Journal- and Time- normalization of the Fat-tailed Citation Distributions

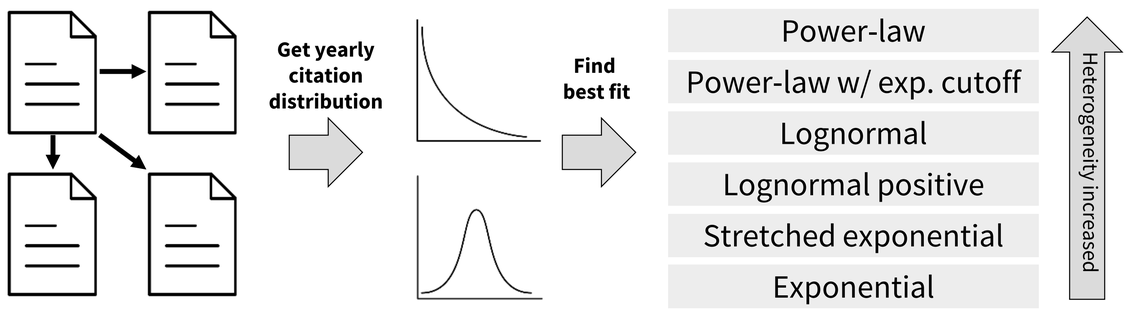
Jinhyuk Yun*, Sejung Ahn , June Young Lee

Future Technology Analysis Center, Korea Institute of Science and Technology Information (KISTI), 66 Hoegiro, Dongdaemun-gu, Seoul, 02456, Korea
*Jinhyuk.yun@kisti.re.kr, http://stat.kaist.ac.kr/~jhyun

Significance

- In this study, we investigate **21 years of annual citation evolution** from 1996 to 2016 of entire citation history of **42 423 644 scientific literatures** in SCOPUS.
- We observe the **raw number** of annual citation acquisitions seems to follow the log-normal distribution for all disciplines, yet **hard to conclude**.
- We also find the **significant disparity of yearly acquired citation numbers** among the journals
- Our simple method that separating citation preference of individual article from the inherited citation of the journals reveals **unexpected regularity on the normalized citation** for the entire fields of science with the probability distributions of power-law with an exponential cut-off
- We also find **strong memory effect** of the citation behaviors, which can be compensated by journal normalization
- Journal normalization outperforms county or discipline normalization
- We design **simple mechanistic model** to reproduce the observations

Motivation Find the best distribution of the citation!



Dataset Scopus and SJR

Scopus®

- April 22, 2017 XML dump
- 42,423,644 scientific literatures
- from 1996 to 2016

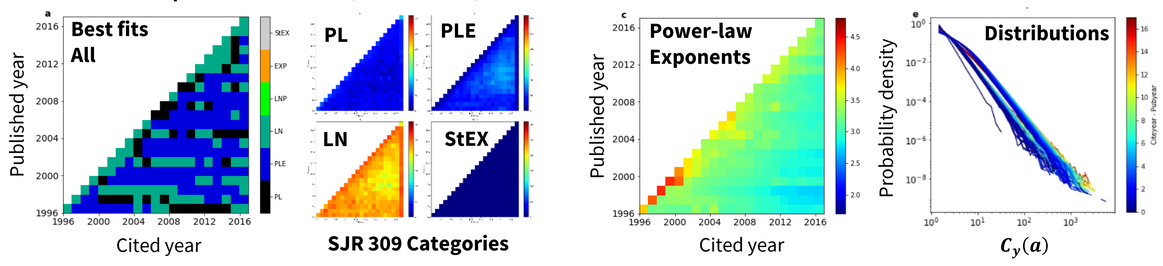
SJR Scimago Journal & Country Rank

- 2017 SJR Classification System
- 27 Subject Areas
- 309 Subject Categories

Best fit Power law or lognormal?

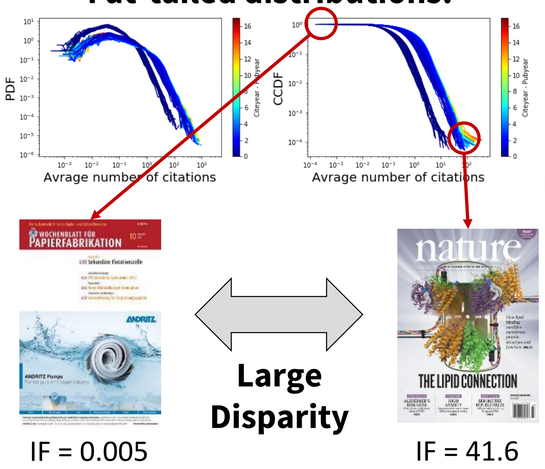
Answer
Mixture of lognormal (LN) and power-laws (PL / PLE)

Exponents
largely varies from ~2.7 to ~4.7



Normalization Remove the journal background

Fat-tailed distributions!



