Quantifying Long-Term Scientific Impact

Scientific Impact → Citations

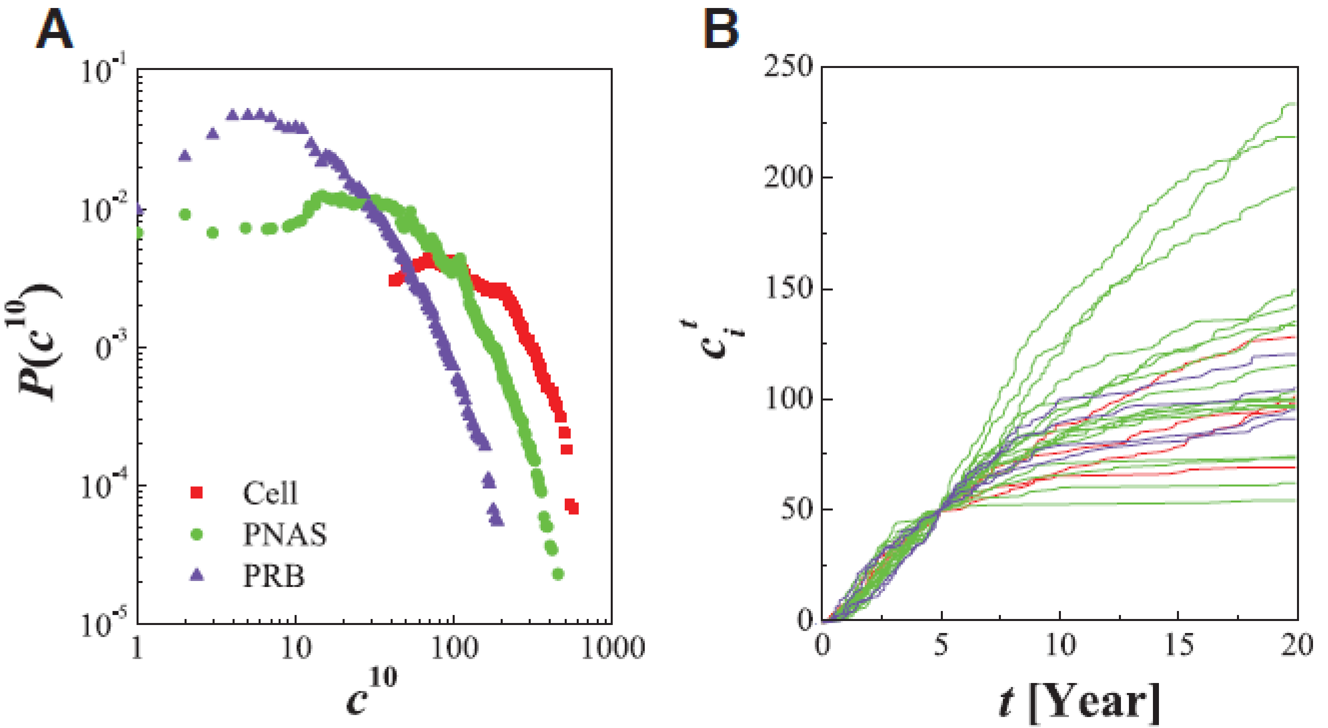
Citation-based measures:

- Impact factor (IF)
- The number of citations
- Hirsch index
- Others

Citation-based measures lack long-term predictability

Papers published in the same journal a decade later acquire widely different number of citations, from one to thousands

A group of papers that within a 5-year span collect the same number of citations are found to have widely different long-term impacts



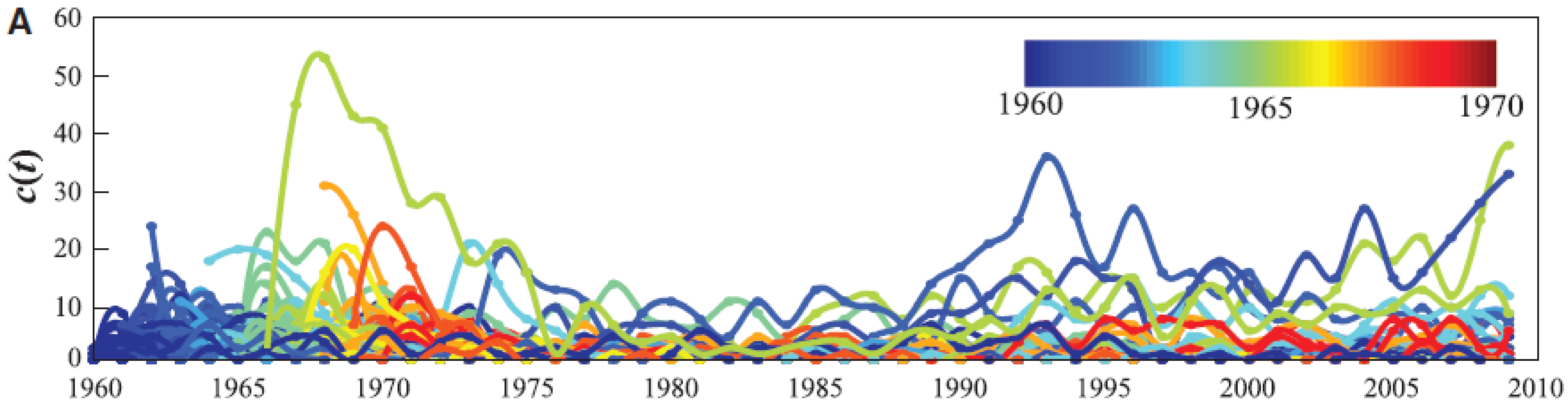
- Citation distributions
- Rescale discipline-dependent variables

Aggregated citation patterns are characterized by generic scaling laws

---

Little is known about the mechanisms governing the citation histories of individual papers

Randomly select 200 papers published between 1960 and 1970 in the (Physical Review) PR corpus.



