Quantifying Long-Term Scientific Impact

Ruiqi Yang

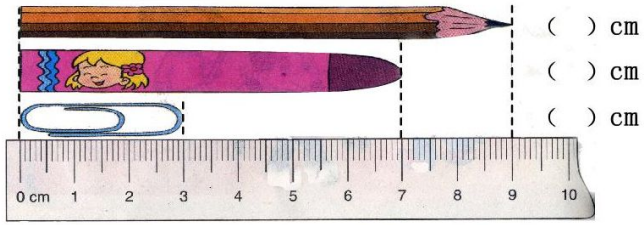
20/Jan/2016

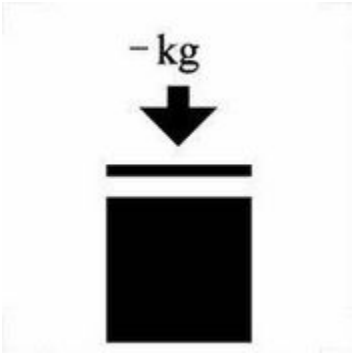
Data Mining Lab, Big Data Research Center, UESTC
Email：junmshao@uestc.edu.cn
http://staff.uestc.edu.cn/shaojunming

- **Outline:**
  - 1.Introduction
  - 2.Minimal Citation Model
  - 3.Model Validation
  - 4.Other Models

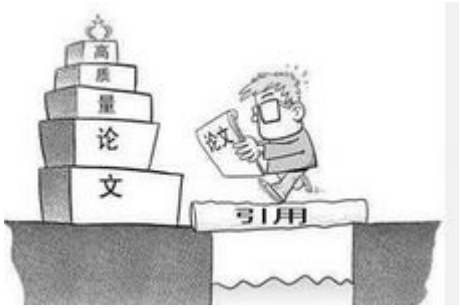
1.Introduction

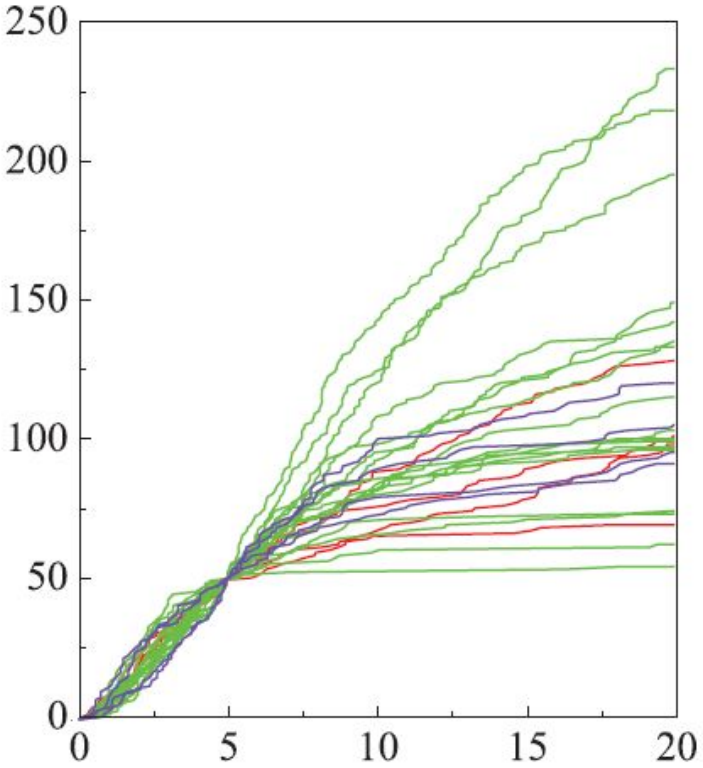
In life, we measure length by cm, dm, m, etc; and measure weight by kg...



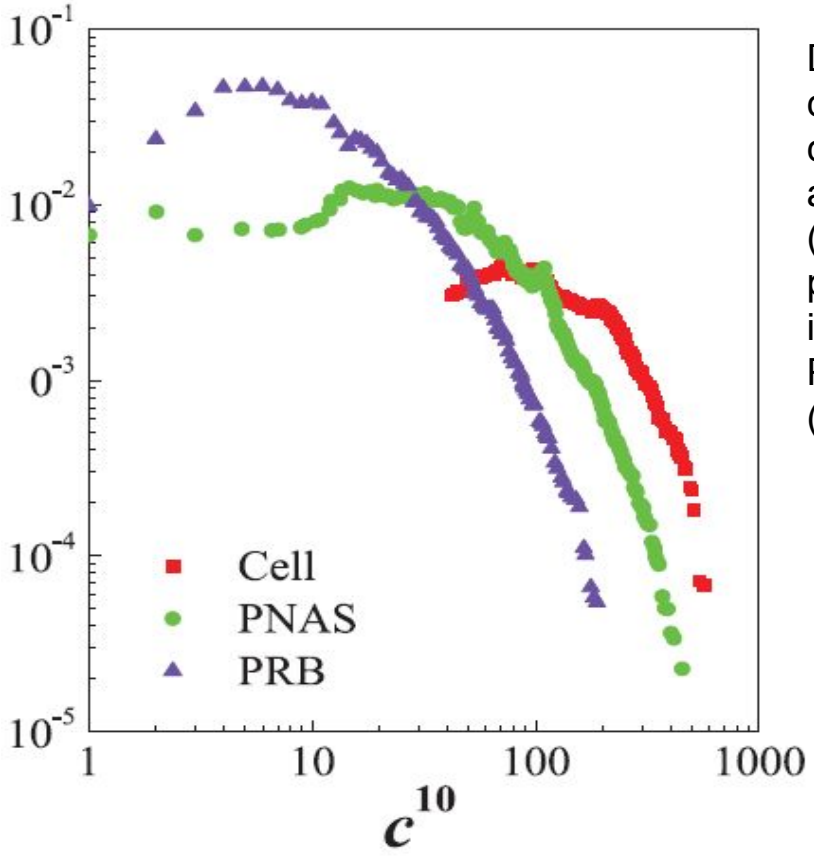


- How do we quantify the scientific impact of a paper？
  - Citation
- Example of citation-based measures
  - The number of citations, the impact factor, etc.





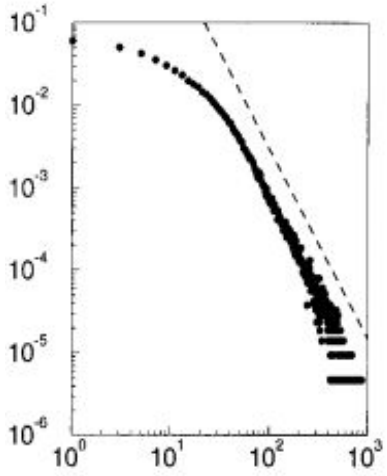
Citation history of all papers published in Cell, PNAS, and Physical Review B (PRB) in 1990 and acquired 50 citations 5 years after publication, illustrating the dierent long-term impact despite their equal early impact.



Distribution of the cumulative citations ten years after publication (c10) for all papers published in Cell, PNAS, and Physical Review B (PRB) in 1990.

2.1 Scale-Free Model $\Pi_i(t) \propto k_i^t$

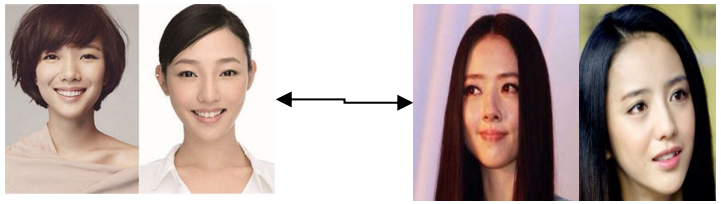
also known as Barabási-Albert(BA) model, is designed to reproduce the degree distribution of complex networks.



