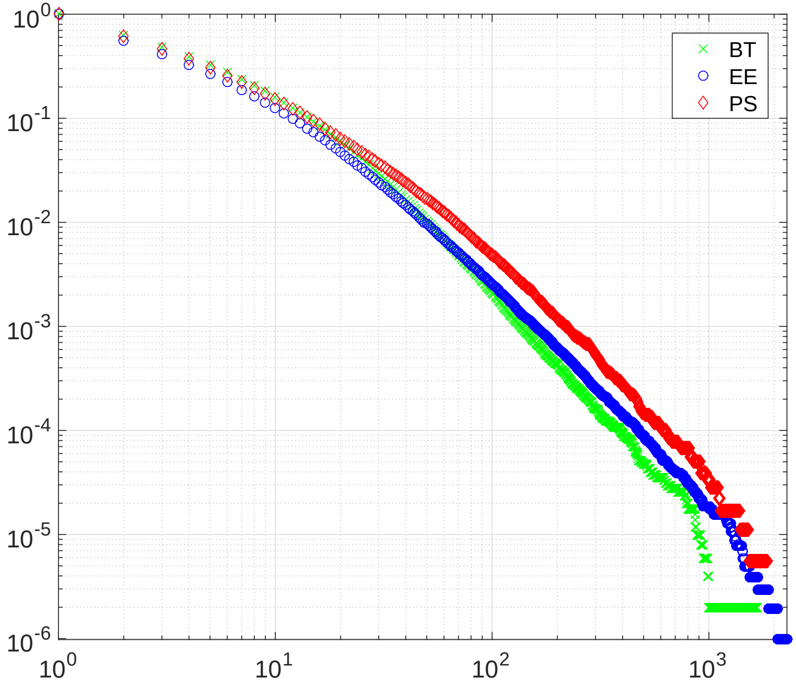
On basis of Web of Science data [not good for certain fields such as CS]:
40 M papers with 500 M citations starting in 1980.
Courtesy of CWTS Leiden (Ludo Waltman)

Number of papers



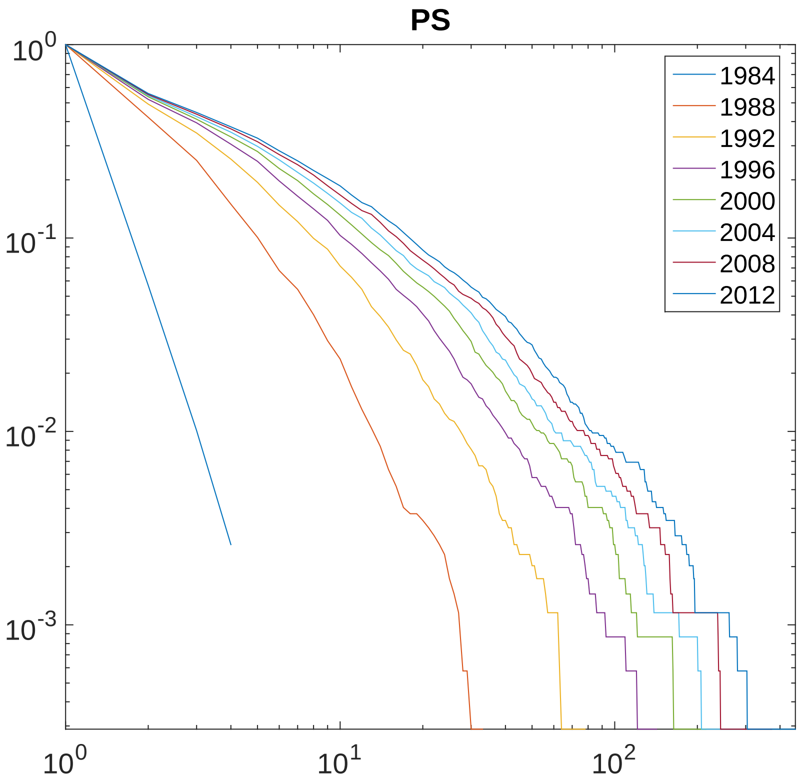
Exponential growth of number of publications.
Already observed by Derek De Solla Price in his 1963 book
'Little Science, big science'

