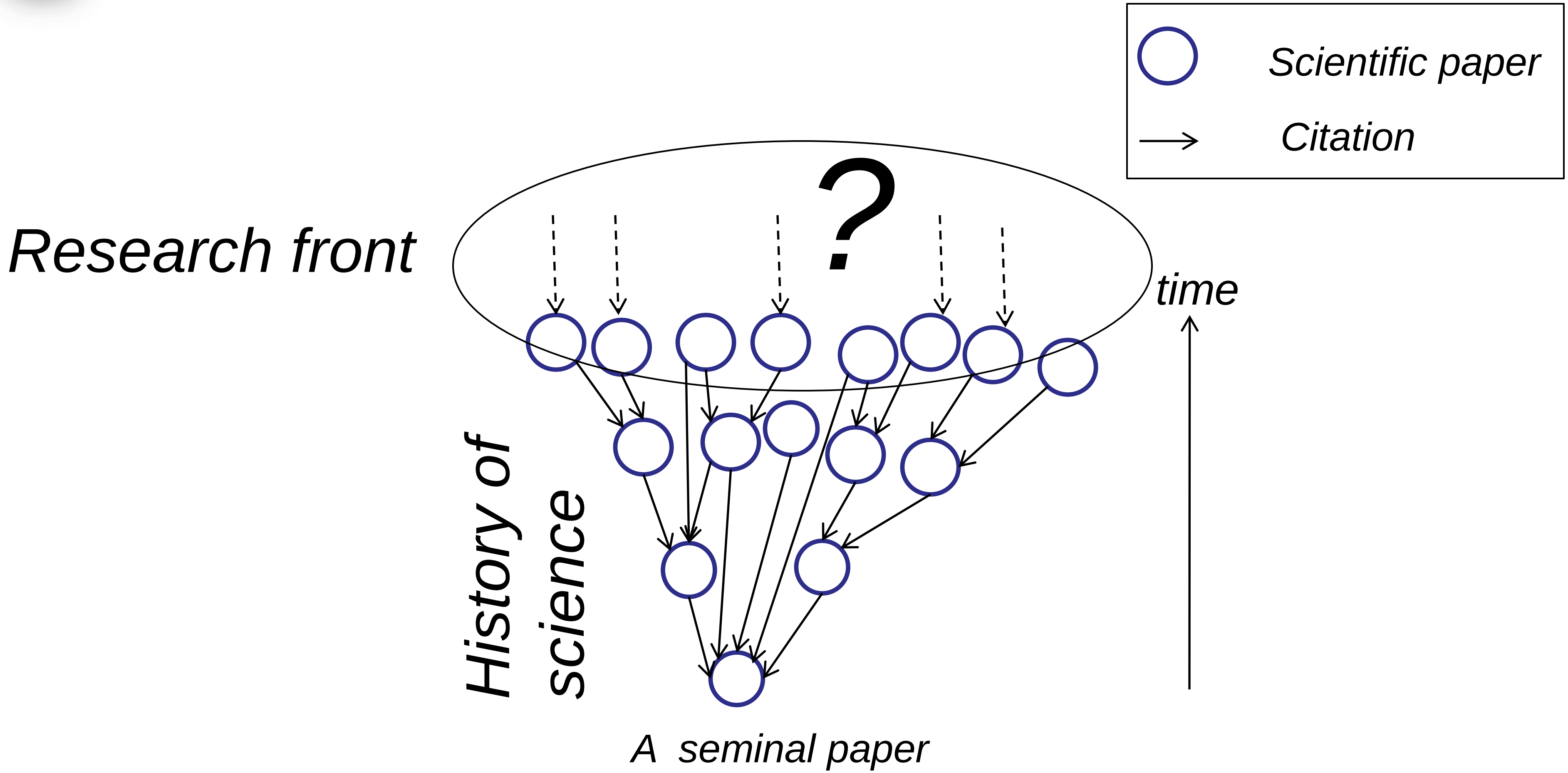
Motivation for studying citation dynamics



Map of science shall be complemented by dynamics.
Both are obtained from citations.
Our goal is to establish a reliable model of citation dynamics, preferably non-empirical one

Citation Dynamics and Bass model of diffusion of innovations

Michael Golosovsky & **Sorin Solomon**

(experimental physics) *(physics of complexity)*

Racah Institute of Physics,

Hebrew University of Jerusalem, Israel

- *The model of citation dynamics of scientific papers*
- *Measurements to verify the model*
- *Discussion*

Our previous publications
M. Golosovsky and S. Solomon,
- *J. Stat. Phys. Special Issue Statistical Mechanics in Social Sciences 151 340 (2013);*
- *Phys. Rev. Lett. 109 098701 (2012)*
- *Eur. Phys. J. Special Topics 205 303 (2012)*

Models of Citation Dynamics

- *From the author's perspective*
- From the paper's perspective

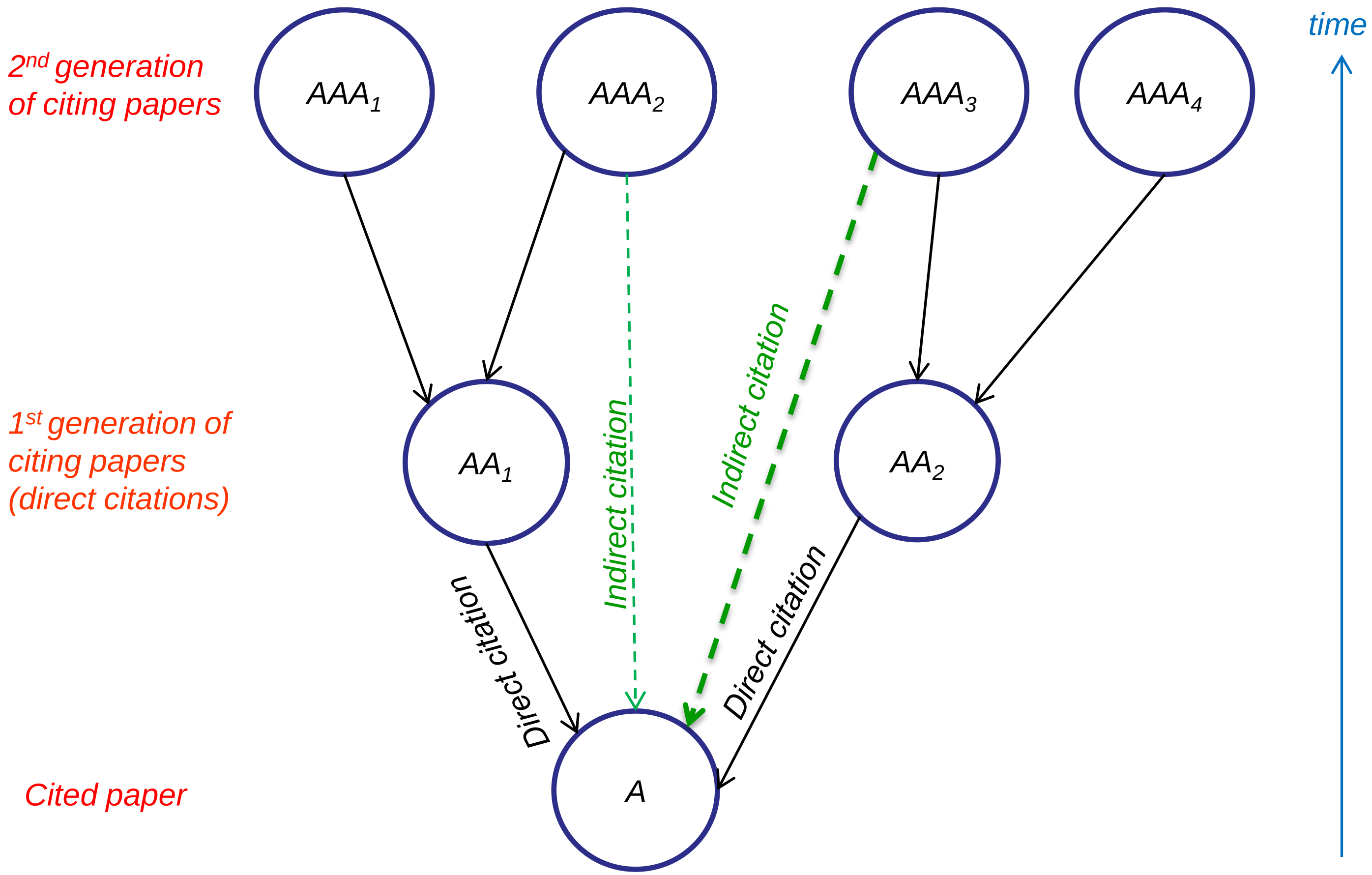
Main idea: there are **two** kinds of citations