Actor collaboration graph with N=212,250 vertices and average connectivity <k>=28.78

$$\ln P(k) = -\gamma \cdot \ln k + C \Rightarrow P(k) = C^{'} \cdot k^{-\gamma}$$

$$P(k) \sim k^{-\gamma}$$



2.2 Fitness Model

also known as Bianconi-Barabási(BB) model, besides the PA mechanism each node i has an initial fitness η_i capturing its unique likelihood to be cited in the future.

$$\Pi_i \propto \eta_i k_i^t$$

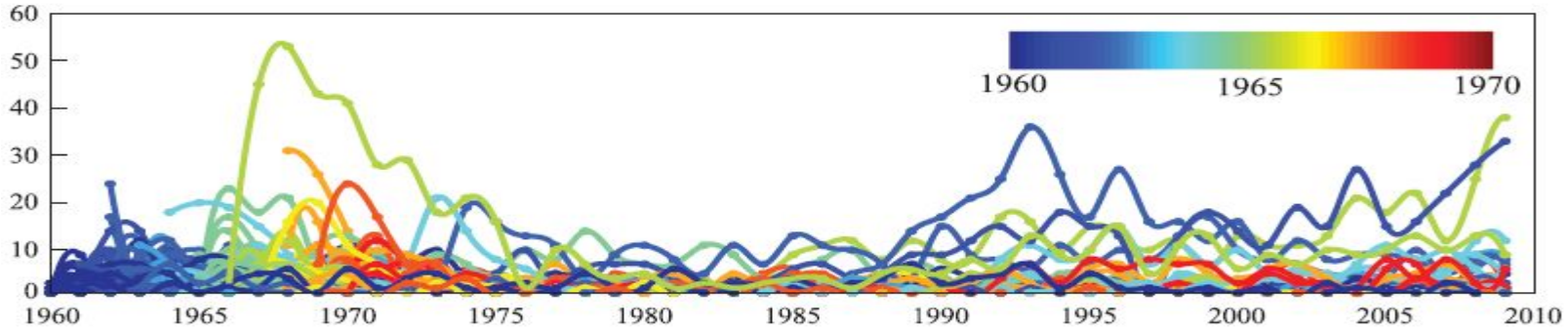
it helps explain why the degrees of some nodes with low initial degrees grow faster than others, even faster than the old nodes with high degrees.

2.3Minimal Citation Model

input: citation history of the target paper $\{t_i\}$

- The citation probability $\Pi_i(\Delta t_i) \sim \eta_i c_i^t P(\Delta t_i)$
- elaps time, Δt_i
- fitness, η_i
- current citations, c_i^t
- aging function, P(·)

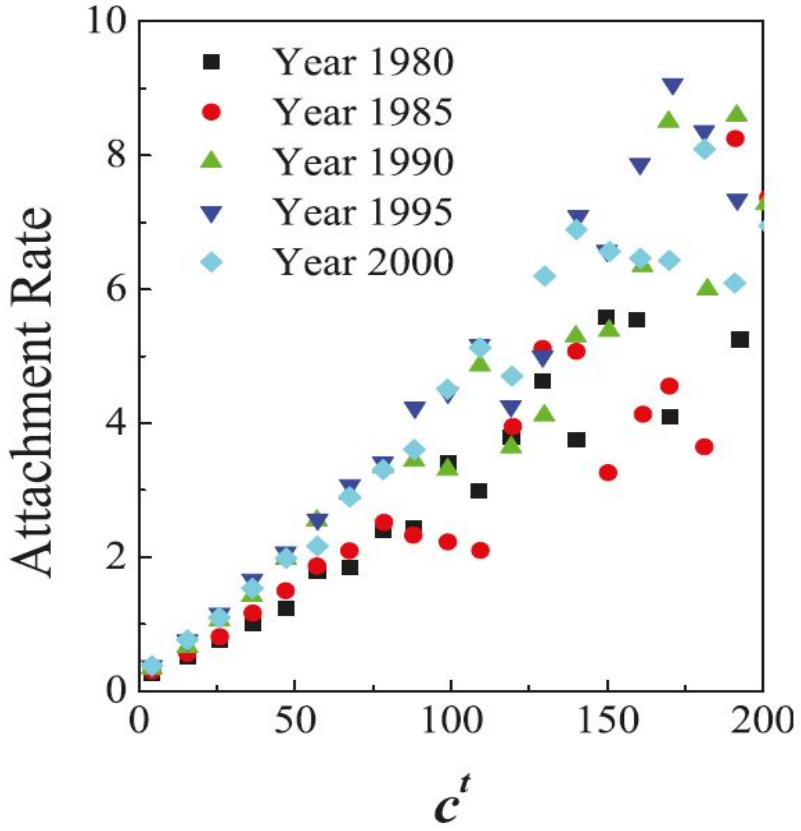
$$c_i^{\Delta t_i} = m(e^{\lambda_i \Phi(\frac{\ln \Delta t_i - \mu_i}{\sigma_i})} - 1)$$



2.4 Solving the model

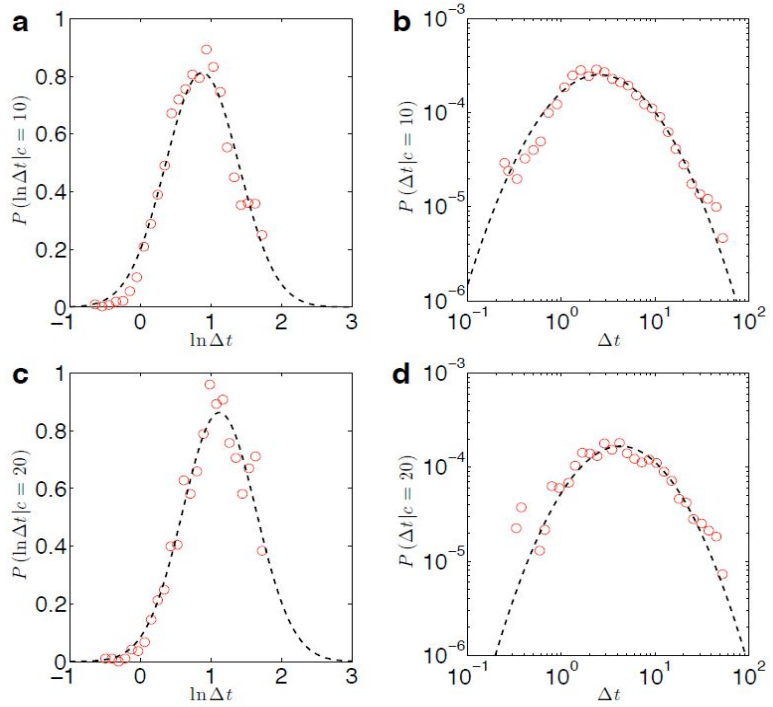
2.4.1 PA

$$\Pi_i(\Delta t_i) \sim \eta_i c_i^t P(\Delta t_i)$$



Empirical validation of preferential attachment. Attachment rate measures the likelihood for new papers published in different years (color coded) to cite an old paper with c^t citations. That is, for each year, c^t measures the citations of each paper before this year, and attachment rate measures the average number of times each paper with c^t citations was cited in this year. The linearity of the curves offers evidence for preferential attachment.

2.4.2 The temporal relaxation function



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

if i.i.d $\{x_i\}$, $\Delta t = \prod_i^n x_i$, then $\ln\Delta t = \sum_i^n x_i$ converges to a normal distribution due to central limit theorem

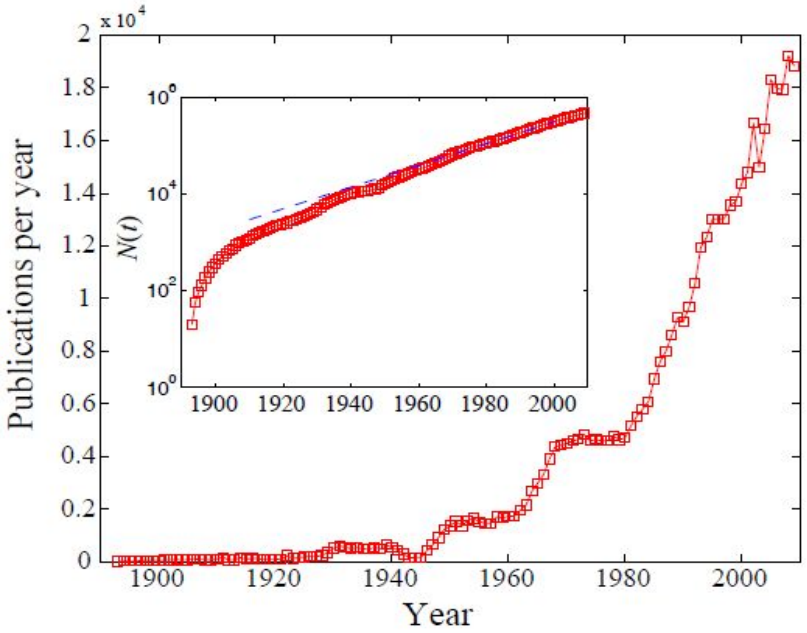
$P(\ln\Delta t)$ when papers change from 10 citations to 11 citations. The dashed line corresponds to the best gaussian fitting. (b) Same as (a) but for $P(\Delta t)$. Dashed line corresponds to the best lognormal fitting ($\mu=7.85$ and $\sigma=1.01$). Here Δt is measured in unit of years. (c) $P(\ln\Delta t)$ when papers change from 20 citations to 21 citations. The dashed line corresponds to the best gaussian fitting. (d) Same as (c) but for $P(\Delta t)$. The dashed line corresponds to the best lognormal fitting ($\mu=8.29$ and $\sigma=0.93$)