Citations of papers



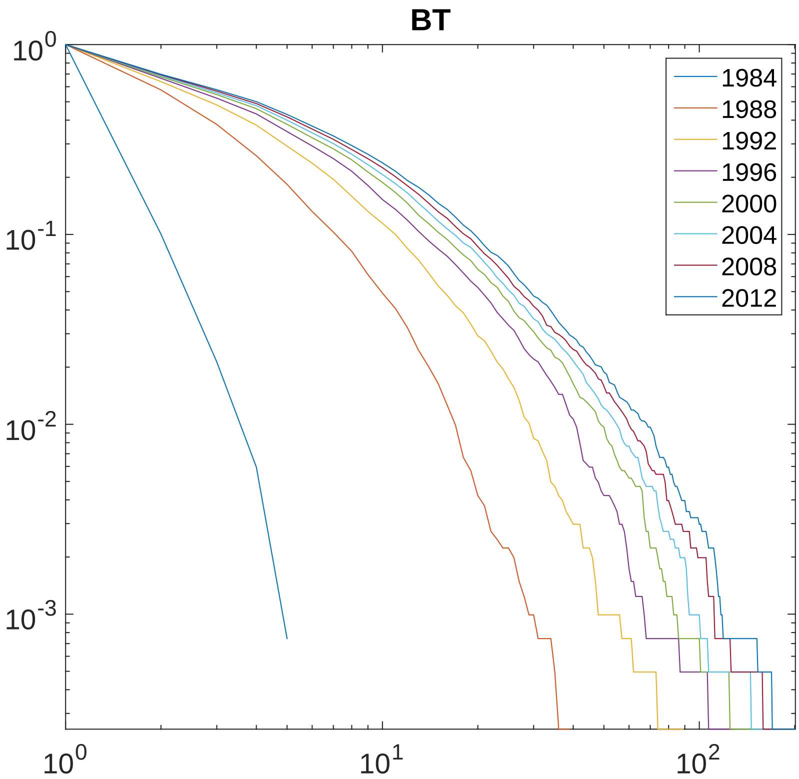
Extreme variability in citation distributions: Power laws?

Dynamic power laws



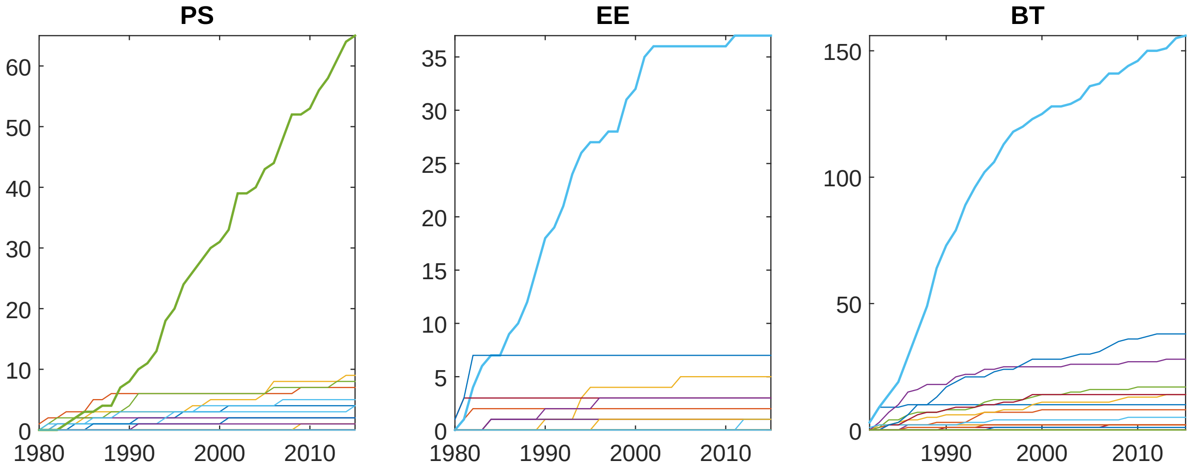
Power laws change over time:
Citation distribution of papers from 1984 in Probability and Statistics