Generic models assuming two kinds of citations:

•*Recursive search – Vasquez (2001)*
•*Redirection - Krapivsky&Redner (2005)*
•*Link copying- Simkin&Roychowdhury (2007)*
•*Direct and indirect citations – Petersen, Presse&Dill (2009)*

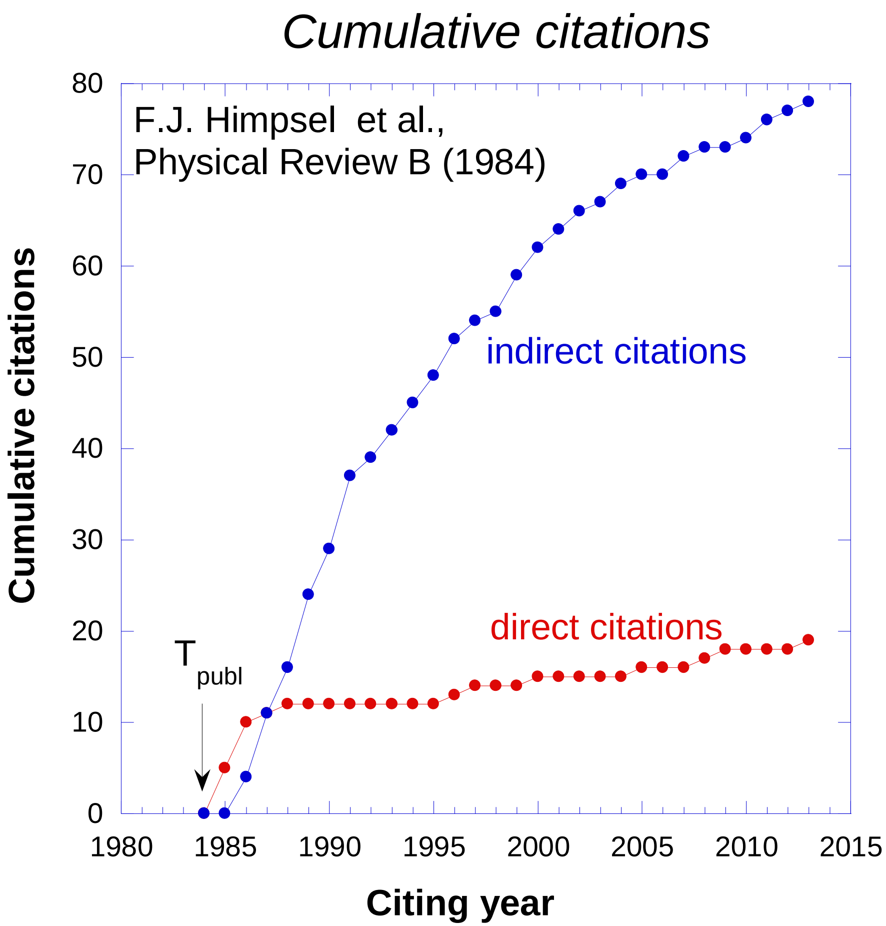
The overview of the models of citation dynamics-
Vitanov &Ausloos in
"Models of Science Dynamics" ,A. Scharnhorst, K. Borner&P. Besselaar

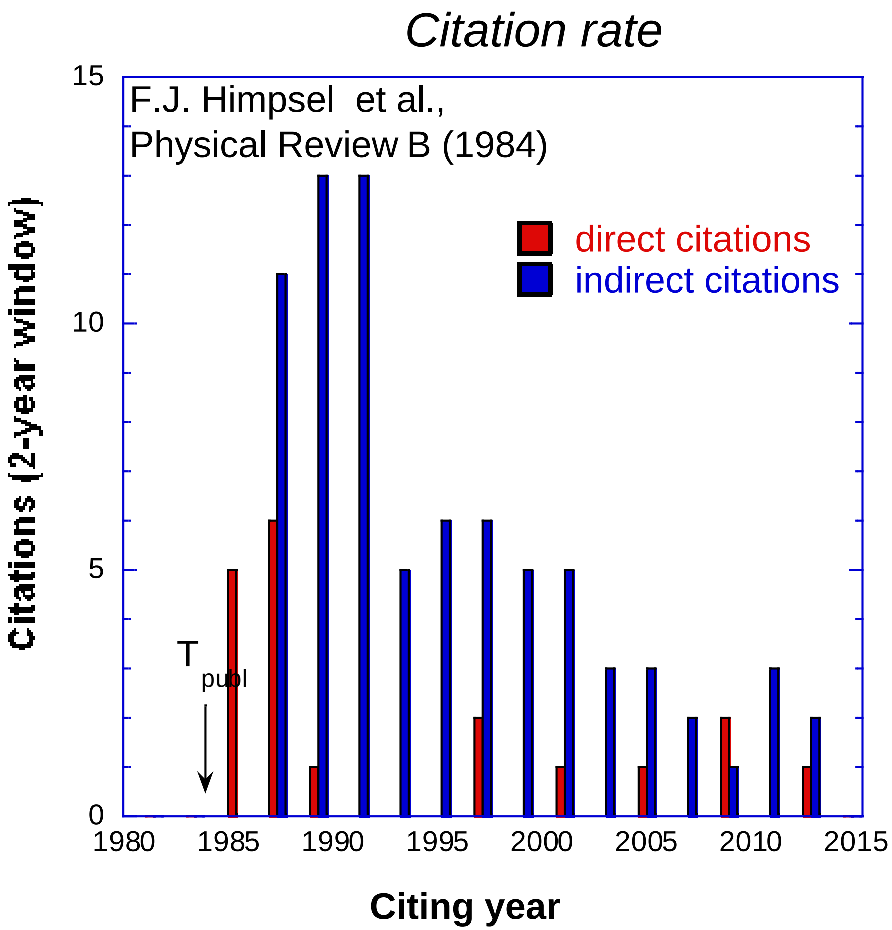
Direct and indirect citations



Direct citation: The citing author reads the paper and cites it
Indirect citation: The citing author reached to this paper through the reference list of another paper that he cites anyway

