Journal- and Time-normalization of Fat-tailed Citations Distributions

Yun, J.; Ahn, S.; Lee, J.Y.

STI 2018 Conference Proceedings

Proceedings of the 23rd International Conference on Science and Technology Indicators

All papers published in this conference proceedings have been peer reviewed through a peer review process administered by the proceedings Editors. Reviews were conducted by expert referees to the professional and scientific standards expected of a conference proceedings.

Chair of the Conference

Paul Wouters

Scientific Editors

Rodrigo Costas
Thomas Franssen
Alfredo Yegros-Yegros

Layout

Andrea Reyes Elizondo
Suze van der Luijt-Jansen

The articles of this collection can be accessed at https://hdl.handle.net/1887/64521

ISBN: 978-90-9031204-0

23rd International Conference on Science and Technology Indicators (STI 2018)
"Science, Technology and Innovation indicators in transition"
12 - 14 September 2018 | Leiden, The Netherlands
#STI18LDN

Journal- and Time-normalization of Fat-tailed Citations Distributions

Jinhyuk Yun*, Sejung Ahn*, June Young Lee*
* *jinhyuk.yun@kisti.re.kr; sjahn@kisti.re.kr; road2you@kisti.re.kr*

Department of Scientometric Research, Korea Institute of Science and Technology Information, 66 Hoegiro, Dongdaemun-gu, Seoul, 02456 (Korea)

Introduction

For decades, it has been observed that citation of scientific literature follows a heterogeneous and fat-tailed distribution. Unfortunately, identifying the distribution is challenging, because many citation distributions are characterized as a long-tail with rare events; it thus essentially accompanying large fluctuations on observed distributions (Clauset et al., 2009). Although observed data behaves like a particular model distribution, it is hard to deny the possibility of the alternative distributions. Indeed, scholars proposed several kinds of model distributions for the citation. One candidate was power-law and its siblings (Redner, 1998; Price, 1976; Brzezinski, 2015). Exponential and stretched-exponential were also reported (Wallace et al., 2009). Moreover, recent studies indicated citation distribution as the (discretized) log-normal (Thelwall & Wilson, 2014; Thelwall, 2016). Unfortunately, many studies are limited to small-scale approaches; it is thus hard to generalize. Tackling this issue, we investigate 21 years of citation evolution through systematic analysis entire citation history of entrie 42,423,644 scientific literature published from 1996 to 2016 in SCOPUS. We also suggest a new normalization method which may reduce the imbalance derived from the fame of the journals.

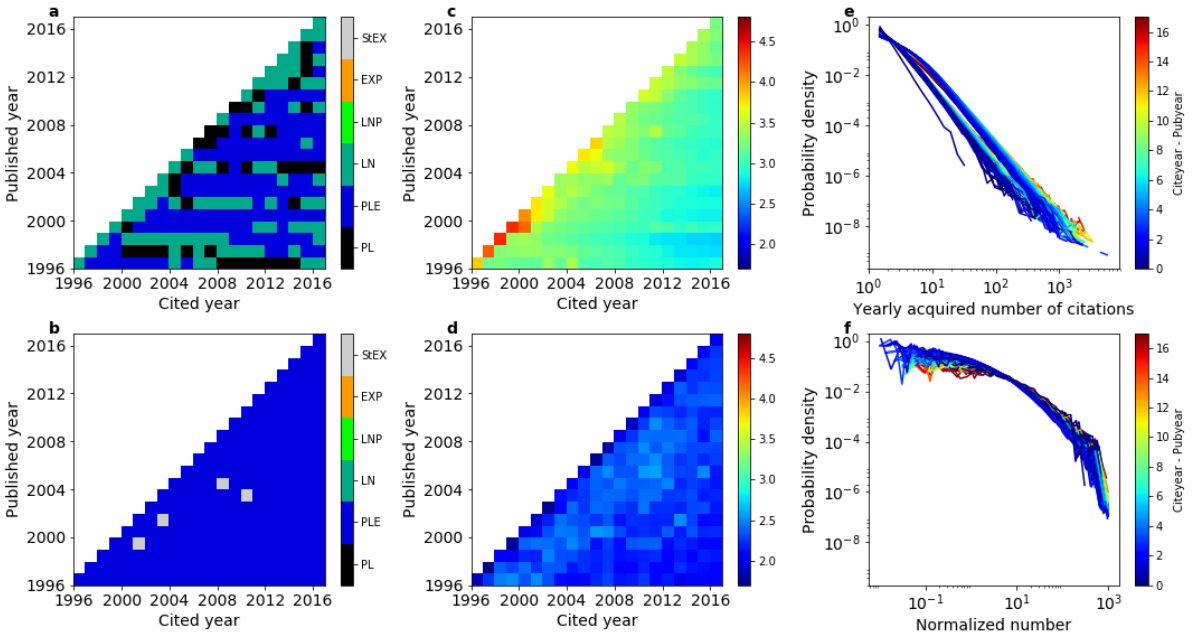
Best distributions of the citation.