2.4.3 formula derivation

$$\frac{\mathrm{d}c_i^t}{dN(t)} = \mathrm{m} \cdot \frac{\Pi_i}{\sum_{i=1}^{N(t)} \Pi_i}$$

$$\sum_{\mathrm{i}=1}^{N(t)} \frac{\Pi_i}{\sum_{i=1}^{N(t)} \Pi_i} = \frac{\sum_{i=1}^{N(t)} \Pi_i}{\sum_{i=1}^{N(t)} \Pi_i} = 1$$

$$m \cdot \sum_{i=1}^{N(t)} \frac{dc_i^t}{dt} = \frac{dN(t)}{dt}$$



The number of papers published each year in the PR corpus. Inset: cumulative number of papers N(t) published up to year t.

$$N(t) \sim \exp(\beta t) \qquad \Pi_i(\Delta t_i) \sim \eta_i c_i^t P(\Delta t_i)$$

$$\frac{dc_i^t}{dt} = m \cdot \beta \frac{\eta_i c_i^t P(\Delta t_i)}{\frac{1}{N(t)} \sum_{i=1}^{N(t)} \Pi_i} \qquad \Delta \mathrm{t}_i = t - t_i, t = \Delta t_i + t_i$$

$$let \;\; A = \frac{1}{N(t)} \sum_{i=1}^{N(t)} \Pi_i, \lambda_i = \frac{\beta \eta_i}{A} \qquad \frac{1}{\mathrm{c}_i^t} dc_i^t = m \cdot \lambda_i P(\Delta t_i) d\Delta t_i$$

$$\mathrm{c}_i^t = C' e^{m \lambda_i \Phi(\frac{\ln \Delta t_i - \mu_i}{\sigma_i})}$$

assuming $c_i = m(f(\eta_i, \Delta t_i) - 1)$

$$\frac{df(\eta_i, \Delta t_i)}{d\Delta t_i} = \beta \frac{\eta_i f(\eta_i, \Delta t_i) P_t(\Delta t_i)}{A}$$

$$c_i^{\Delta t_i} = m(e^{\lambda_i \Phi(\frac{\ln \Delta t_i - \mu_i}{\sigma_i})} - 1), c_i^{\infty} = m(e^{\lambda_i} - 1)$$

so, here is a question, how to estimate the three parameters $(\lambda_i, \mu_i, \sigma_i)$

2.5 Survival Analysis

- the survival function $S(t) = P(T > t)$
- lifetime distribution function $F(t) = P(T \le t) = 1 - S(t)$
- the density function of the lifetime distribution $f(t) = F'(t) = \frac{d}{dt}F(t)$
- hazard function $\lambda(t) = \lim_{dt \to 0} \frac{P(t \le T < t + dt)}{dt \cdot S(t)} = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)}$
- cumulative hazard function $\Lambda(t) = -\log S(t)$ $\frac{d}{dt}\Lambda(t) = -\frac{S'(t)}{S(t)} = \lambda(t), \Lambda(t) = \int_0^t \lambda(u)du$

Future life time at a given time t_0 is the time remaining until death, given survival to age t_0 , then the probability of death at or before age $t_0 + t$: $P(T \le t_0 + t \mid T > t_0) = \frac{P(t_0 < t \le t_0 + t)}{P(T > t_0)} = \frac{F(t_0 + t) - F(t_0)}{S(t_0)}$

Therefore the probability density of future lifetime is:

$$\frac{d}{dt}\frac{F(t_0 + t) - F(t_0)}{S(t_0)} = \frac{f(t_0 + t)}{S(t_0)}$$

$$P(T > t_2 \mid T > t_1) = \frac{S(t_2)}{S(t_1)} = \frac{e^{-\log S(t_1)}}{e^{-\log S(t_2)}} = \frac{e^{\Lambda t_1}}{e^{\Lambda t_2}} = e^{-(\Lambda(t_2) - \Lambda(t_1))} = e^{-(\int_0^{t_2} \lambda(u)du - \int_0^{t_1} \lambda(u)du)} = e^{-\int_{t_1}^{t_2} \lambda(u)du}$$

$$\frac{f(t_2)}{S(t_1)} = -\frac{S'(t_2)}{S(t_2)} \cdot \frac{S(t_2)}{S(t_1)} = \lambda(t_2) \cdot e^{-\int_{t_1}^{t_2} \lambda(u)du}$$

Imagine a stochastic process $\{x(t)\}$ where $x(t)$ represents the number of events by time t, satisfying:

$$P(x(t+h)-x(t)=1)=\lambda_0(x,t)h+O(h)$$

where $\lambda_0(x,t)$ is a time dependent rate parameter.

$$L=\prod_{i=1}^{N}P(t_i\mid t_{i-1})=\prod_{i=1}^{N}[\lambda_0(i-1,t_i)\cdot e^{-\int_{t_{i-1}}^{t_i}\lambda_0(i-1,t)dt}]=\prod_{i=1}^{N}\lambda_0(i-1,t_i)\cdot e^{-\int_0^T\lambda_0(x(t),t)dt}$$

$$\frac{df(\eta_i,\Delta t_i)}{d\Delta t_i}=\beta\frac{\eta_i f(\eta_i,\Delta t_i)P_t(\Delta t_i)}{A}$$

$$\frac{f(t_i)}{S(t_{i-1})}$$

$$\lambda_0(x,t)=\frac{dc_i^t}{dt}=\frac{d(m(f(\eta_i,t)-1))}{dt}=\frac{\beta\eta_i}{A}\cdot(m(f(\eta_i,t)-1)+m)P_t(t)=\frac{\beta\eta_i}{A}\cdot(x+m)P_t(\Delta t_i)=\frac{\lambda_i(x+m)}{\sqrt{2\pi}\sigma t}\exp(-\frac{(\ln(t)-\mu)^2}{2\sigma^2})$$

$$(\lambda^*, \mu^*, \sigma^*) = \arg\max_{\lambda,\mu,\sigma} L(\lambda, \mu, \sigma)$$

$$\frac{\partial L(\lambda^*, \mu^*, \sigma^*)}{\partial \lambda^*} = 0, \frac{\partial L(\lambda^*, \mu^*, \sigma^*)}{\partial \mu^*} = 0, \frac{\partial L(\lambda^*, \mu^*, \sigma^*)}{\partial \sigma^*} = 0$$

$$\lambda^* = \left[(1+\hat{m})\Phi\left(\frac{\ln(T)-\mu^*}{\sigma^*}\right) - \left\langle \Phi\left(\frac{\ln(t_i)-\mu^*}{\sigma^*}\right) \right\rangle \right]^{-1}.$$

$$\left\langle \frac{\ln(t_i)-\mu^*}{\sigma^*} - \lambda^* P_G\left(\frac{\ln(t_i)-\mu^*}{\sigma^*}\right) \right\rangle + \lambda^*(1+\hat{m}) P_G\left(\frac{\ln(T)-\mu^*}{\sigma^*}\right) = 0$$