An example





Direct citations appear immediately after publication and saturate after ~3-5 years.
Indirect citations start to grow after delay of ~1-2 years and come to saturation slowly (*if at all*)

The model and its context

•Population dynamics:

age-dependent birth-death model with immigration

•Epidemiology

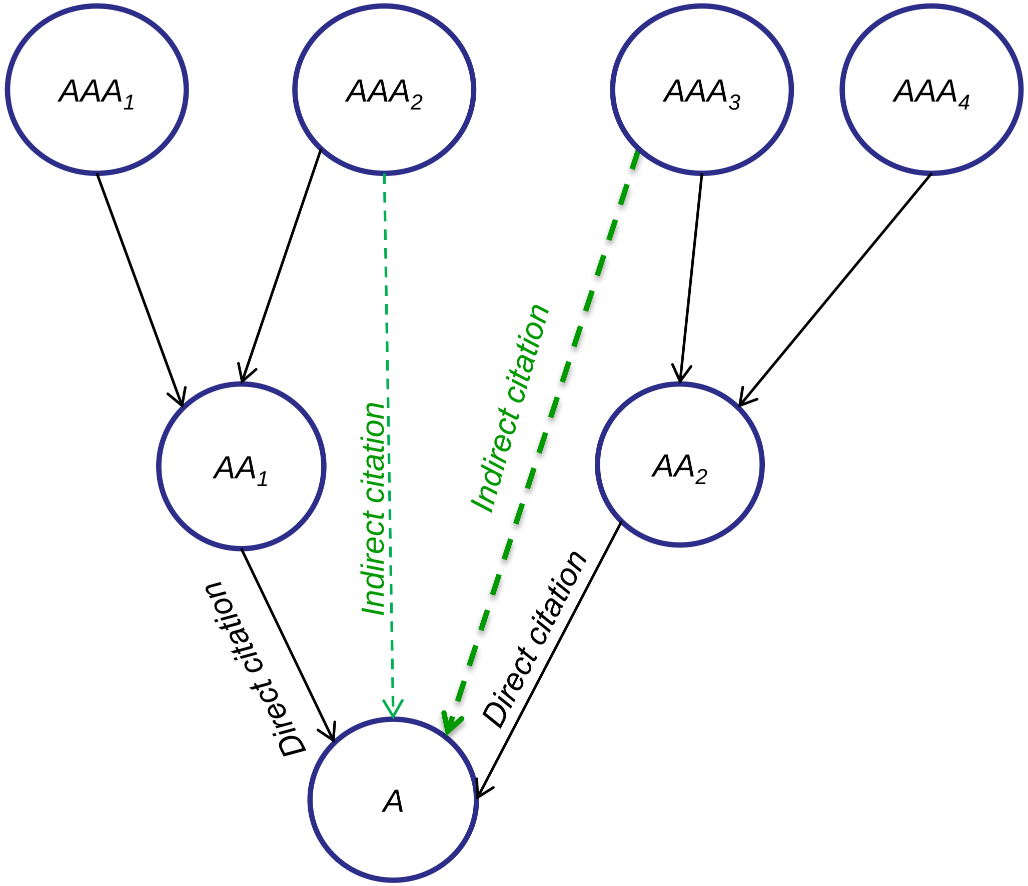
SI (susceptible-infected) model

•Rumor propagation

•Diffusion of innovations

•Marketing:

product penetration into the market



Bass model of diffusion of innovations

F. Bass (1969)

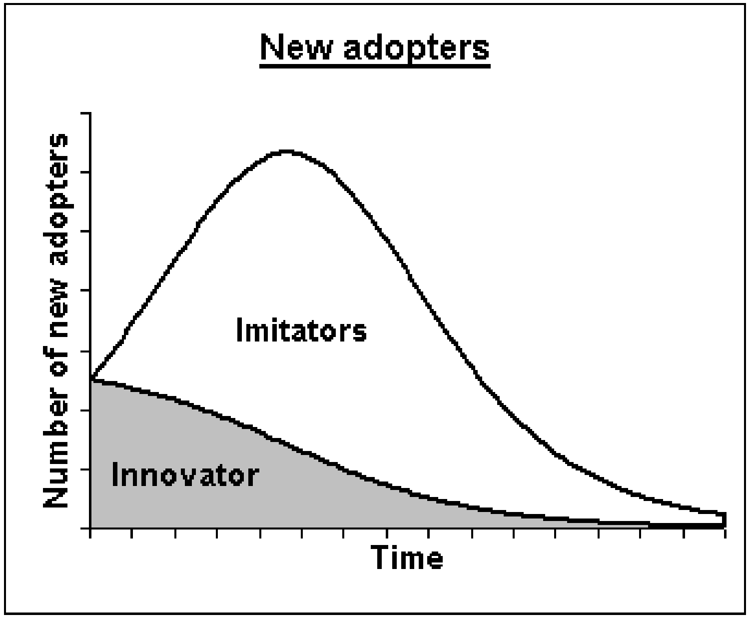
$$\frac{dN}{dt} = \left(p + \frac{q}{\overline{N}} N \right)\left(1 - \frac{N}{\overline{N}} \right)$$

N - total number of adopters

$\overline{N}$ - number of potential adopters (market size)

p – coefficient of innovation

q – coefficient of imitation



Innovators are reached by broadcasting
Imitators are reached by word-of-mouth

Bass model of diffusion of innovations

Innovations

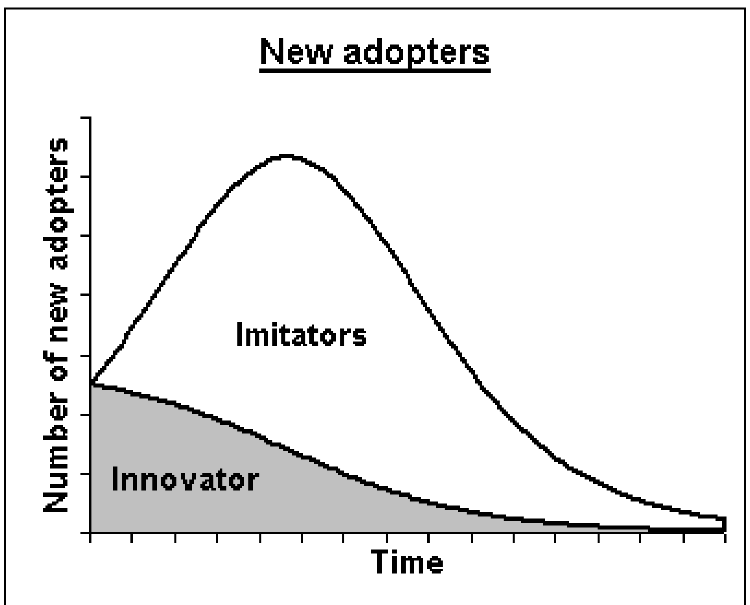
$$\frac{dN}{dt} = \left(p + \frac{q}{\overline{N}}N\right)\left(1 - \frac{N}{\overline{N}}\right)$$