Dynamic power laws



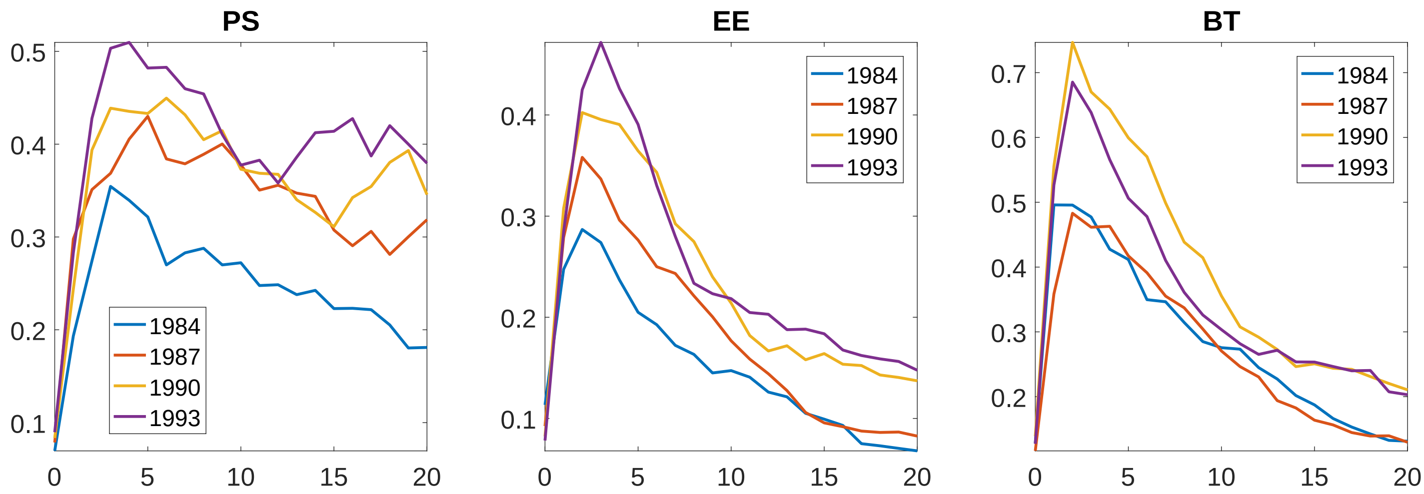
Power laws change over time:
Citation distribution of papers from 1984 in Biomedical Technology

Evolution of citations



Evolution citations over time:
Random sample of 20 papers from 1980

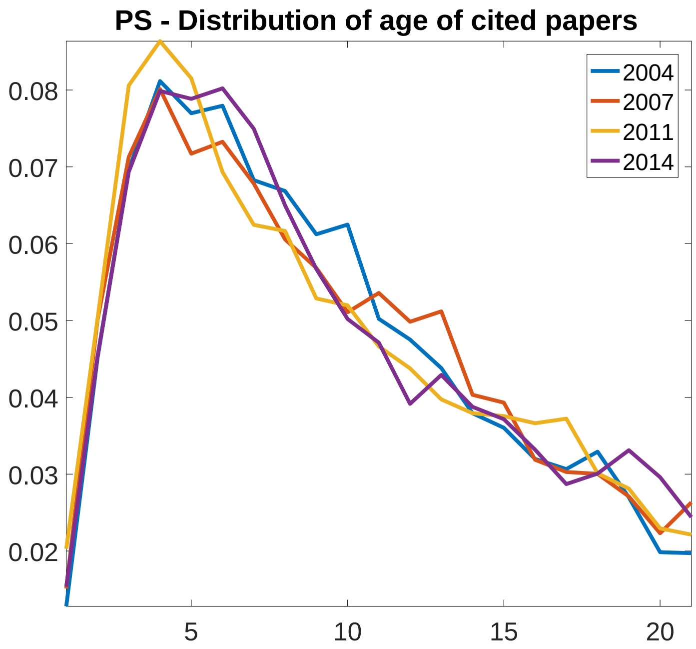
Evolution of citations



Average citation increment over a 20-years time window for papers published in different years