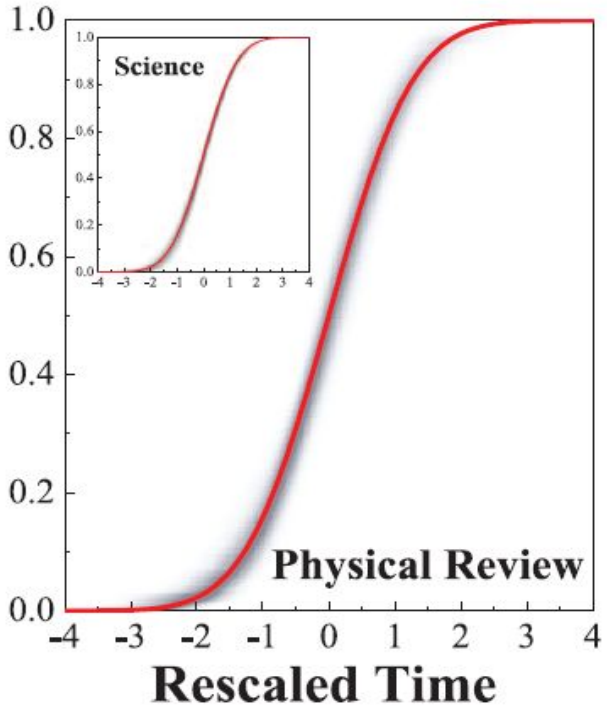
$$\left\langle \frac{\ln(t_i)-\mu^*}{\sigma^*} \left(\frac{\ln(t_i)-\mu^*}{\sigma^*} - \lambda^* P_G\left(\frac{\ln(t_i)-\mu^*}{\sigma^*}\right) \right) \right\rangle + \lambda^*(1+\hat{m}) \frac{\ln(T)-\mu^*}{\sigma^*} P_G\left(\frac{\ln(T)-\mu^*}{\sigma^*}\right) = 1$$

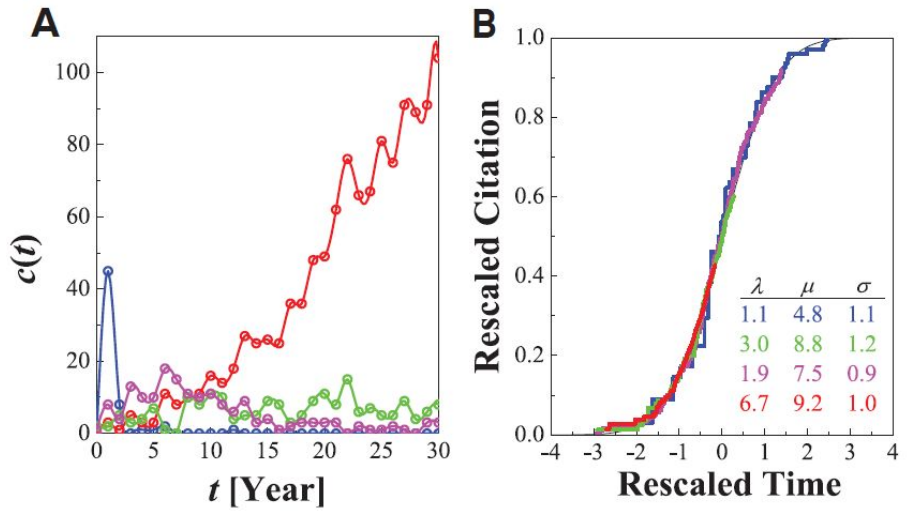
3. Model Validation

3.1. Fitting the original data

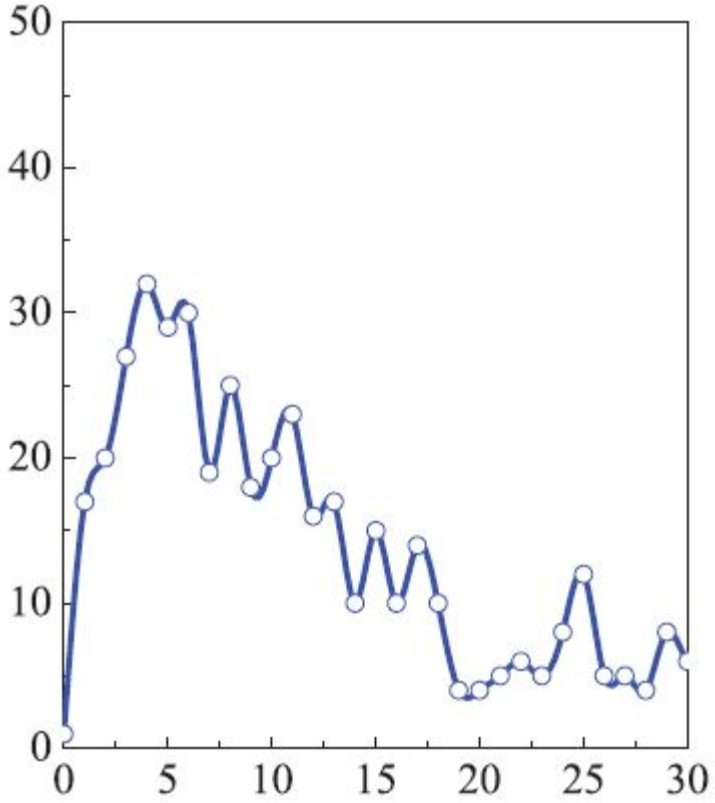
$$c_i^{\Delta t_i} = m(e^{\lambda_i \Phi(\frac{\ln \Delta t_i - \mu_i}{\sigma_i})} - 1)$$

let $\tilde{t} = (\ln t - \mu_i), \tilde{c} = \ln(1 + c_i^t / m) / \lambda_i \Rightarrow \tilde{c} = \Phi(\tilde{t})$