N - total number of adopters

$\overline{N}$ - number of potential adopters (market size)

p – coefficient of innovation

q – coefficient of imitation

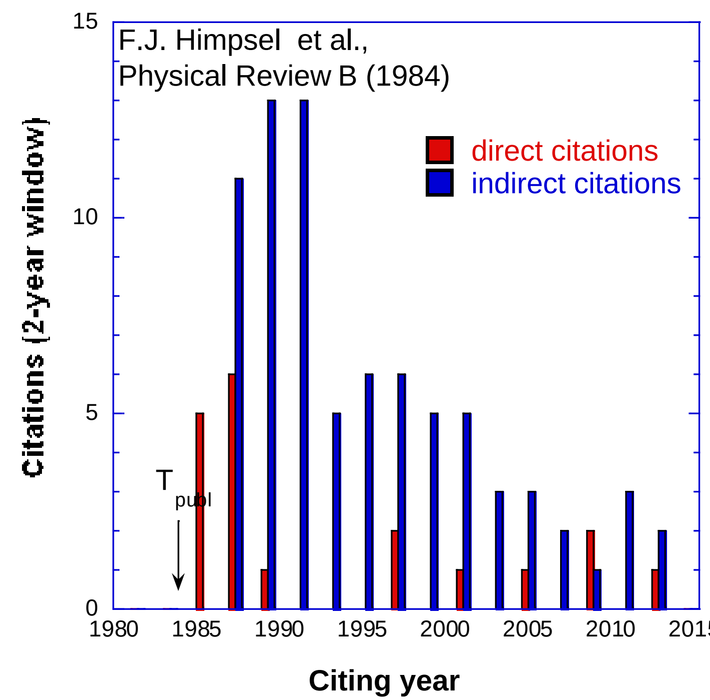


Citations

$$\frac{dk}{dt} = p(t) + \int_0^t q(t-\tau)\frac{dk}{d\tau}d\tau$$

Adopters = innovators & imitators

Citations = direct & indirect



Epidemic model: an application to the transmission of ideas *Goffman-Newill (1964)*

Epidemiology- SI model

$$\frac{dS}{dt} = \mu - \beta IS$$
$$\frac{dI}{dt} = \nu + \beta IS$$

S – susceptible

I – infected

$$\text{if } S = const \text{ then } \frac{dI}{dt} = \nu + \beta SI$$

Citations

$$\frac{dk}{dt} = p(t) + \int_0^t q(k, t-\tau)\frac{dk}{d\tau}d\tau$$

Total citations — *Direct citations* — *Indirect citations*

$$\text{if } q = const \text{ then } \frac{dk}{dt} = p(t) + qk$$

Total citations- infected
Direct citations- v
Indirect citation – βIS

The model and the measurements

Self-exciting (non-Markovian) stochastic process

λ – latent citation rate:

$$\lambda = p(t) + \int_0^t q(t-\tau)\frac{dk}{d\tau}d\tau$$ ← memory

actual citations in the time window Δt follow a

Poisson process where $\Delta k = \lambda \Delta t$ and $P(\Delta k) = \frac{\lambda^{\Delta k}}{(\Delta k)!}e^{-\lambda}$

To apprehend the model we used ISI Web of Science and considered Physics papers published in 1984.

1st step: analysis of a small dataset of 36 papers

- *We measured direct and indirect citations separately*
- *Found functions p(t) and q(t-τ)*

2nd step: verification of the model using large data set of 40,195 papers

- *Verified parameters p(t) and q(t-τ)*
- *Verified that the noise is indeed Poissonian*