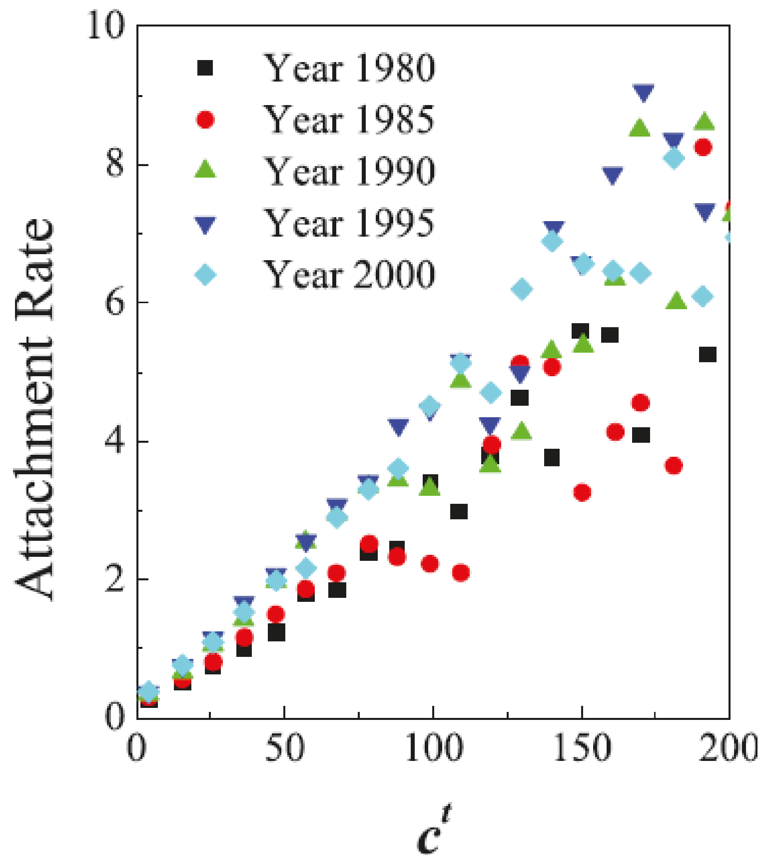
- Lack of order
- Hence lack of predictability

This lack of order in citation histories is only apparent, because citations follow widely reproducible dynamical patterns that span research fields.

Three fundamental mechanisms that drive the citation history of individual papers:

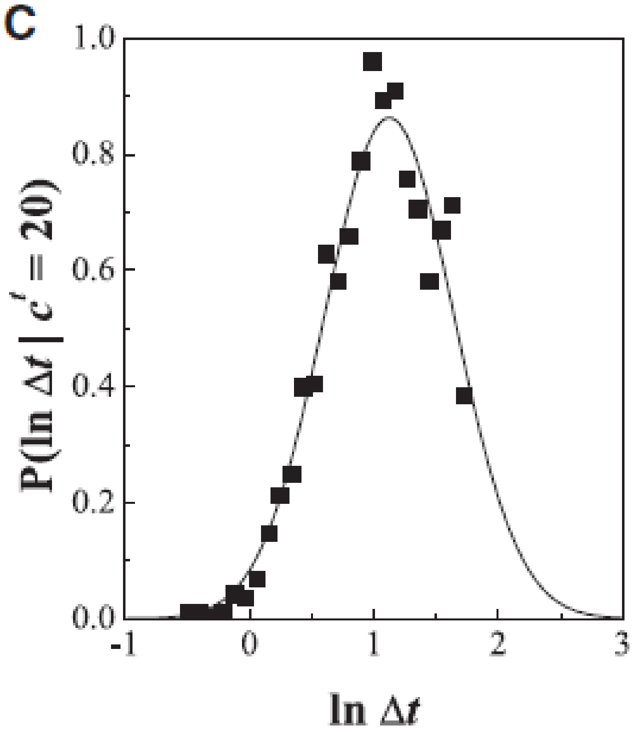
- Preferential attachment
- Aging
- Fitness

Preferential Attachment



c_i^t: the citations of each paper before this year

Aging



$$P(\Delta t) = \frac{1}{\sqrt{2\pi}\sigma\Delta t}\exp\left(-\frac{(\ln\Delta t - \mu)^2}{2\sigma^2}\right).$$

μ governes the time for a paper to reach its citation peak

σ captures the decay rate

$P_i(t)$: log-normal survival probability

Fitness

Inherent differences between papers
novelty and importance of a discovery

η_i :   the community's response to a work.