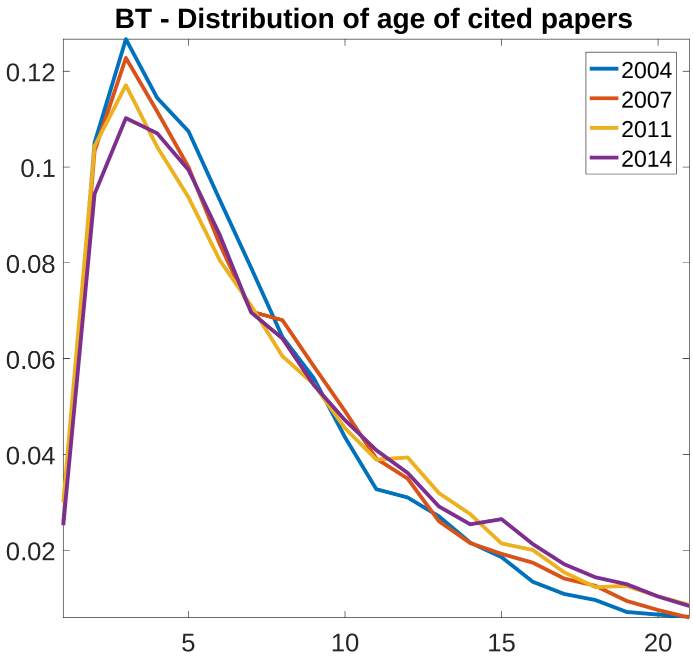
Looks relatively homogeneous

Aging of citations



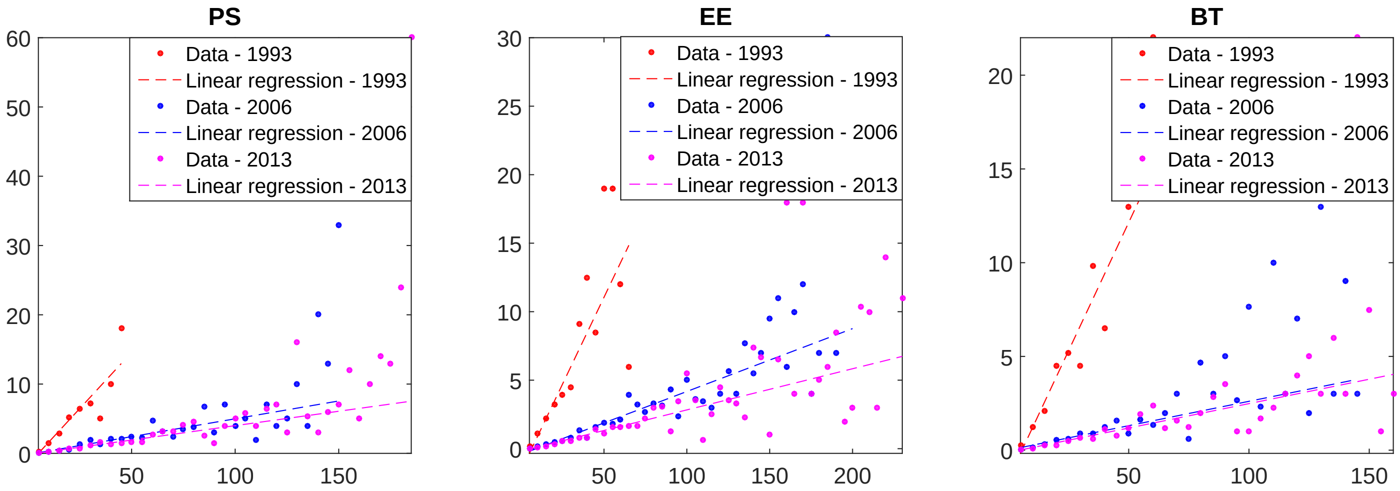
Distribution of age of cited papers for different publication years:
Probability and Statistics: log-normal?

Aging of citations



Distribution of age of cited papers for different publication years:
Biomedical Engineering: log-normal?

Almost linear growth citations



Average number of citations received by papers published in 1984 in 1993, 2006 and 2013 according to total citations up to same year.

Conclusions

▷ Number of papers grows almost exponentially;

▷ Citations per paper vary tremendously;

▷ Citation counts follow approximate
power-law distribution,
with exponent changing over time;

▷ Papers stop receiving citations after (variable) time;

▷ Age of cited papers looks roughly log-normal;

▷ Reasonable prediction that citations grow
almost linearly in time given past.

Modeling networks

Use random graphs to model uncertainty in formation connections between elements.

▷ Static models:
Graph has fixed number of elements:

Erdős-Rényi random graph and **configuration model.**

▷ Dynamic models:
Graph has evolving number of elements:

Preferential attachment model

Due to highly dynamic nature of citation networks, focus on dynamic models.