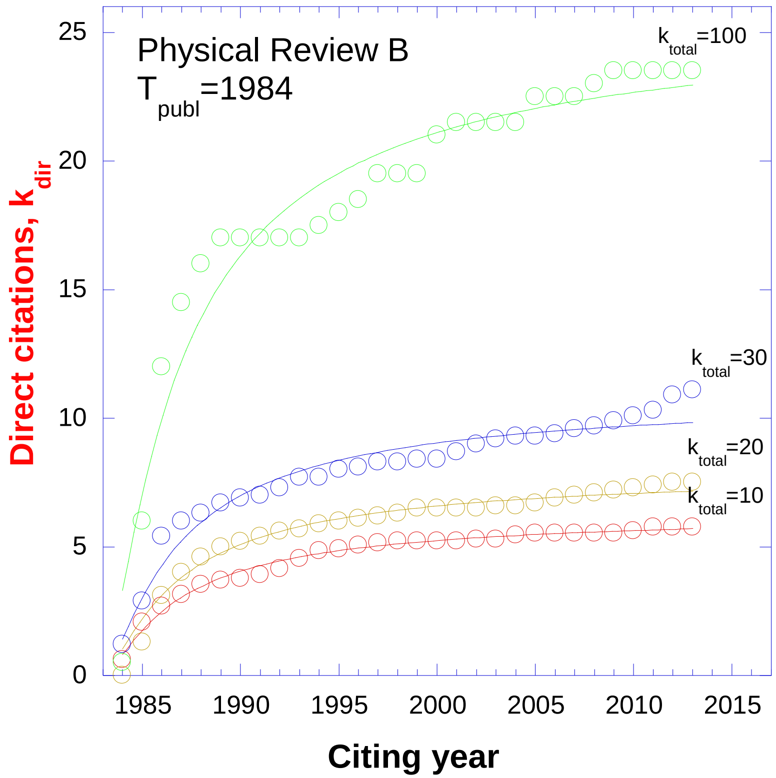
Comparison to the model

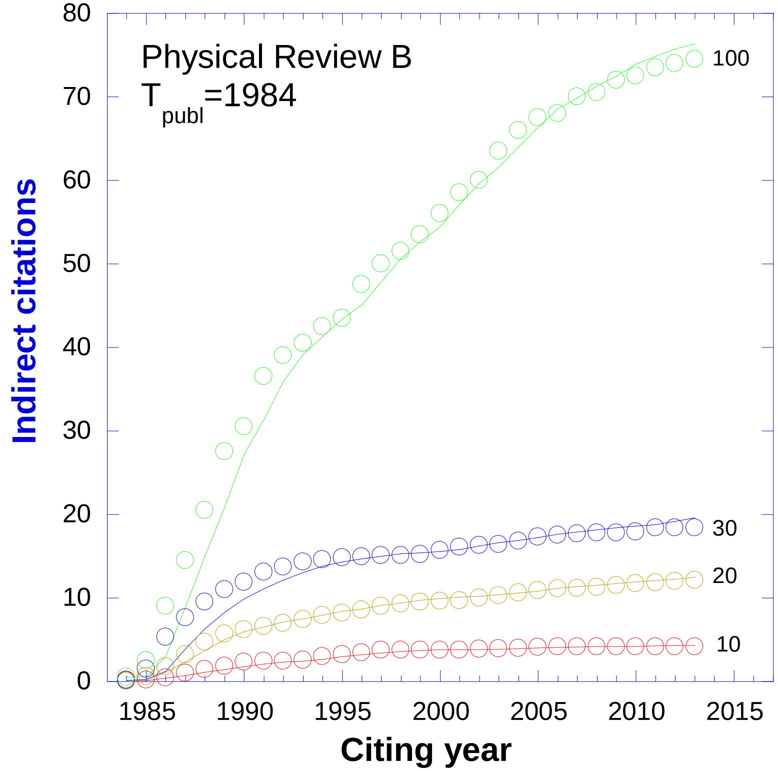
$$\frac{dk}{dt} = p(t) + \int_0^t q(t-\tau)\frac{dk}{d\tau}d\tau$$

$$p = p_0 e^{-\frac{3.4}{t+0.7}}$$

$$q = q_0 e^{-0.8(t-\tau)}$$

small dataset of 36 papers





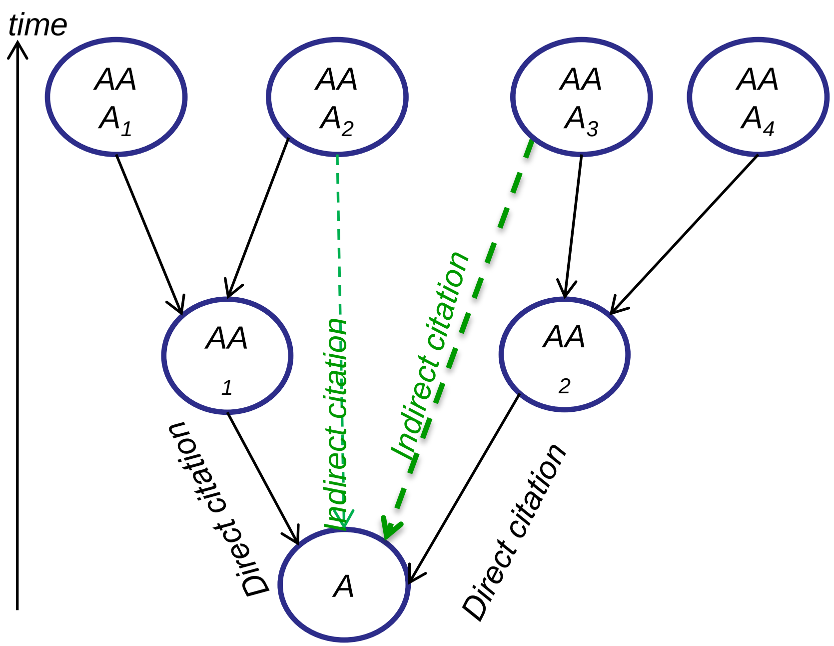
Analysis of results: nonlinearity

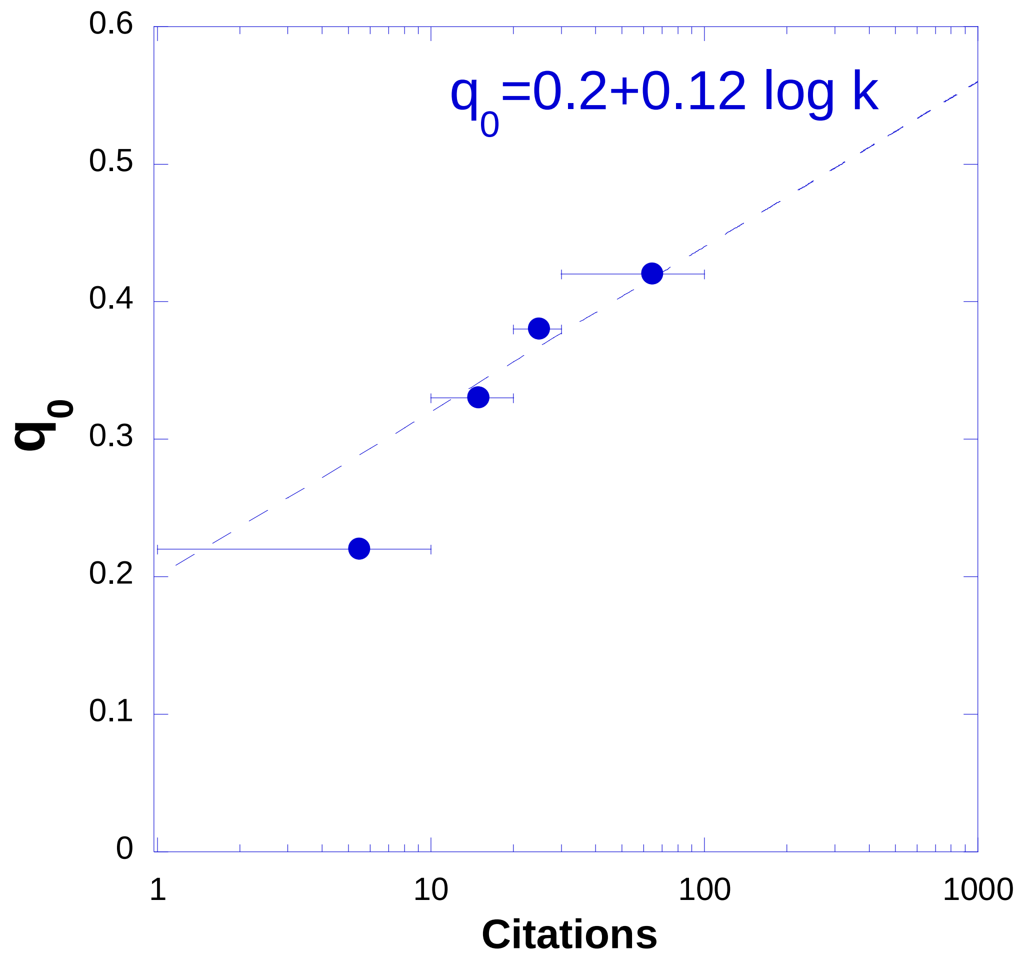
$$\frac{dk}{dt} = p(t) + \int_0^t q(k, t-\tau)\frac{dk}{d\tau}d\tau$$