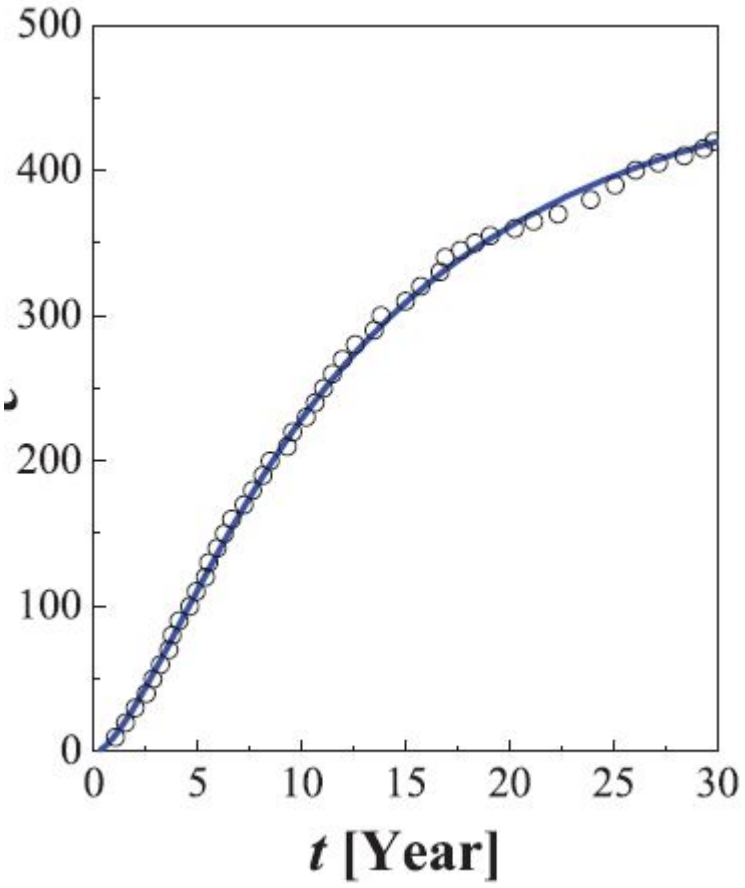




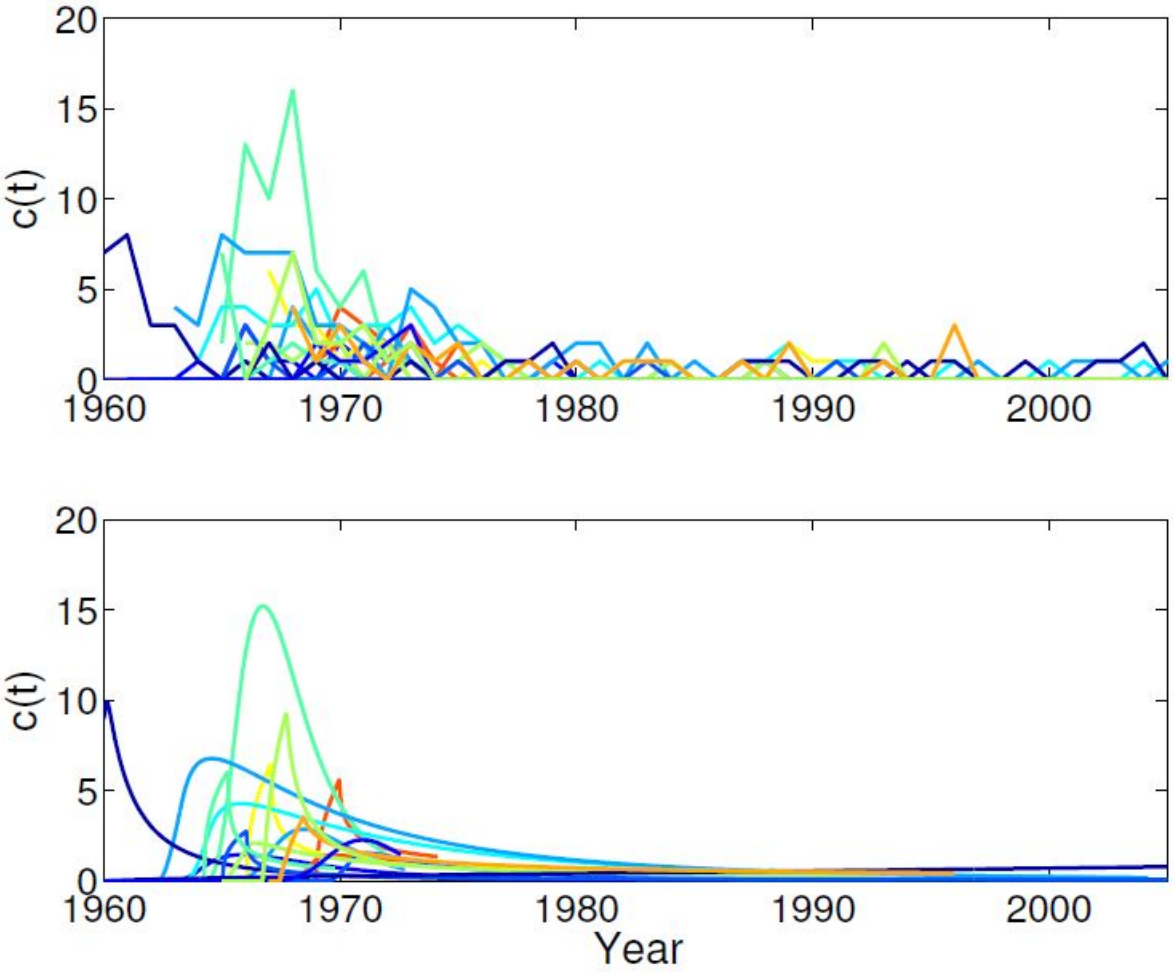
Citation history of four papers published in PR in 1964, selected for their distinct dynamics, displaying a 'jump-decay' pattern (blue); delayed peak (magenta); attracting a constant number of citation over time (green), or acquiring an increasing number of citations each year (red).



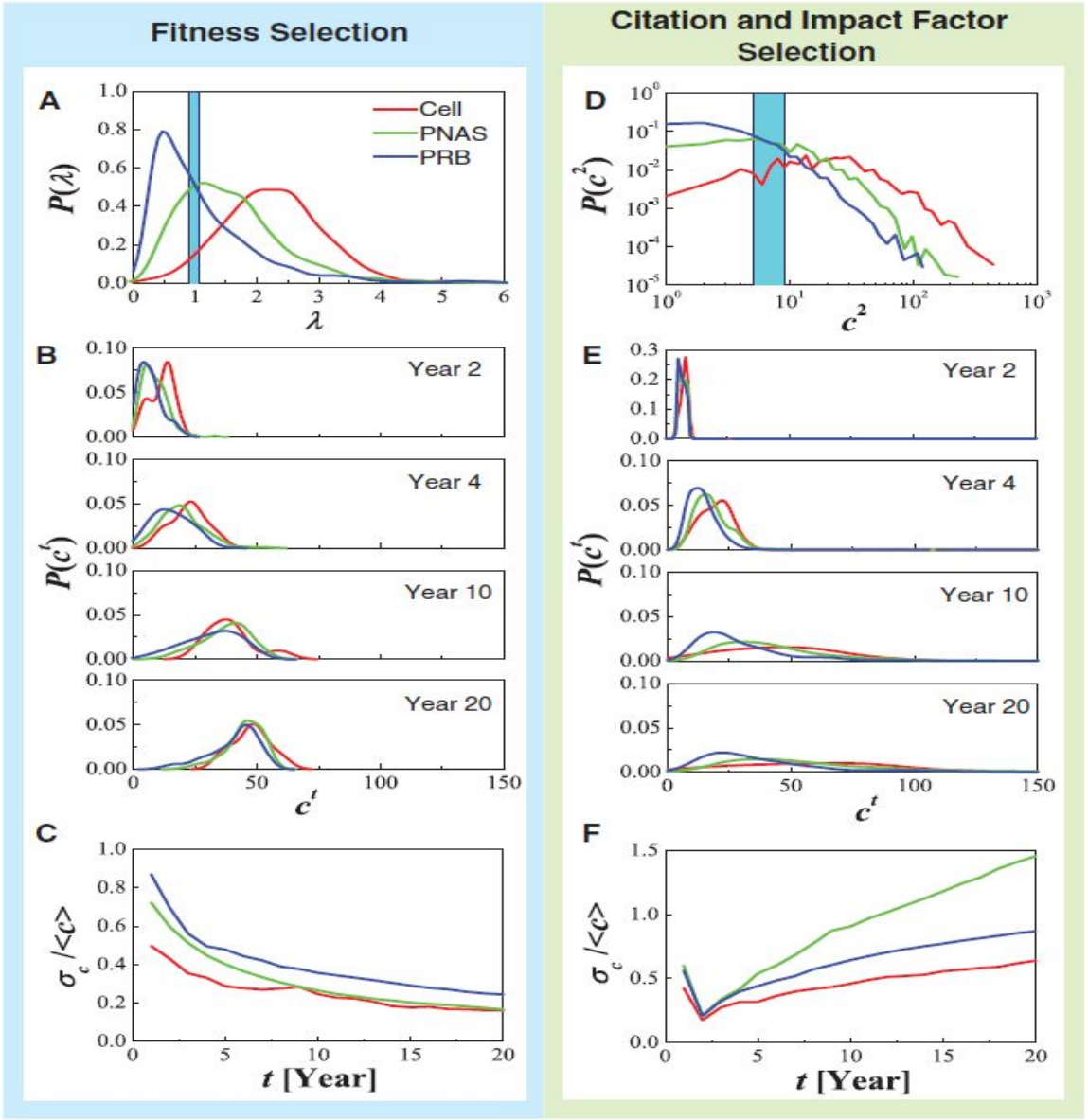
Yearly citation c(t) for a research paper from the PR corpus



Cumulative citations ct for the paper in the left graph.



We randomly selected two papers each year between 1960 to 1970 from the PR corpus. Their citation histories are shown on the top panel.
Color code corresponding to the publication year.



$$c_i^{\Delta t_i} = m(e^{\lambda_i \Phi(\frac{\ln \Delta t_i - \mu_i}{\sigma_i})} - 1), c_i^{\infty} = m(e^{\lambda_i} - 1)$$

Evaluating long-term impact. (A) Fitness distribution P(l) for papers published by Cell, PNAS, and PRB in 1990. Shaded area indicates papers in the l $\approx$ 1 range, which were selected for further study.(B) Citation distributions for papers with fitness l $\approx$ 1, highlighted in (A), for years 2, 4, 10, and 20 after publication. (C) Time-dependent relative variance of citations for papers selected in (A). (D) Citation distribution 2 years after publication [P(c2)] for papers published by Cell, PNAS, and PRB. Shaded area highlights papers with c2$\in$[5,9] that were selected for further study. (E) Citation distributions for papers with c2$\in$[5,9], selected in (D), after 2, 4, 10, and 20 years. (F) Time-dependent relative variance of citations for papers selected in (D).