Preferential attachment

(Barabási-Albert (1999): 31402 citations @24-06-2018)

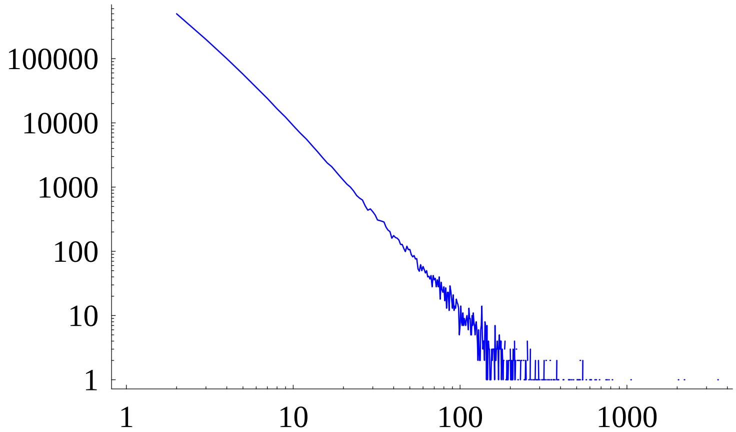
At time n, single vertex is added with m edges emanating from it. Probability that edge connects to ith vertex is proportional to

$$D_i(n-1) + \delta,$$

where $D_i(n)$ is degree vertex i at time n, $\delta > -m$ is parameter.

Yields power-law degree sequence with exponent $\tau = 3 + \delta/m > 2$.

Rich get richer!



$m = 2, \delta = 0, \tau = 3, n = 10^6$

Preferential attachment

▷ Preferential attachment models (PAMs) grow **linearly** in time:

Embed in continuous time,
where growth becomes exponential.

▷ Idea fruitful also for regular PAMs: [Rudas, Tóth and Valko (2007), K. Athreya et al. (2007)]
powerful tools of continuous-time branching processes:

Jagers-Nerman (84), Jagers (75).

▷ **Old-get-richer** phenomenon leading in PAMs:

Introduce random fitness for each vertex in graph.

▷ Vertices keep on receiving citations in PAMs:

Introduce aging effect.

Model

▷ Preferential attachment model in continuous time where rate of growth at time t of links to vertex v that is born at time s is

$$\eta_v(D_v(t)+\delta)g(t-s),$$

where

▷ η_v is fitness of vertex v:

Citation counts become highly variable;

▷ g is (integrable) aging function:

Vertices receive finite number of citations in lifetime;

▷ $D_v(t)$ is degree of vertex v at time t:

Increments of citation counts roughly linear;

▷ δ is parameter allowing for fine tuning.

Result

Focus here on tree setting in continuous time.

Denote

$$N(t) = \{\text{number of individuals in system}\},$$

$$N_k(t) = \{\text{number of individuals in system having } k \text{ children}\}.$$

Theorem 1. (Garavaglia-vdH-Woeginger 17)
There exist $\alpha > 0$ and random variable $\Theta > 0$, such that

$$\mathrm{e}^{-\alpha t} N(t) \xrightarrow{a.s.} \Theta,$$

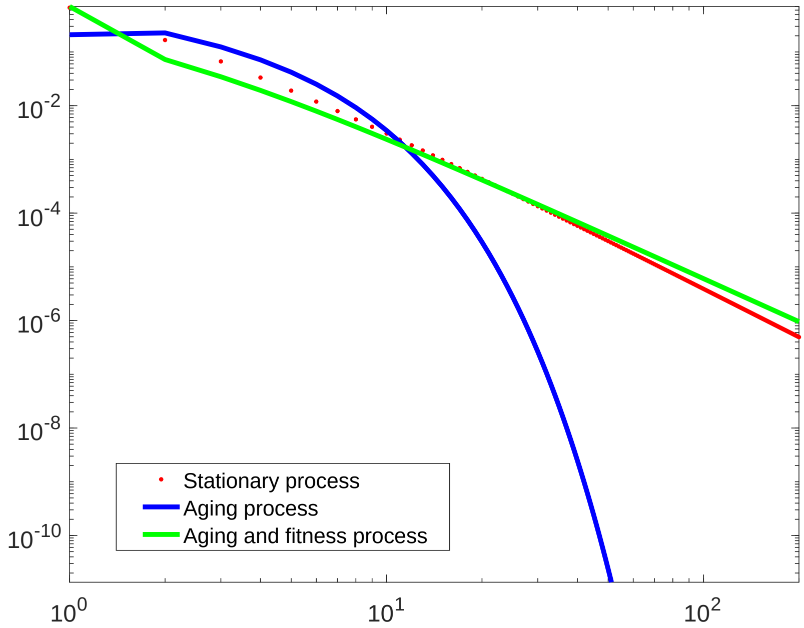
and probability mass function $(p_k)_{k \geq 0}$ such that