The probability that paper i is cited at time t after publication

$$\prod_i(t) \sim \eta_i c_i^t P_i(t)$$

$$\Delta t_i = t - t_i = \beta^{-1} \ln(N/i),$$

$$c_i^t = m\left(e^{\lambda_i \Phi\left(\frac{\ln t - \mu_i}{\sigma_i}\right)} - 1\right).$$

$$\Phi(x) \equiv (2\pi)^{-1/2} \int_{-\infty} e^{-y^2/2} dy. \qquad \lambda_i \equiv \eta_i \beta / A$$

- $\boldsymbol{m}$: the average number of references each new paper contains
- $\boldsymbol{\beta}$: the growth rate of the total number of publications
- $\boldsymbol{A}$: is a normalization constant

$$c_i^t = m\left(e^{\lambda_i \Phi\left(\frac{\ln t - \mu_i}{\sigma_i}\right)} - 1\right).$$

$t \rightarrow \infty \quad c_i^t \rightarrow c_i^\infty \quad \varphi \rightarrow 1$

$$c_i^\infty = m\left(e^{\lambda_i} - 1\right)$$

The ultimate impact of a paper is only determined by the relative fitness.

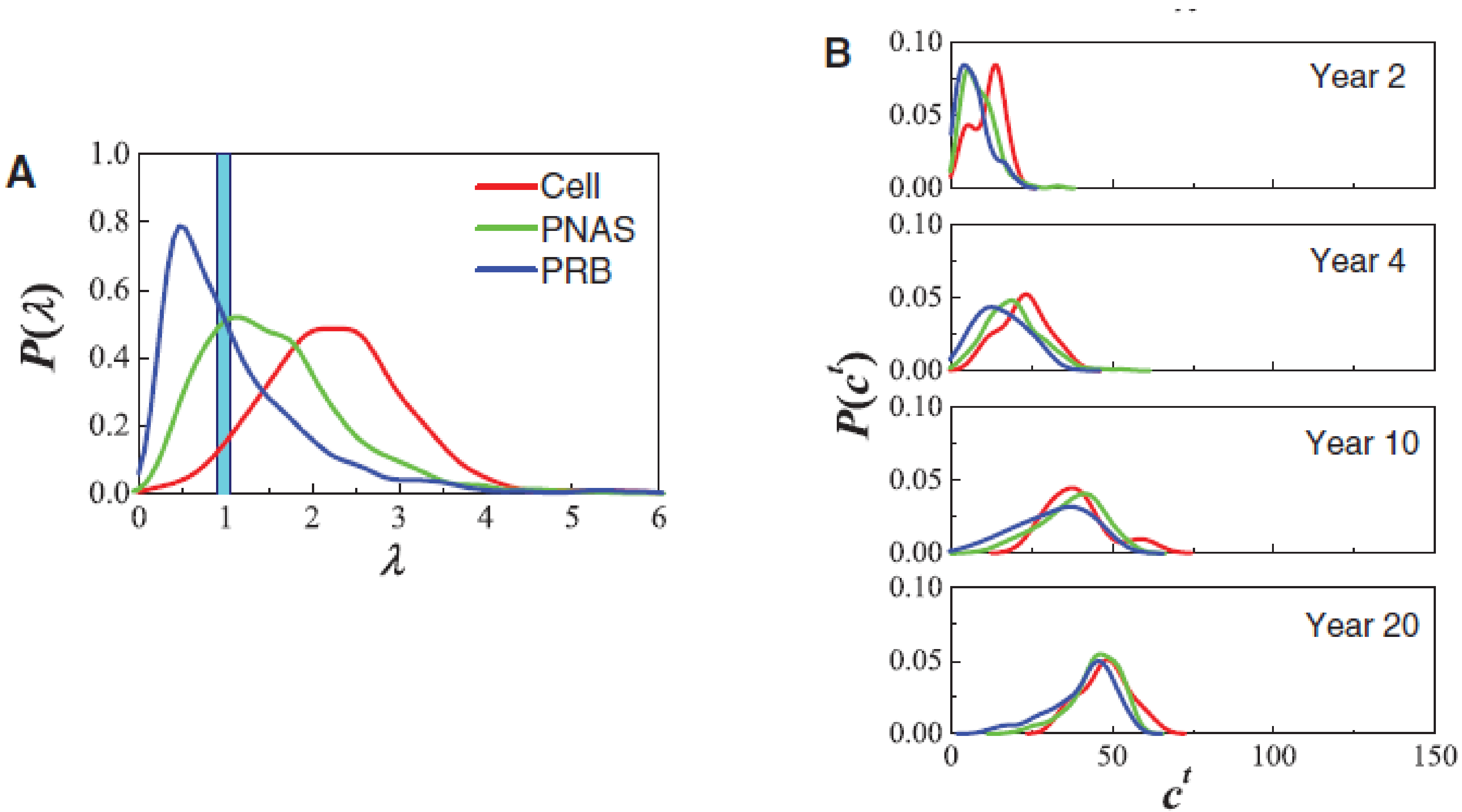
$$T_i^* \approx \exp(\mu_i).$$

The impact time is mainly determined by μ_i

The proposed model offers a journal-free methodology to evaluate long term impact.