$$p = p_0 e^{-\frac{3.4}{t+0.7}} \quad ; p_0 \text{ - initial conditions}$$

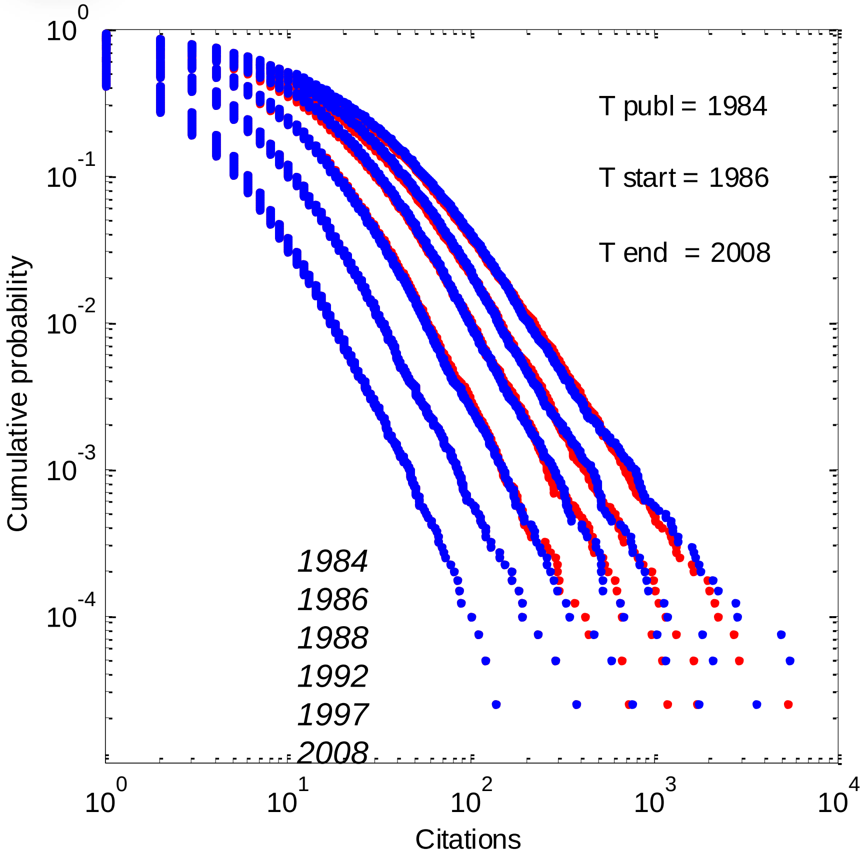
$$q = q_0 e^{-0.8(t-\tau)} \; ; q_0 \text{ depends on } k?!$$

The memory is short- only 1.5-2 years



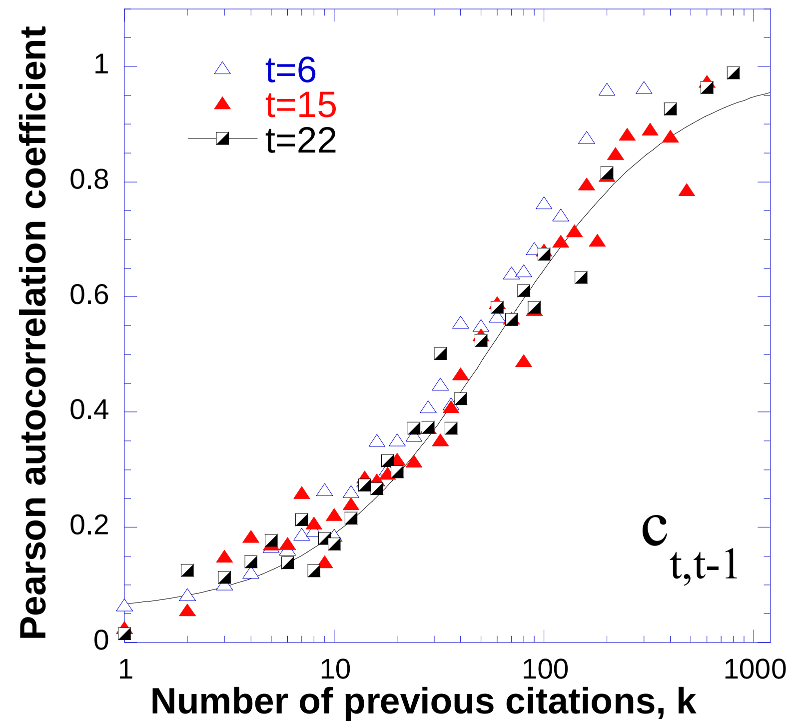


Verification of the model



Cumulative citation distributions to 40,195 Physics papers published in 1984. The model takes the data for 1984, 1985, and 1986 as initial conditions and calculates citations further on.

Blue circles- model, red circles- data



Pearson **autocorrelation** coefficient. It captures memory effect.

large dataset of 40,195 papers

Analysis of the model

$$P(\Delta k) = \frac{\lambda^{\Delta k}}{(\Delta k)!} e^{-\lambda}$$

Latent citation rate

$$\lambda = p(t) + \int_0^t q(k, t-\tau) \frac{dk}{d\tau} d\tau$$

positive feedback due to nonlinearity + self-excitation

The theoretical models completely missed the nonlinearity

"There are more things in heaven and earth, Horatio,
Than are dreamt of in your philosophy."
-William Shakespeare, *Hamlet*

Consequences of nonlinearity

- *Citation distributions are non-stationary*
- *Runaways*

For simplicity, consider impulse response

$$\frac{dk}{dt} = P\delta(t) + \int_0^t q_0 e^{-\beta(t-\tau)} \frac{dk}{d\tau} d\tau$$

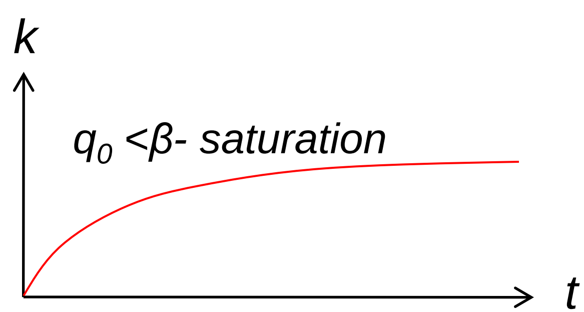
solution: $\frac{dk}{dt} = P q_0 e^{-(\beta - q_0)t}$

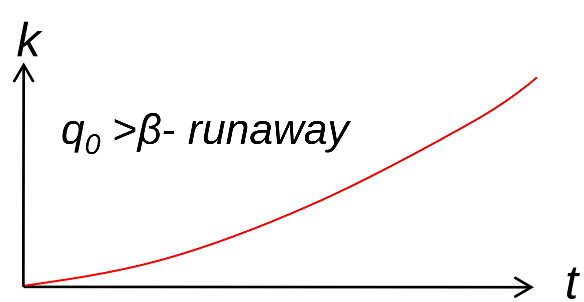
$$k = P\left[1 + \frac{q_0}{\beta - q_0}\left(1 - e^{-(\beta - q_0)t}\right)\right]$$

$q_0 = \beta$ - branching (tipping) point

since q_0 increases with k, highly-cited papers can pass tipping point and become runaways