$$\frac{N_k(t)}{N(t)} \xrightarrow{\mathbb{P}} p_k.$$

▷ Power-law behavior characterized by asymptotics p_k for k large.

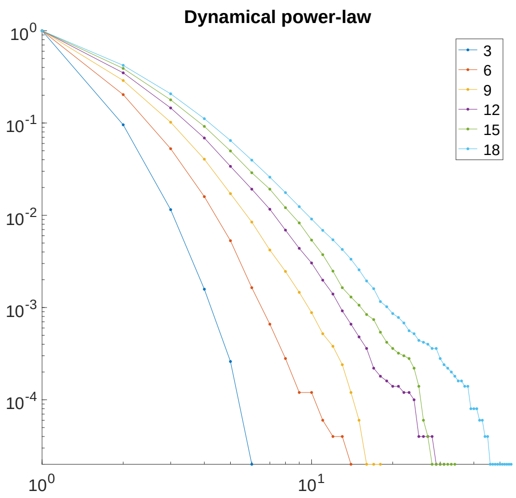
Picture-based discussion!

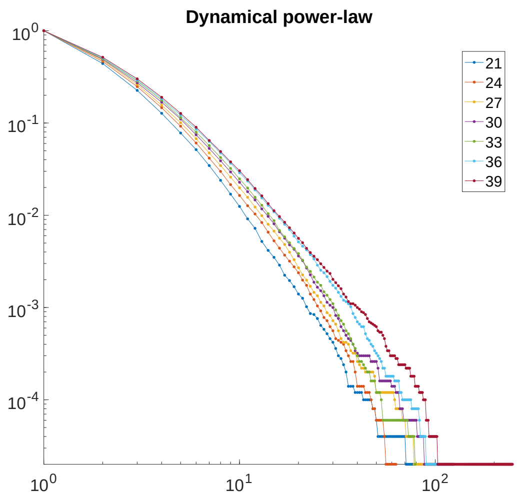
Degree distribution with/without fitness/aging



Examples of degree distributions with/without aging and with/without exponential fitness

Dynamic power law





Simulation of dynamic power law for exponential fitness

Conclusion

Reasonable **qualitative** comparison model/data:

▷ Exponential growth;

▷ Integrable aging;

▷ Highly-variable fitnesses.

Rigorous results in tree case:

▷ Exponential growth is **typical behavior;**

▷ Power laws with aging and fitness **only** when fitness has **at most** exponential tail.

Importance measures citation networks

Above model provides insight into

mechanisms driving citation networks.

Does not yet provide insight into how to

measure quality/popularity paper beyond citation counts.

Effective measure on the World-Wide Web is

PageRank

Is stationary distribution of random walk with random restarts:

bored surfer

Invented by Brin-Page 1998 to
bring order to the Web.

PageRank

Below, we assume that

$G = (V, E)$ is directed graph.

Fix damping factor c (or restart probability $1 - c$,) and let P denote random walk transition probability. Then, PageRank $\pi^{(G)}$ satisfies

$$\pi^{(G)} = c\pi^{(G)}P + \frac{1-c}{n}\mathbf{1}.$$

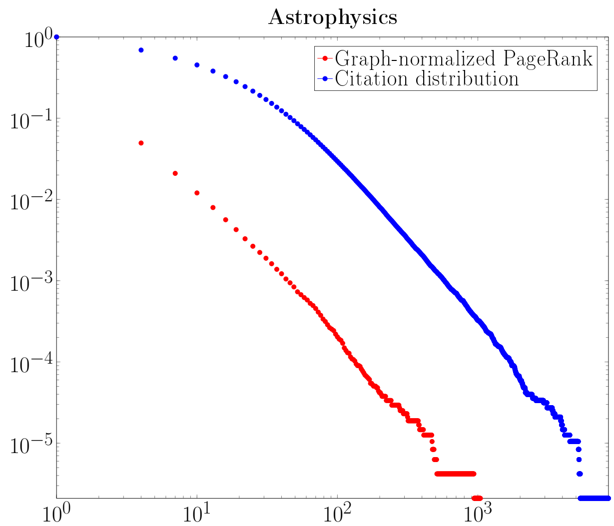
Often more convenient to deal with graph-normalized PageRank, which is just $R^{(G)} = n\pi^{(G)}$, and which satisfies