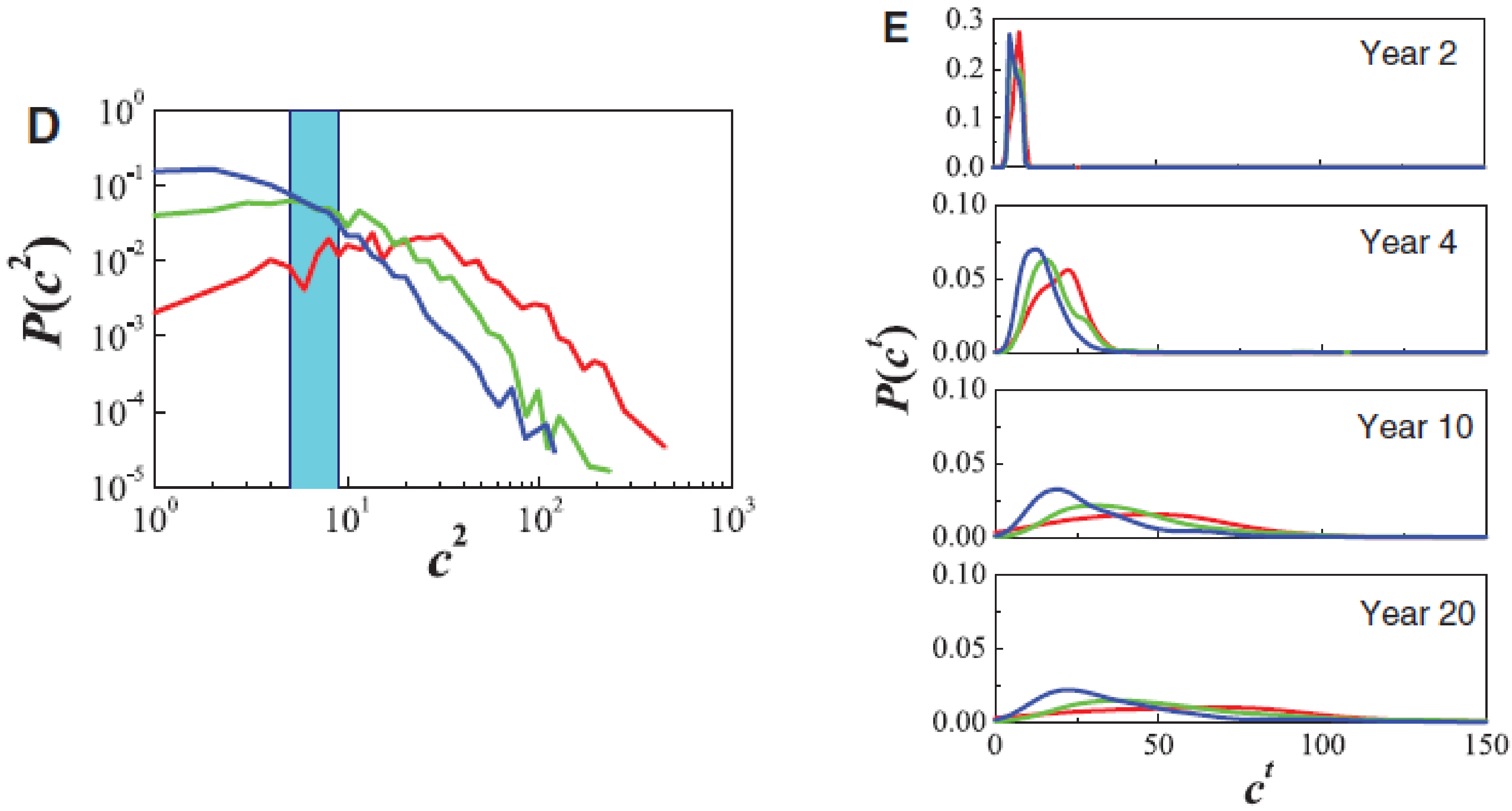
Journal	IF
Physical Review B (PRB)	3.26
Proceedings of the National Academy of Sciences USA (PNAS)	10.48
Cell	33.62