Quantifying Journal Impact

$$C_j^t = \frac{1}{N_j}\sum_i^{N_j} c_i^t \qquad C_j^t = m\left(e^{\Lambda_j \Phi\left(\frac{\ln T - M_j}{\Sigma_j}\right)} - 1\right).$$

$$\mathrm{IF}(T) = \frac{\sum_i^{N_1} c_i(T|T_1) + \sum_i^{N_2} c_i(T|T_2)}{N_1 + N_2}$$

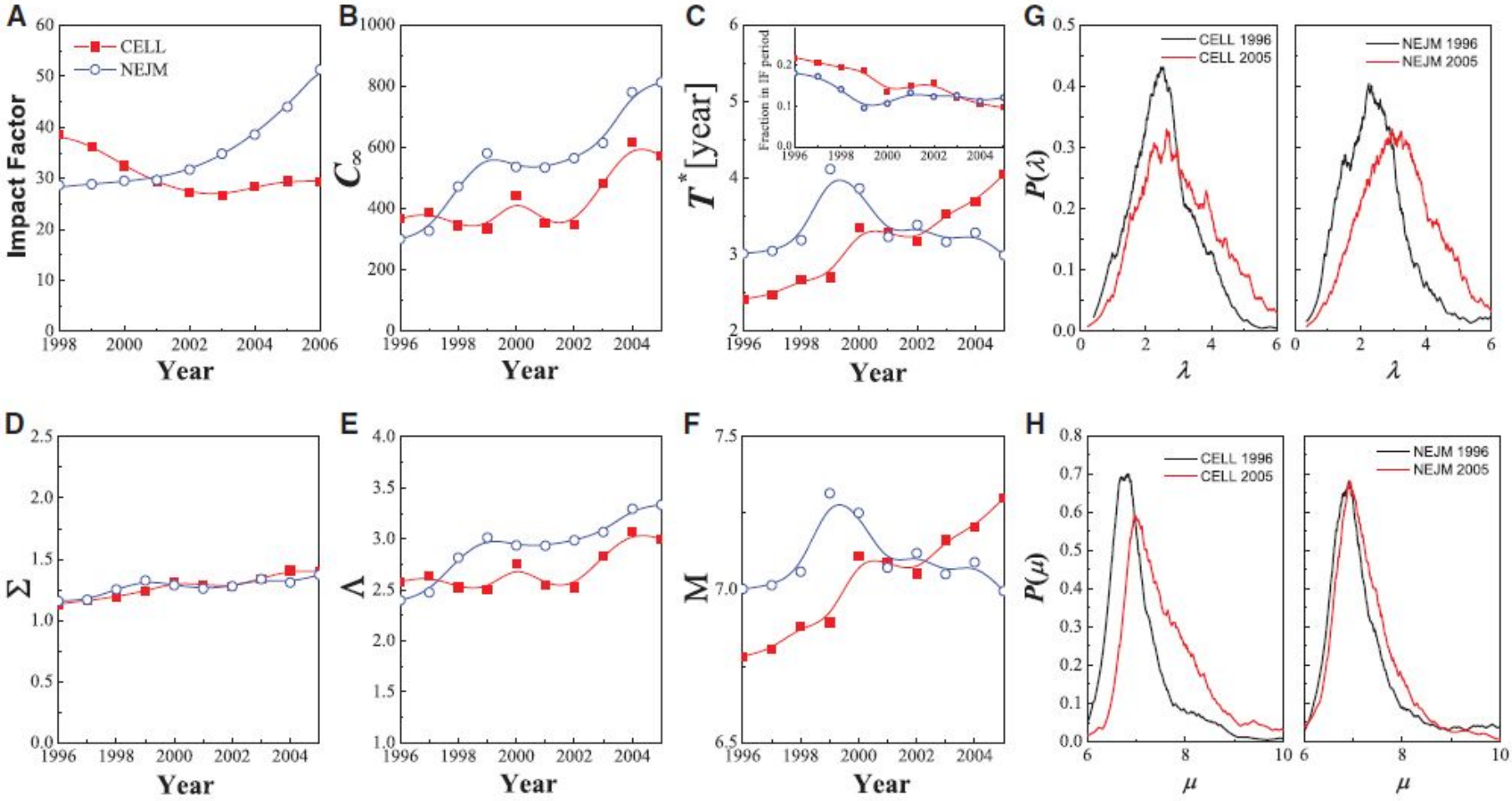
$$\mathrm{IF}(T) = \frac{N_1 C(T|T_1) + N_2 C(T|T_2)}{N_1 + N_2}$$

$$= \frac{mN_1}{N_1+N_2}\left(e^{\Lambda(T_1)\Phi\left(\frac{M_1 - M(T_1)}{\Sigma(T_1)}\right)} - e^{\Lambda(T_1)\Phi\left(\frac{M_3 - M(T_1)}{\Sigma(T_1)}\right)}\right) + \frac{mN_2}{N_1+N_2}\left(e^{\Lambda(T_2)\Phi\left(\frac{M_3 - M(T_2)}{\Sigma(T_2)}\right)} - e^{\Lambda(T_2)\Phi\left(\frac{M_2 - M(T_2)}{\Sigma(T_2)}\right)}\right).$$

where $T_1 = T-2, T_2 = T-1$ $\quad (\Lambda, M, \Sigma) \equiv (\Lambda(T_1), M(T_1), \Sigma(T_1)) = (\Lambda(T_2), M(T_2), \Sigma(T_2))$

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

In 1998, the IFs of Cell and NEJM were 38.7 and 28.7, respectively. Over the next decade, there was a remarkable reversal: NEJM became the first journal to reach IF = 50, whereas Cell's IF decreased to around 30.



3. 2. Predicting Citations

Using the training period T_t and k_t sampling citations, to predict the number of citations at a future time T_p

the expected increment of citations between $(T_t, T_p]$:

$$\overline{\Delta k} = (k_t + m)\left(e^{\eta(\Phi((\ln T_p - \mu)/\sigma) - \Phi((\ln T_t - \mu)/\sigma))} - 1\right).$$

Hence, the expected citation at time T_p :

$$\overline{k}(\eta, \mu, \sigma) = (k_t + m)e^{\eta(\Phi((\ln T_p - \mu)/\sigma) - \Phi((\ln T_t - \mu)/\sigma))} - m$$

assuming uniform prior distribution of (η, μ, σ)