Branching process
Citation cascades
q_0/β – reproduction number



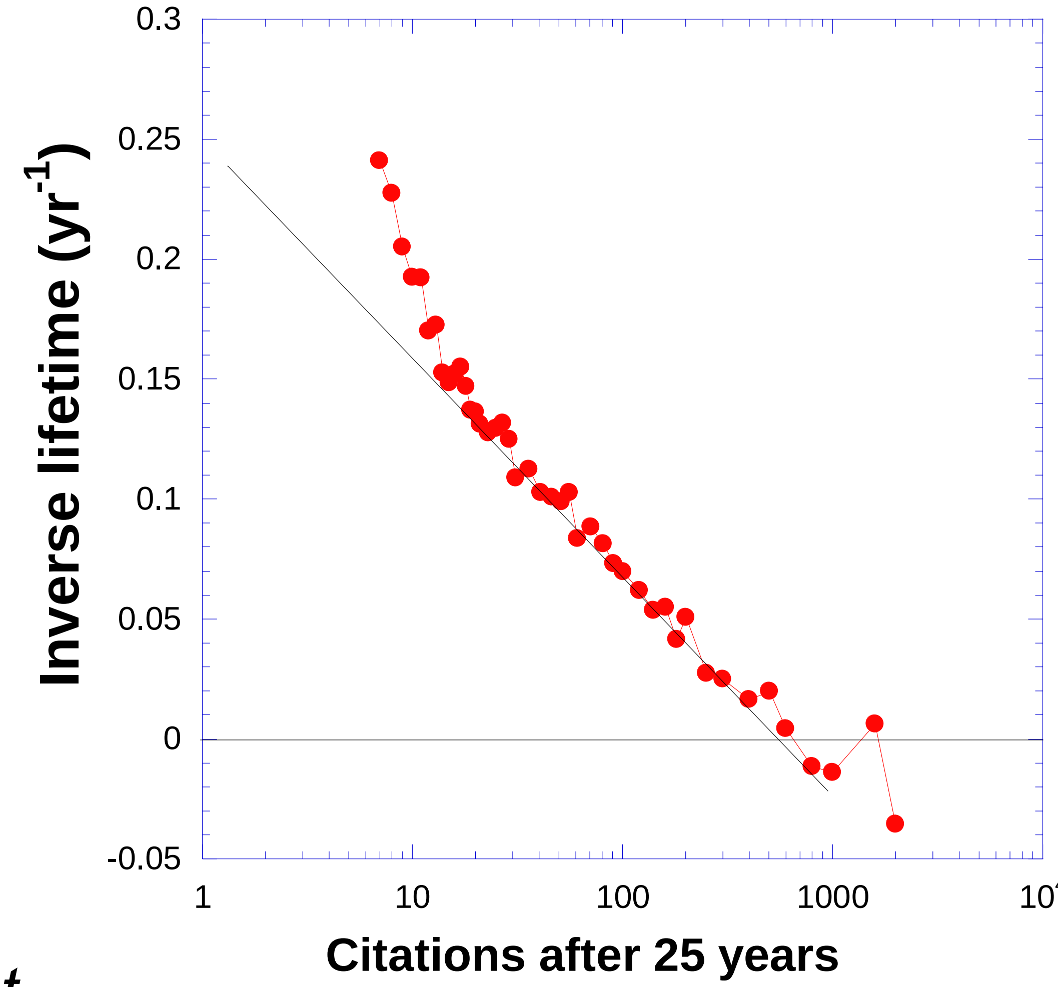


Runaways

$$\frac{dk}{dt} = Pq_0 e^{-(\beta - q_0)t}$$

$$(\beta - q_0) - \text{inverse lifetime}$$

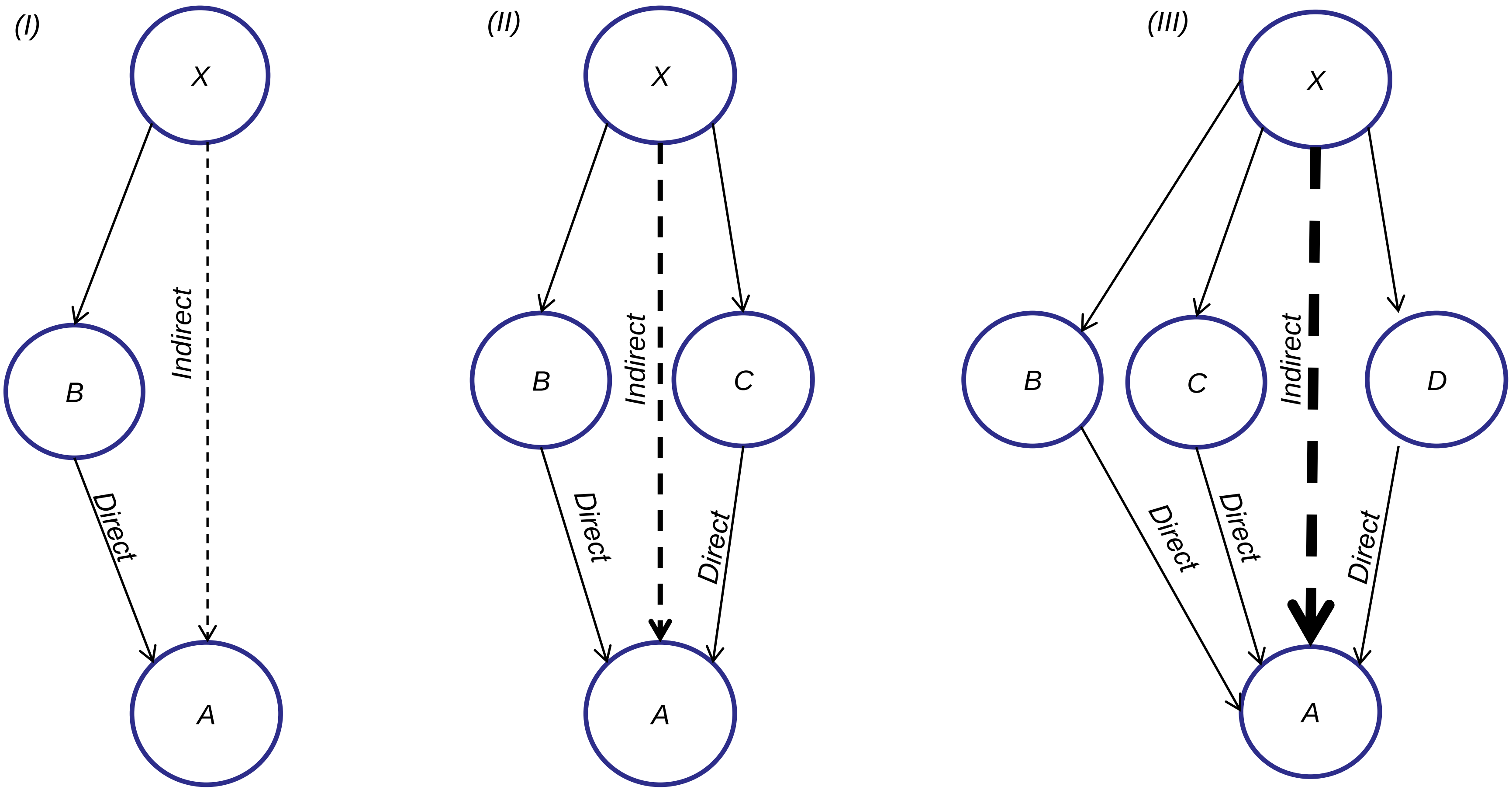
$$q_0 = 0.2 + 0.12 \log k$$



Not only "rich get richer" but

Rich live longer!