Fitness Selection



Convergence

Citation & IF Selection



Diverge

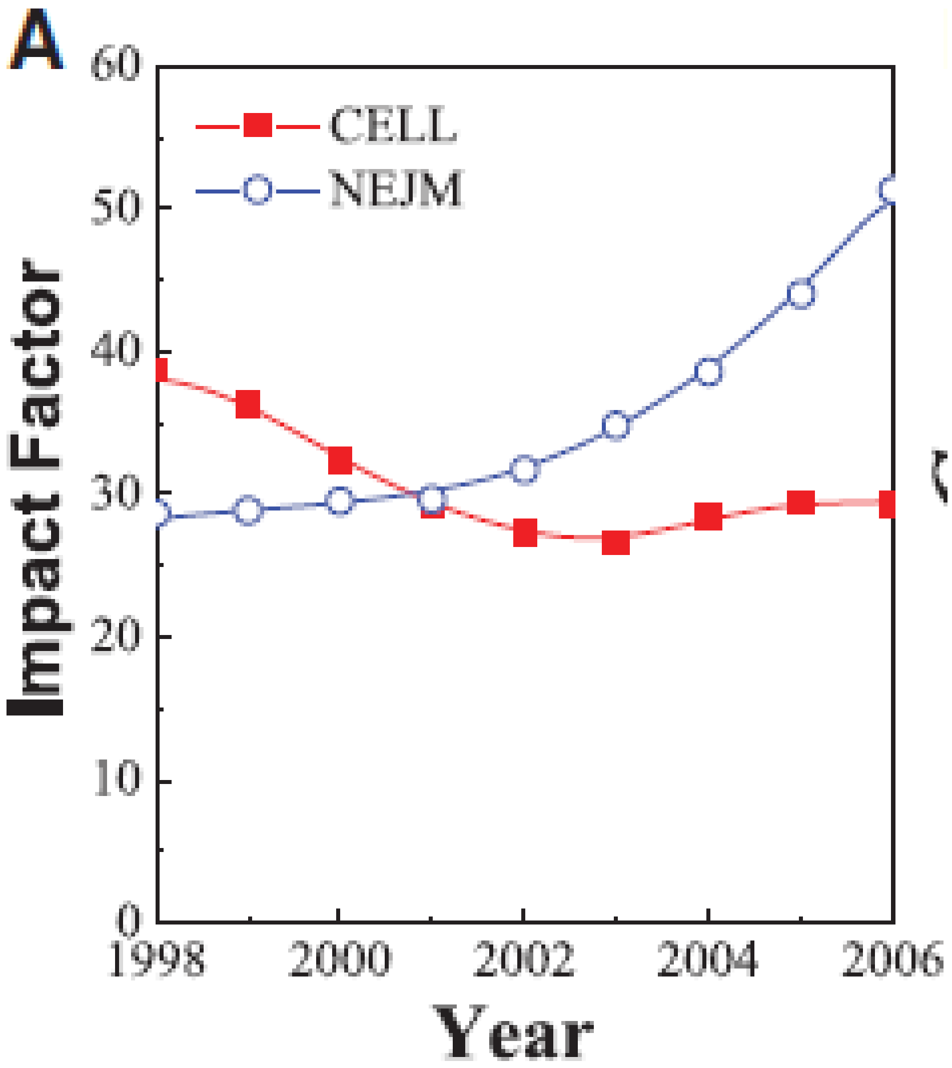
Calculating the IF

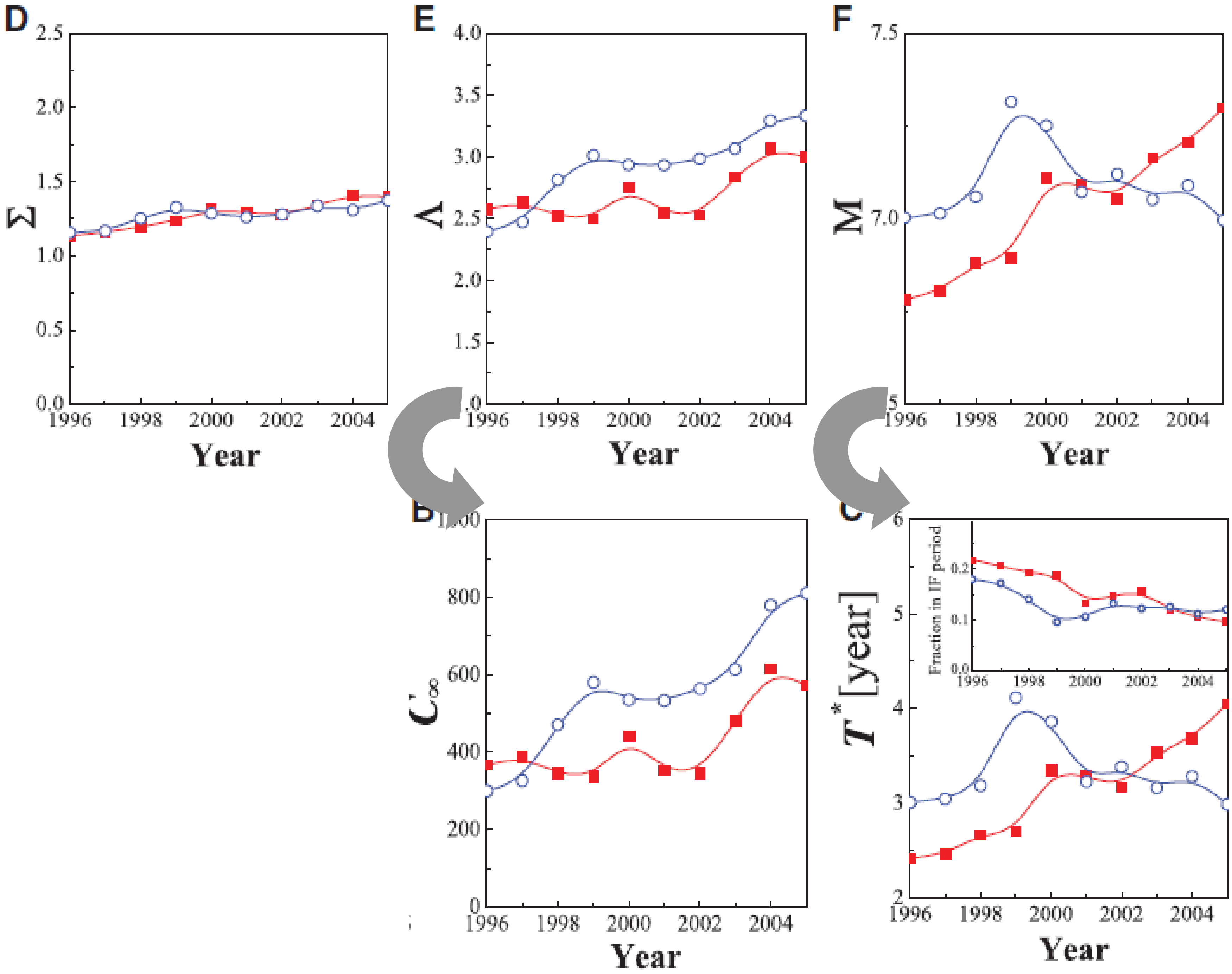
Λ M Σ ⟵ λ μ σ

$$\mathrm{IF} \approx \frac{m}{2}\left(\exp\left[\Lambda\Phi\left(\frac{M_1 - M}{\Sigma}\right)\right] - \exp\left[\Lambda\Phi\left(\frac{M_2 - M}{\Sigma}\right)\right]\right).$$