Large Disparity

Normalize

Journal normalized citation:

$$\frac{\text{Citations}}{\text{Mean citation of the journal}}$$

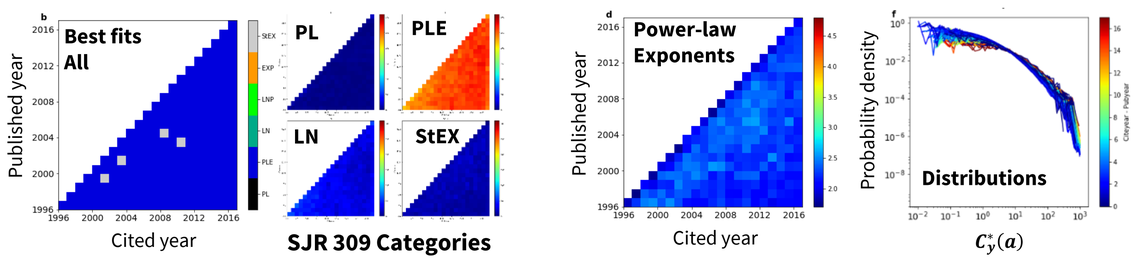
$$C_y^*(a) = \frac{C_y(a)}{\sum_{a \in j(a,y_p)} C_y(a) / N[j(a,y_p)]}$$

- $C_y(a)$**:** The number of citation for an article a of year y.
- $j(a, y_p)$: The set of the articles published in the same journal in the same year of publication (y_p) for an article a.
- $N[j(a, y_p)]$: Total number of publications in the journal

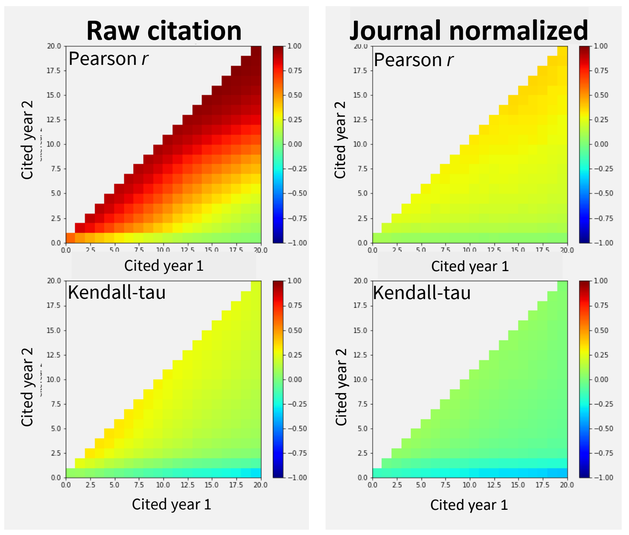
Best fit after normalization! Power-law w/ exp. cutoff!

Best fit
Power law with an exponential cutoff for all 21 years!

Exponents ~2.3



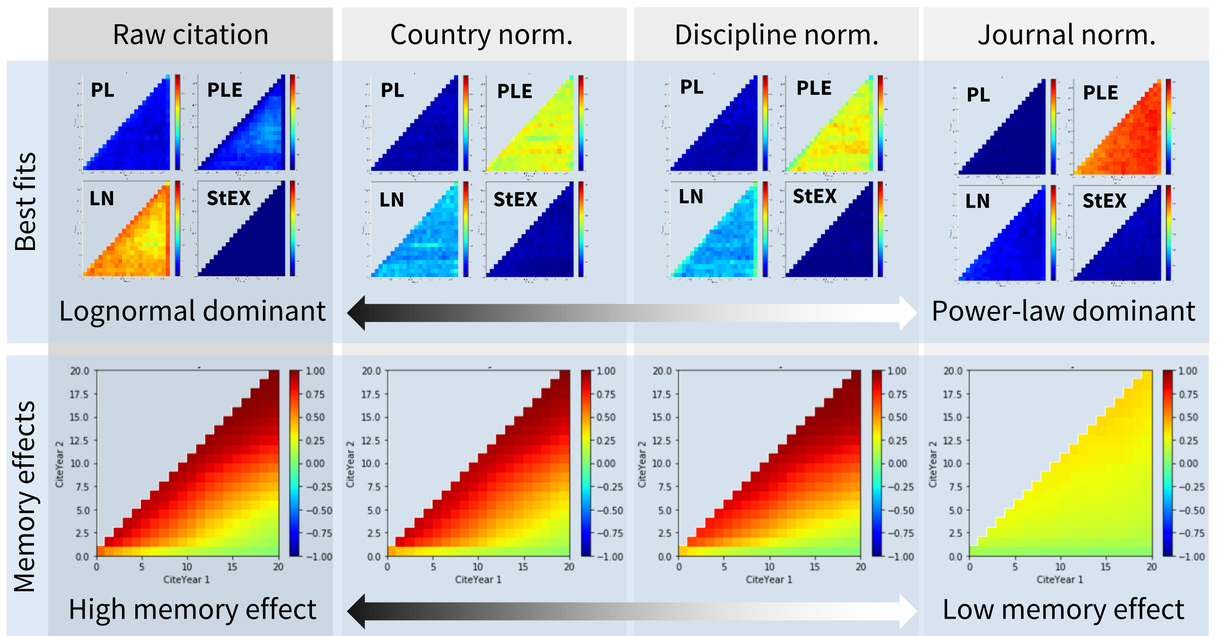
Temporal correlations or memory effect?