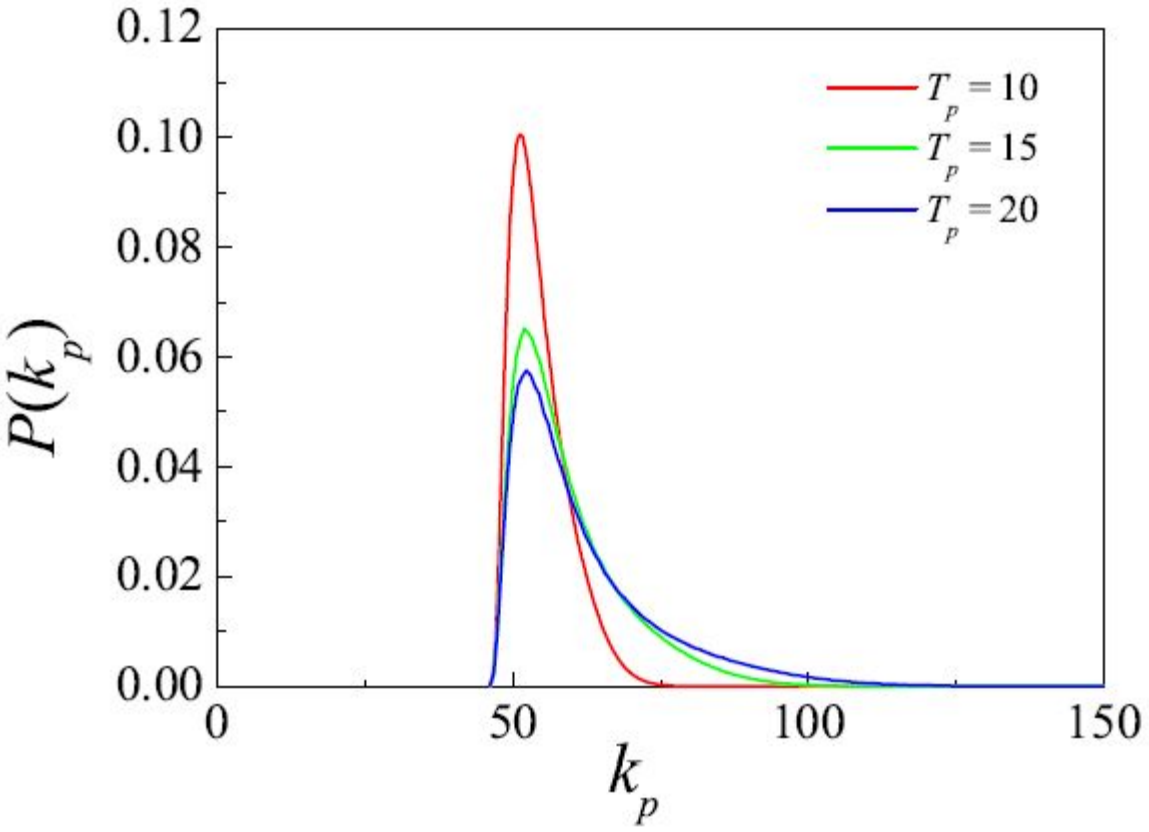
$$P(\eta,\mu,\sigma) \propto L = e^{\ln L(\eta,\mu,\sigma)} \qquad L = \prod_{i=1}^{N} \lambda_0(i-1,t_i)\cdot e^{-\int_0^T \lambda_0(x(t),t)dt}$$

Therefore, given a citation history, we can use the model to predict the probability for the paper to have k_p citations at the time T_p ,

$$P(k_p) = \int \delta\left(\overline{k}(\eta,\mu,\sigma) - k_p\right) P(\eta,\mu,\sigma) d\eta d\mu d\sigma.$$

Hence, the most probable future citation kp can be obtained from

$$\left.\frac{dP(k_p)}{dk_p}\right|_{k_p=k_p^*} = 0,$$



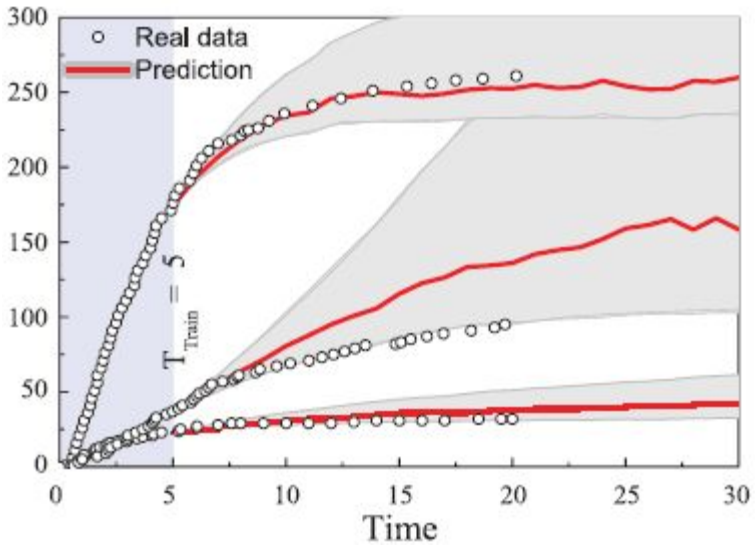
Illustrative example of P(kp) for a randomly selected paper. Different lines correspond to different testing period (Tp).

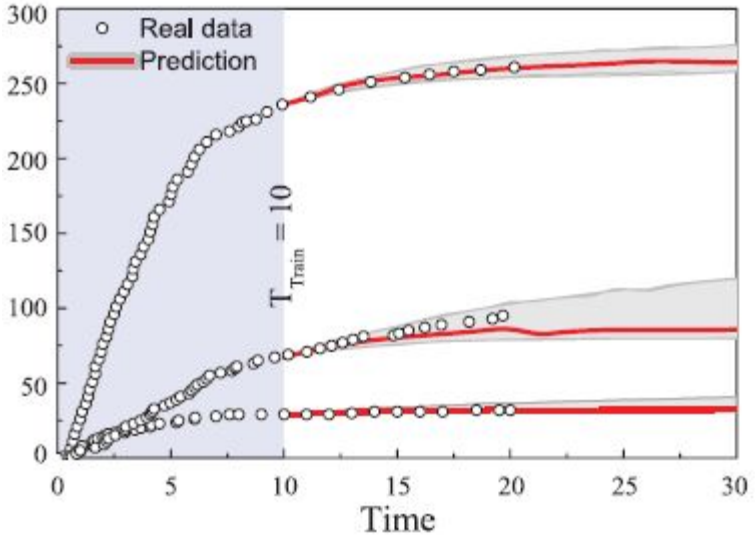
the upper/lower uncertainty can be obtained by

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

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

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





BaseLines(Diffusion of Innotations)

- Logistic Model

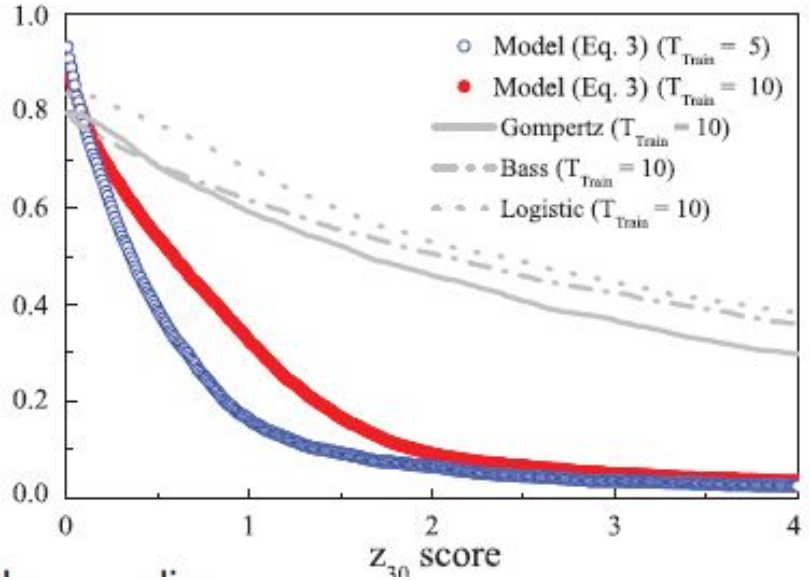
$$c_i^t = \frac{c_i^\infty}{1 + e^{-r_i(t-\tau_i)}}$$