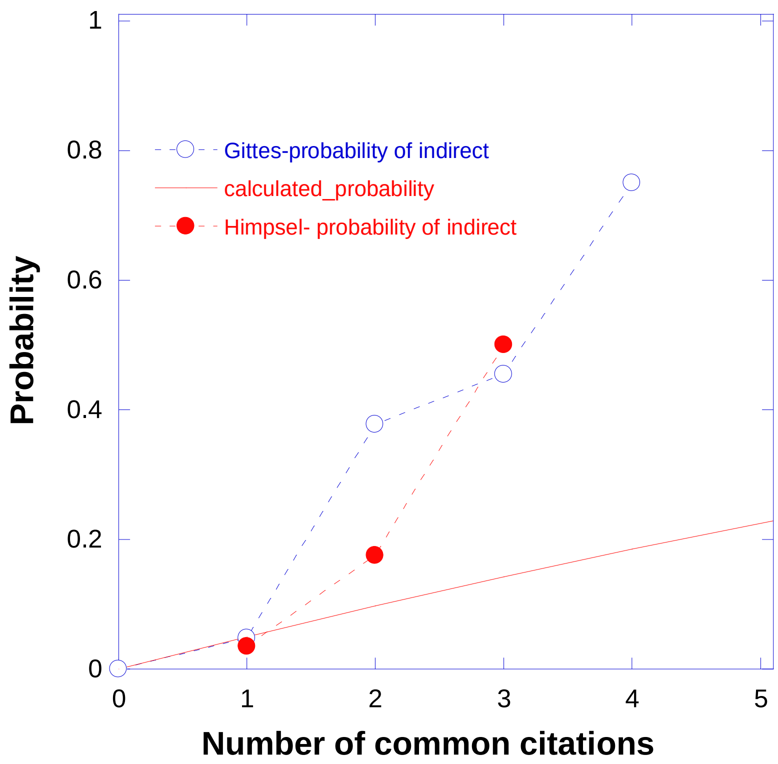
What is the source of nonlinearity? Look at network motifs

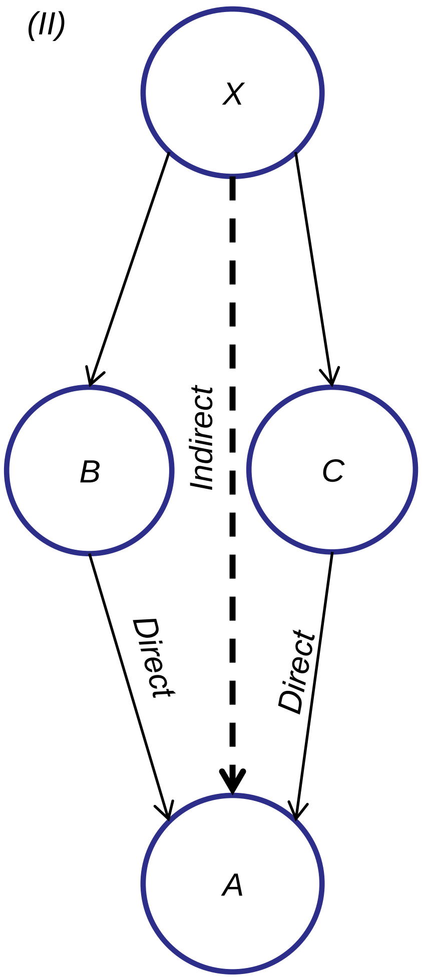


$3q^{III}_{indirect} > 2q^{II}_{indirect} > q^{I}_{indirect}$

Homophily? Peer-to-peer influence?

Multipath interference

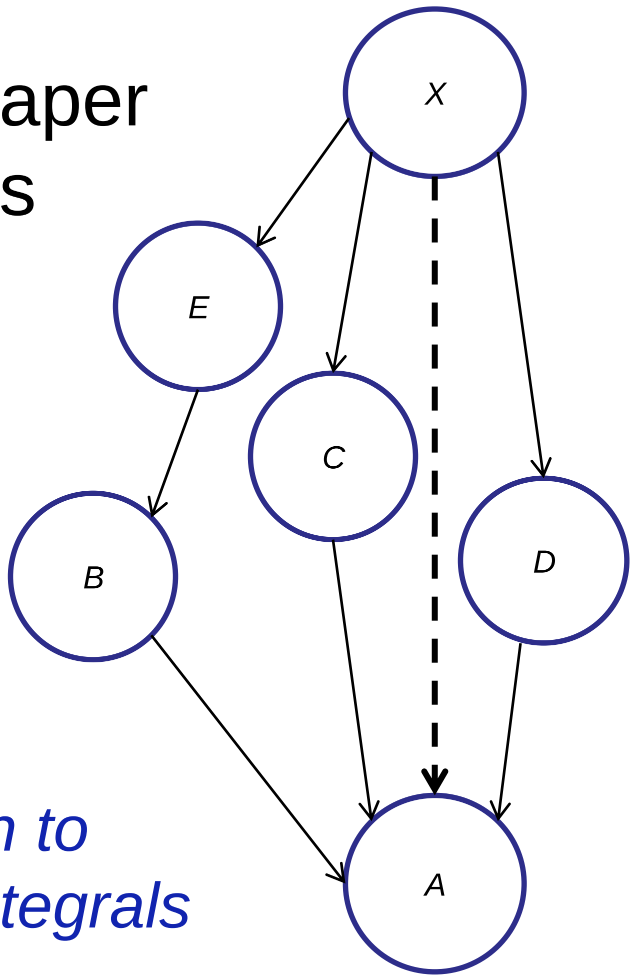




Probability of indirect citations in dependence on the number of common direct citations. Continuous line shows linear approximation (no multipath interference).

The probability that paper X cites paper A is the weighted sum over all paths leading from X to A

Pathways are important!



Physicists can appreciate possible relation to multipath interference or Feynman path integrals

Conclusions

- Microscopic model of citation dynamics based on nonlinear self-exciting process
- Verification of the model

- Deep analogy between citation dynamics and diffusion of innovations/products:

  Citation dynamics of a paper captures penetration to market of ideas: broadcasting (direct) and word-of-mouth (indirect) mechanisms are clearly seen

Future plans and applications

- Forecasting
  prediction of the future citation history of a given set of papers
  - *impact factor of a journal*
  - *h-index of a researcher*

- Research frontiers
  Early identification of important papers

Relation to previous talks

Memory, non-Markovian processes- Martin Rosvall, Marton Karsai, Renaud Lambiotte, Ingo Scholtes

Economic analogies, Knowledge diffusion- Bruno Goncalvez

Runaways (in the context of institutions) –Nikolay Vitanov

h-index- and how to affect/predict it- Serge Galam, Raj Kumar Pan

Stochastic process, autocorrelation/autocovariance – Oleg Yordanov

Homophily- Sanja Scepanovic

Pathways- Renaud Lambiotte

Predictability- Ingo Scholtes