Raw count
Strong memory effect

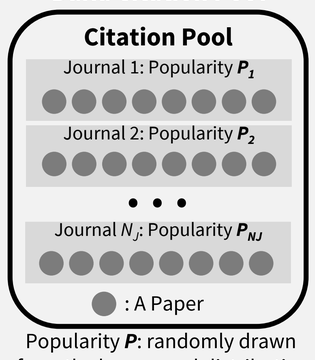
Journal normalized count
weak (or no) memory effect

Other normalizations? country and discipline



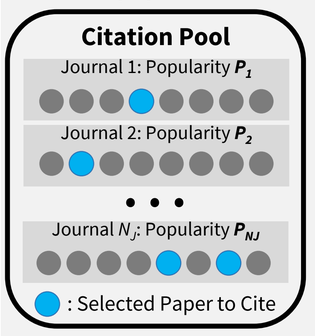
Mechanistic model introduce the journal popularity

1. Initialization
Build Citation Pool



Popularity P: randomly drawn from the log-normal distribution

2. Update rule (for every timestep)
A new paper cites k articles in the citation pool with probability of

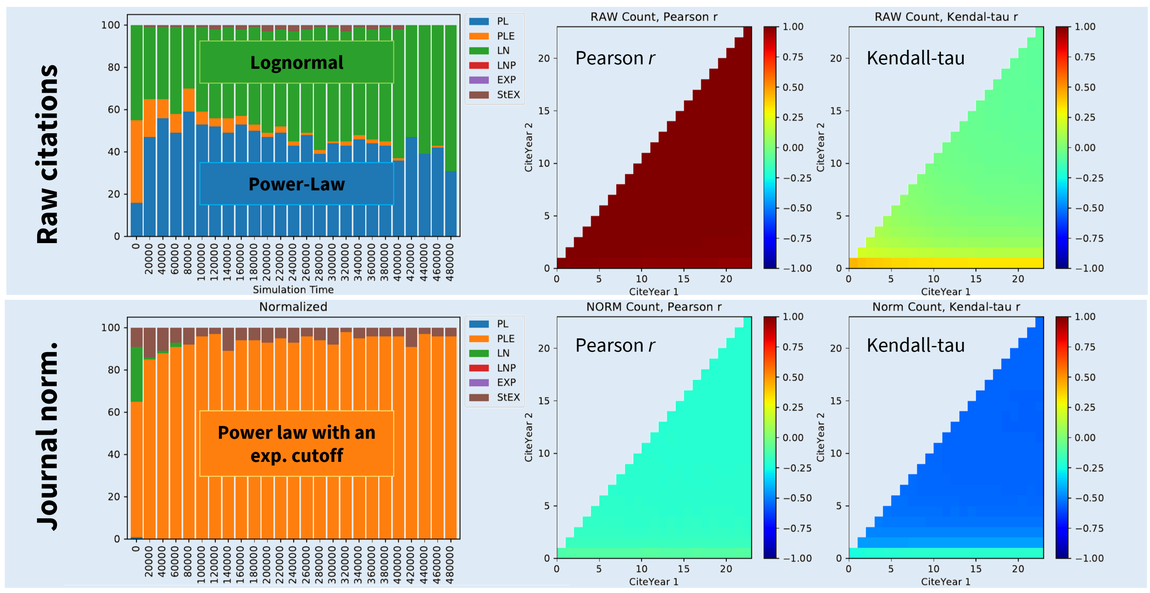


$$Prob(i) \propto P(i) \times C^*(i;t)$$

$P(i)$: Popularity of Journal paper i belonging
$C^*(i;t)$: Sum of citation of paper i with decaying

Here, $C^*(i;t) = 1 + \sum_{c \in C(i)} \frac{\exp\left(-\frac{a(c;t)}{T}\right)}{k}$

- $C(i)$: Set of citations that paper i received since the onset
- $a(c)$: Age of the citation c at time t
- T**:** Temporal decaying rate for the citation history
- k**:** Number of reference per a new article



Basic parameters Number of papers = 20000; Number of journals = 500; Ensembles = 100;
Popularity distribution parameters Mean : $\mu = 0.0$; Standard deviation: $\sigma = 1.6$;
Ageing parameters Decay parameter $T = 3 \times 10^4$

References

[1] SCOPUS CUSTOM DATA, https://www.scopus.com.
[2] SCImago Journal & Country Rank, http://www.scimagojr.com.
[3] Anna D. Broido, Aaron Clauset, Scale-free networks are rare, *arXiv:1801.03400.*
[4] Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman, Power-Law Distributions in Empirical Data, *SIAM Rev., 51(4).*
[5] Alexander Michael Petersen et al., Reputation and impact in academic careers, *PNAS, 111(43).*

Acknowledgement: This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government through Grant No. NRF-2017R1E1A1A03070975 (JY; SA) and Korea Institute of Science and Technology Information(KISTI)