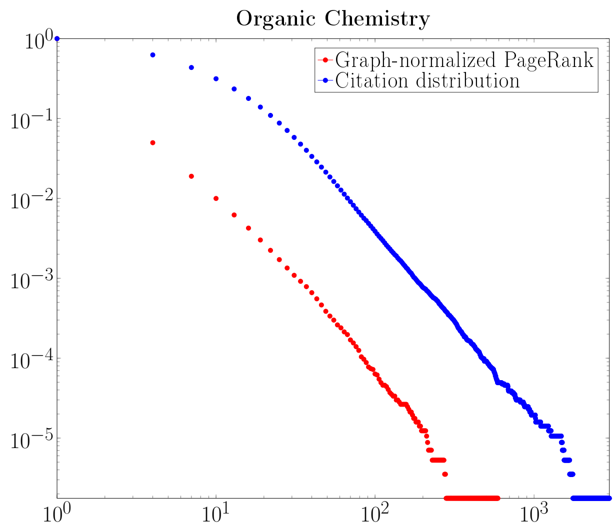
$$R^{(G)}(v) = c\sum_{u \to v} \frac{1}{d_v^{(\mathrm{out})}} R^{(G)}(u) + 1 - c.$$

Then, denoting V_n vertex chosen uniformly at random from $[n]$,

$$\mathbb{E}[R^{(G)}(V_n)] = 1.$$

Power-law hypothesis





PageRank and In-degree in citation networks from Web of Science in Astrophysics and Organic chemistry

In-degree and PageRank have same power-law exponents

Power-law hypothesis

Power-law hypothesis

Much previous work on power-law hypothesis:

▷ In-degree and PageRank: why do they follow similar power laws?
Litvak-Scheinhardt-Volkovich Internet Math (2007)

▷ Generalized PageRank on directed configuration networks
Chen-Litvak-Olvera-Cravioto RSA (2017)

▷ PageRank on inhomogeneous random digraphs
Lee-Olvera-Cravioto (2017)

Generally prove weak convergence and power-law hypothesis at once, relying on stochastic fixed-point equations.

Even more work on algorithmic extensions of PageRank!

Local weak convergence

▷ Key technique in analyzing sparse graphs is

local weak convergence.

Makes statement that local neighborhoods in CM are like BP exact. See Section II.1.4 for intro LWC and Section II.3.2 for LWC CM.†

▷ Applies generally, to general IRGs in Section II.2.2, and PAM Berger-Borgs-Chayes-Saberi (14) and Section II.4.2.

▷ LWC holds when

$$\frac{1}{n}\sum_{i\in[n]} \mathbb{1}_{\{B_r(i)\simeq(H,y)\}} \to \mathbb{P}(B_r(\varnothing)\simeq(H,y)),$$

for any rooted graph (H,y), where $B_r(i)$ is r-neighborhood of $i\in[n]$ and $B_r(\varnothing)$ is r-neighborhood of $\varnothing$ in limiting rooted random graph.

▷ Convergence of means is LWC in distribution, convergence in probability is LWC in probability.

Result

Long-term aim is to prove power-law hypothesis in great generality.

Start by giving general condition for weak convergence PageRank:

Theorem 2. (Garavaglia-vdH-Litvak 18)
Assume that the graph $G_n = (V_n, E_n)$ converges in local weak convergence sense.
(a) If LW convergence is in distribution, then there exists a limiting random variable such that

$$R_{V_n}^{(G_n)} \xrightarrow{d} R^{(\infty)}.$$

(b) If LW convergence is in probability, then there exists a limiting random variable such that, for every $x > 0$,

$$\frac{1}{n}\sum_{v\in[n]} \mathbb{1}_{\{R_v^{(G_n)}>x\}} \xrightarrow{\mathbb{P}} \mathbb{1}_{\{R^{(\infty)}>x\}}.$$