- Bass Model

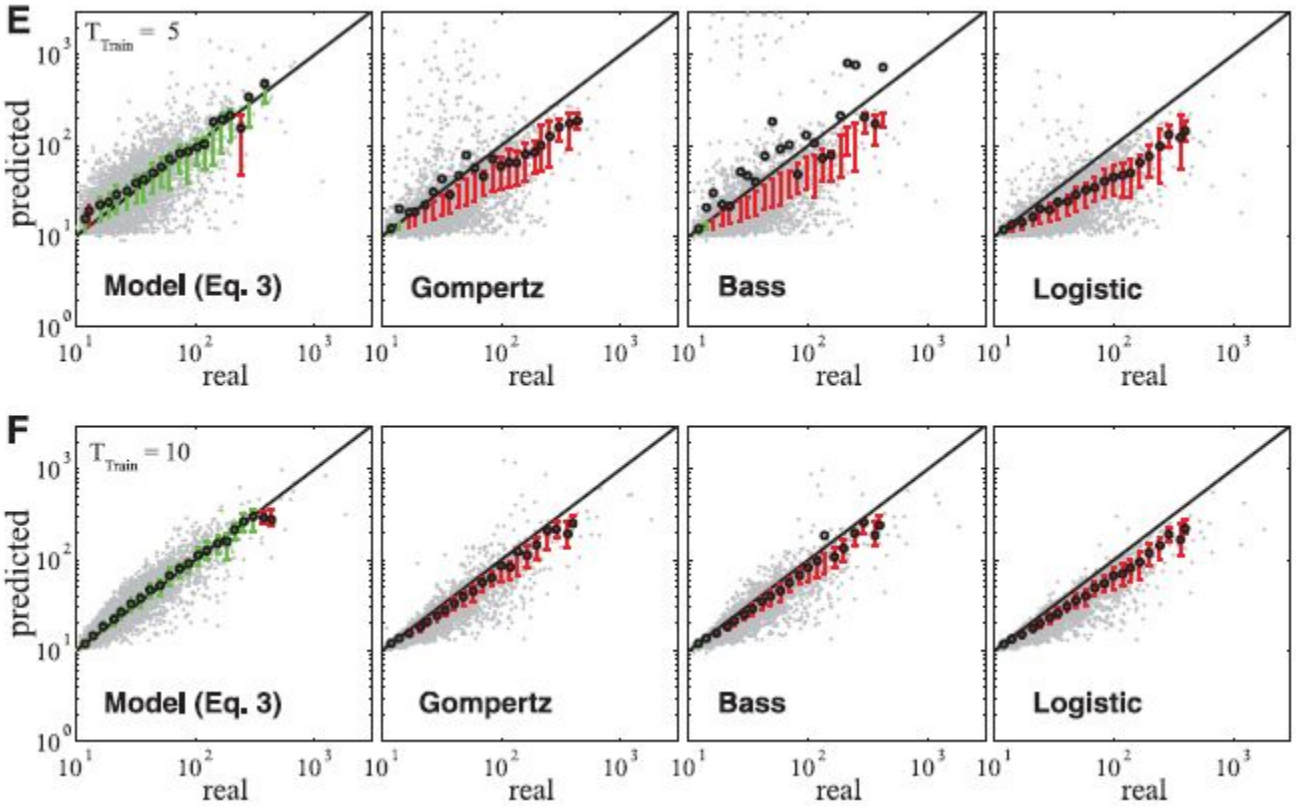
$$c_i^t = c^\infty \frac{1 - e^{-(p+q)t}}{1 + \frac{q}{p} e^{-(p+q)t}}$$

- Gompertz Model

$$c_i^t = c^\infty e^{-e^{-(a+qt)}}$$



We selected papers published in 1960s in the PR corpus that acquired at least 10 citations in 5 years (4492 in total). The red curve captures predictions for 30 years afterpublication for TTrain = 10, indicating that for our model 93.5%papers have z30 $\leqslant$ 2. The blue curve relies on 5-year training. The gray curves capture the predictions of Gompertz, Bass, and logisticmodels for 30 years after publication by using 10 years as training.


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

4. Other Models

4.1 RPP(Reinforced Poisson Process)

$$x_d(t) = \lambda_d f_d(t;\theta) i_d(t) \qquad \frac{dc_i^t}{dt} = m \cdot \lambda_i c_i^t P(\Delta t_i)$$

$$L = P(T \mid t_N) \prod_{i=1}^{N} P(t_i \mid t_{i-1}) \qquad L = \prod_{i=1}^{N} P(t_i \mid t_{i-1})$$

bayes formula: $P(h \mid D) = P(D \mid h) \cdot P(h) / P(D)$

bring in conjugate prior to eliminate over-fitting.

$$\mathcal{L}(\lambda_d, \theta_d) = p_0(T|t_{n_d}^d) \prod_{i=1}^{n_d} p_1(t_i^d|t_{i-1}^d)$$

$$= \lambda_d^{n_d} \prod_{i=1}^{n_d} (m+i-1) f_d(t_i^d;\theta_d) \times e^{-\lambda_d\left((m+n_d)F_d(T;\theta_d) - \sum_{i=1}^{n_d} F_d(t_i^d;\theta_d)\right)}$$

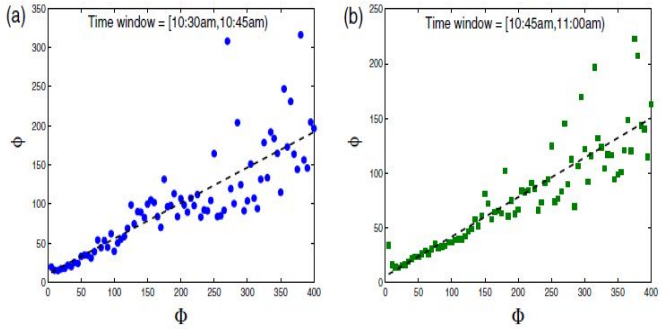
$$p(\lambda_d|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda_d^{\alpha-1} e^{-\beta\lambda_d}.$$

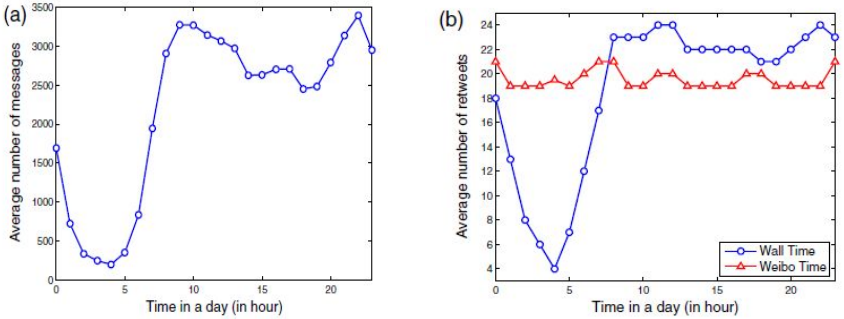
$$\langle c^d(t) \rangle = \int c^d(t) p(\lambda_d|\vec{t^d}, \theta_d, \alpha, \beta) \mathrm{d}\lambda_d$$

4.2 PETM(a reinforced Poisson Process model with Power-law relaxation, Exponential reinforcement and Time Mapping process)

$$r_m(k) = k + 1 + m = \sum_{j=0}^{k} 1 + m$$

$$\hat{\tau} = g(\tau) = \sum_{j=0}^{\tau} M_j / M^*$$





$$r_m(k) = \sum_{j=0}^{j=k} e^{-\alpha_m j} + m$$

$$x_d(t) = \lambda_d t^{-\gamma} (m + \sum_{j=0}^{j=k} e^{-\alpha_m j})$$

4.3. SEHP(Self-Excited Hawkes Process)

now that different retweets can generate different contribution to PA, what about thinking of the triggering effect of each subsequent forwarding.

$$\lambda(t) = ve^{-\beta t} + \alpha \sum_{j=1}^{j_{\max}(t)} e^{-\beta(t-t_j)}$$

$$x_d(t) = \mathrm{m}\lambda_d f_d(t;\theta) + \lambda_d f_d(t;\theta) \sum_{\mathrm{j=0}}^{\mathrm{j=k}} e^{-\alpha_m j}$$

$$if \quad f_d(t;\theta) = e^{-\theta t}, then \quad x_d(t) = \mathrm{m}\lambda_d e^{-\theta t} + \lambda_d \sum_{\mathrm{j=0}}^{\mathrm{j=k}} e^{-\alpha_m j - \theta t}$$

where v is the initial triggering strength, α is the triggering strength of each subsequent forwarding.

4.4. SEISMIC

$$\lambda(t) = p(t) \cdot \sum_{t_i \le t, i \ge 0} n_i \phi(t - t_i)$$

where $p(t)$ is the retweeting probability of time t, n_i is the out-degree of the ith retweeting node, $\phi(t - t_i)$ indicates the probability that the neighbor nodes(nodes which will retweet) of the ith retweeting node retweet the tweet.

what can be done when applying to user-item network?

- I cannot find a dataset containing the complete perchasing data of a certain item.
- the visibility of the sales volumn may generate different PAs.
- serveral ways to modify the model:

1.consider different Network Growth Models

2.according to the special characters of user-item network , modify the model just like the SEHP did.

3.think about the triggering effect.

Thanks

Ruiqi Yang