$$C^{\infty} = m\left(e^{\Lambda} - 1\right) \qquad T^{*} = \exp(M)$$

Illustrate the changes of IF




D
E
F
B
C
Σ
Λ
M
C_∞
T^* [year]
Fraction in IF period
Year

Predict future citations

$$\sigma_p^+ = \sqrt{\int_{k_p^*}^{\infty} (k_p - k_p^*)^2 P(k_p) dk_p}$$

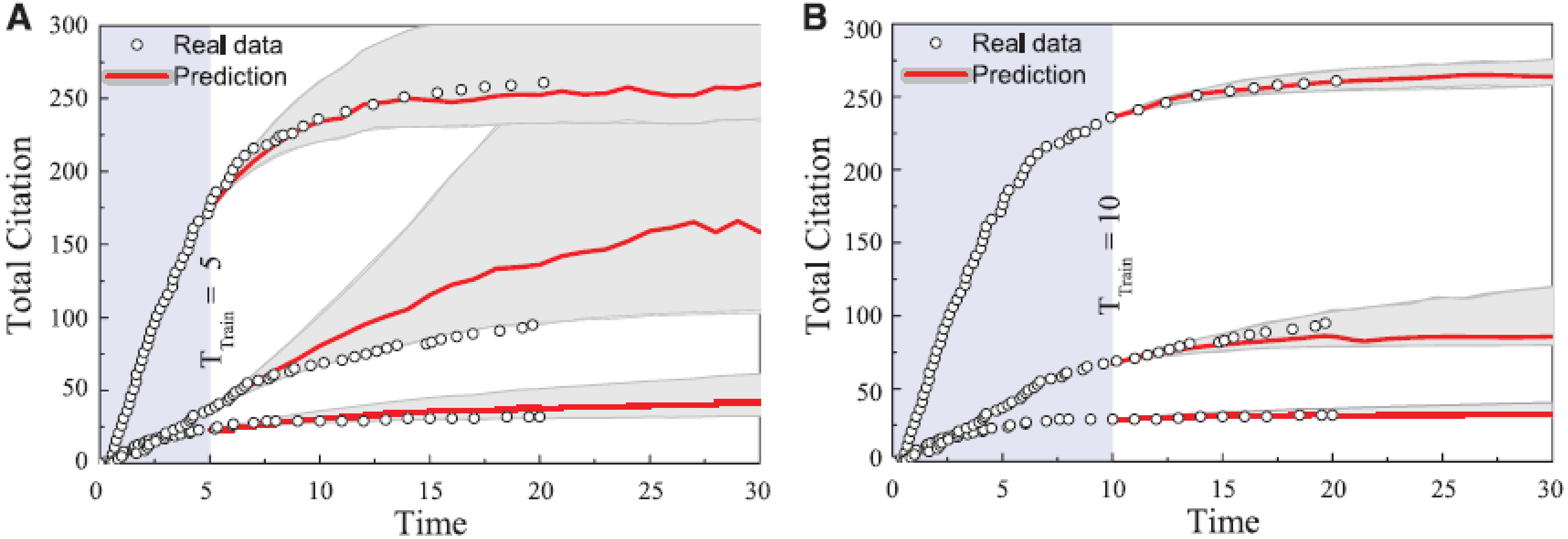
$$\sigma_p^- = \sqrt{\int_{k_t}^{k_p^*} (k_p - k_p^*)^2 P(k_p) dk_p}$$

citation envelope

$$[-\sigma_p^-, \sigma_p^+]$$

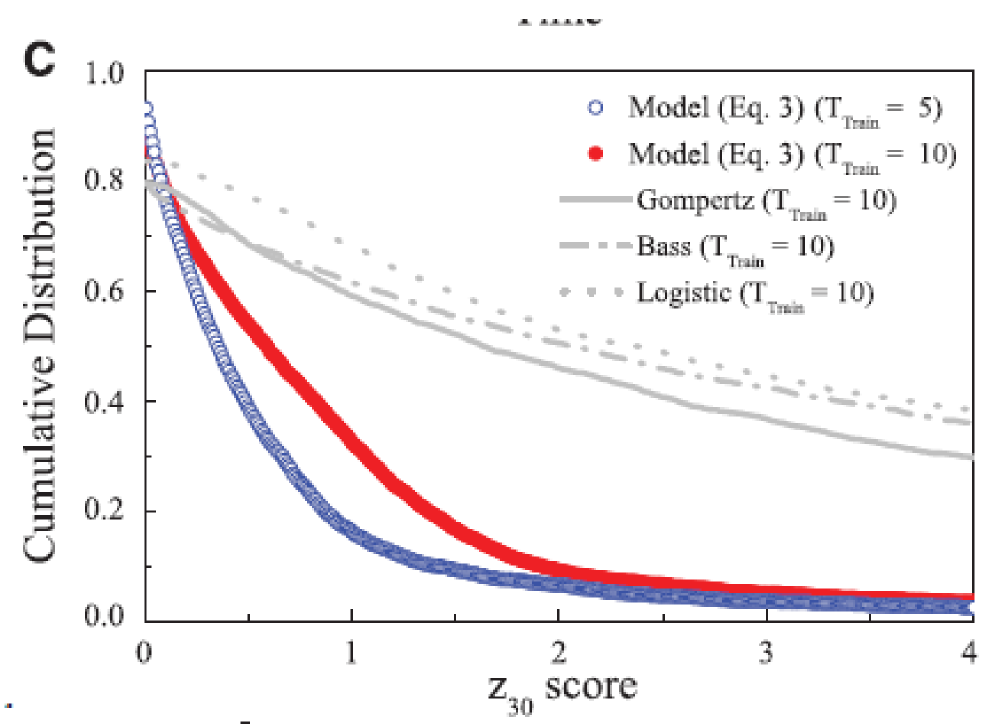
5 years T_{train}

10 years T_{train}

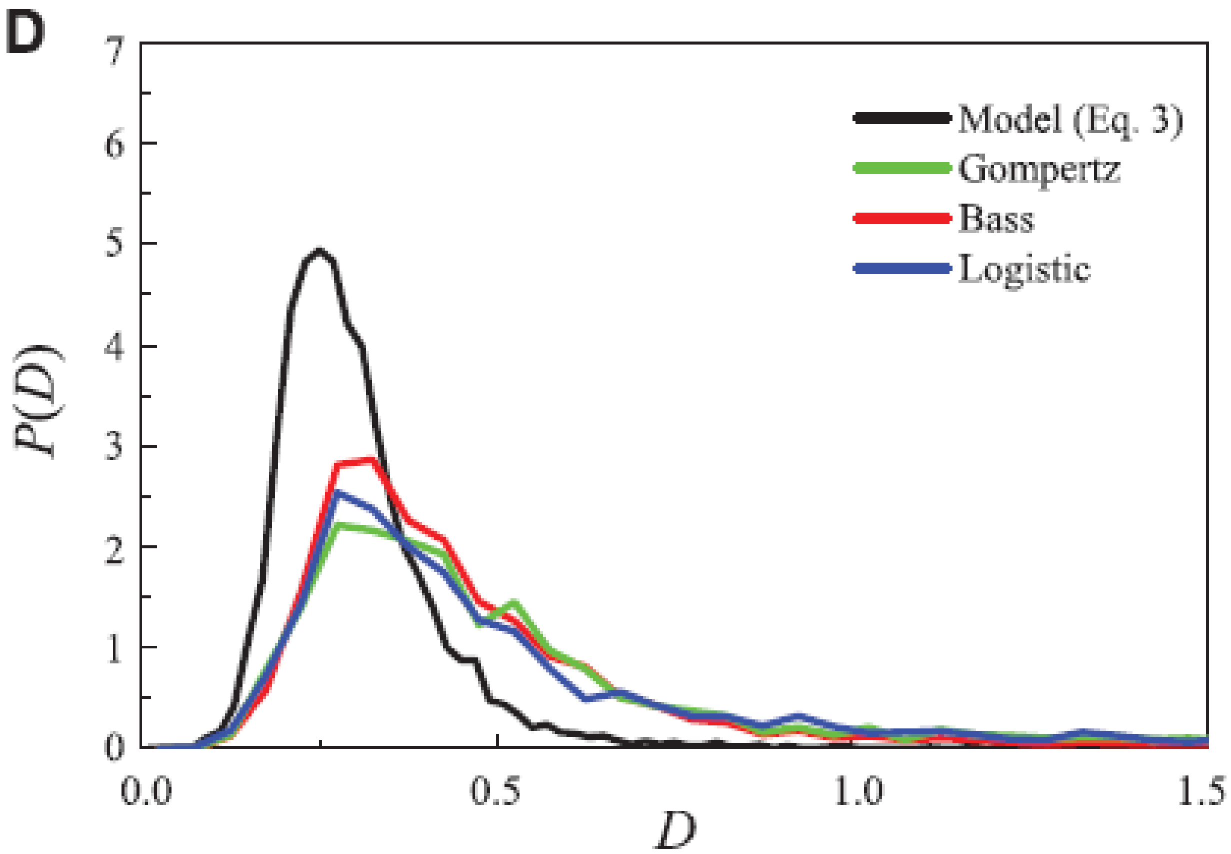


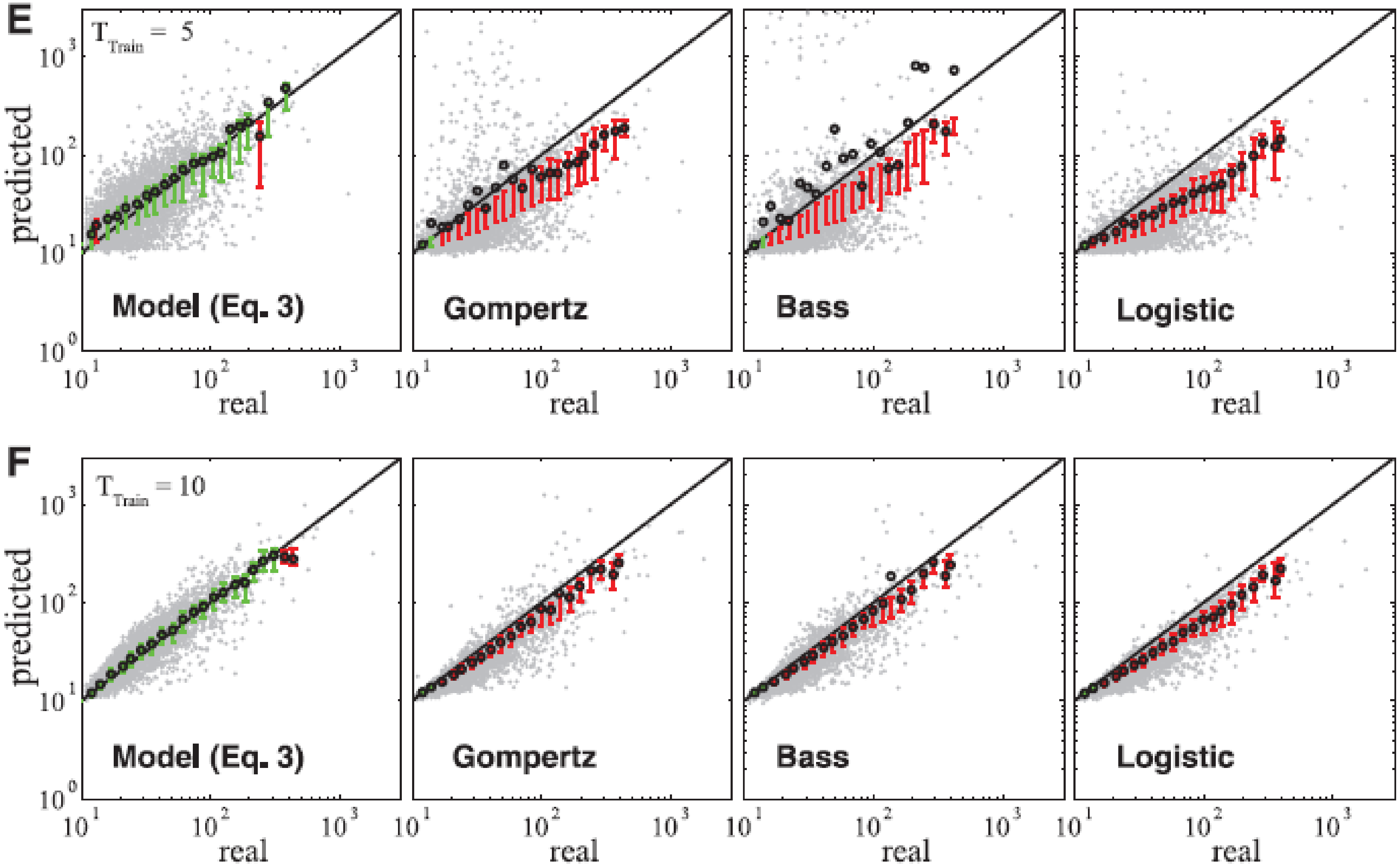
Our Model
Logistic Model
Bass Model
Gompertz Model

$$z_T = |c^T - k_p^*| / \sigma_p^+$$



Kolmogorov-Smirnov (KS) test




E
T_{Train} = 5
F
T_{Train} = 10
predicted
real
Model (Eq. 3)
Gompertz
Bass
Logistic

Limitations

- It cannot account for exogenous “second acts,”
  like the citation bump observed for
  superconductivity
  papers after the discovery of high-temperature
  superconductivity in the 1980s,
- It cannot detect delayed impact,
  like the explosion of citations to Erdős and Rényi’s
  work 4 decades after their publication, following
  the emergence of network science

Thank you for listening!