We begin investigating the empirical evidence for best fit model distributions by the Maximum Likelihood Ratio (MLR) methods (Clauset et al., 2009). We tested six candidate distributions in a various level of heterogeneities: (i)&(ii) power-law (with an exponential cut-off), (iii)&(iv) log-normal (positive), (v)&(vi) (stretched) exponential; and we consider a certain distribution is most suitable model if the Maximum Likelihood Estimator is superior to all other distributions. We use the yearly acquired citations instead of the long-term accumulated citation count. The measure is defined as the number of citation of a certain article acquired in a certain year, implying the level of attention for single academic literature in a certain year.

First, we apply MRL methods for all six model distributions, for each year between 1996 and 2016. Unexpectedly, what we observe is the mixture of three distributions, instead of the single dominant model (see Figure 1a). Specifically, we observe the mixture of log-normal (LN), power-law with an exponential cut-off (PLE), and power-law (PL) as the best fit, whereas the other three distributions are not suspected. The estimated power-law exponent is from ~2.7 to ~4.7 (Figure 1c) and $x_{\min}$ is ranged from 11 to 108 ($\langle x_{\min} \rangle$ ~39.82). This result is also supported

by the visual demonstration of probability density showing widespread lines (Figure 1e). It is thus hard to conclude the existence of the universal distribution across the years.

Figure 1: The raw number of citation and its Journal- and Time- Normalized measure of yearly acquired citation count for the articles in SCOPUS from 1996 to 2016. (a) We observe the mixture of log-normal (LN), power-law with an exponential cut-off (PLE), and basic power-law (PL) as the best fit of the pdf distribution for the raw count; (b) meanwhile power-law with exponential cut-off (PLE) dominates normalized measures. (c) The estimated power-law exponent of raw citation distribution ranges extensively from ~2.7 to ~4.7, yet it is congregated around ~2.3 for the normalized measure (d). This result is also supported by the visual demonstration of probability density [compare (e) for the raw count and (f) for the normalized measures].



Journal- and Time- normalized citation score

One should note that the mean citation per paper was also varied largely by the journal (Seglen, 1997). This heterogeneity implies the existence of inherited citation from the reputation of the journal; it thus makes hard to compare the citation counts of articles from different journals. Also, the preference of citation consistently decreased due to the aging effect (Eom & Fortunato, 2011; Hajra & Sen, 2005). We propose the rescaled measures of citation compensating those effects that lead to very similar patterns across years as follows:

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

where $C_y(a)$ is the citation count of article a in the cited year y, and $j(a, y_p)$ is the set of articles published in the same journal and published year (y_p) of the article a.

Unlike the raw citation, we show the single distribution, namely power-law with an exponential cutoff, dominates with our rescaled citation measure. This observation is stable for entire publication and citation year (Figure 1**b**). The estimated power-law exponent of $C_y^*(a)$

for those distributions congregate around ~2.3 (Figure 1**d**) and x_{min} is ranged from 15.46 to 97 ($<x_{min}>$ ~47.91). This finding also visually supported by probability density itself (Figure 1**f**).

Discussion

In this poster, we explored the structure of academic citation through a massive history of citation metadata over the past two decades highlighting the influence of the journals' prestige. We suggest that in-depth analysis of factors that influence the number of citations, e.g. impact of countries, authors, institutes, and disciplines, may be promised to enhance the impact of our approach (Albarrán *et al.*, 2011). Also, extending our analysis into cumulative citation count is left for further study due to the computational complexity. Going one step forward, if data-driven analysis accompanied by proper normalization is judiciously combined with the citation analysis, the synergy will bring for the methodologies in the social sciences, humanities, and policy-making.

References

Albarrán, P., Crespo J., Ortuño I., and Ruiz-Castillo J. (2011). The skewness of science in 219 sub-fields and a number of aggregates. *Scientometrics*, 88(2), 385–397.

Brzezinski M. (2015) Power laws in citation distributions: evidence from scopus. *Scientometrics*, 103(1), 213–228.

Clauset A., Shalizi C.R., Newman M.E. (2009). Power-law distributions in empirical data. *SIAM review*, 51(4), 661–703.

Eom Y.H., Fortunato S. (2011). Characterizing and modeling citation dynamics. *PloS one*, 6(9), e24926.

Hajra K.B., Sen P. (2005). Aging in citation networks. *Physica A: Statistical Mechanics and its Applications*, 346, 44–48.

Price D.d.S. (1976) A general theory of bibliometric and other cumulative advantage processes. *Journal of the Association for Information Science and Technology*, 27(5), 292–306.

Redner S. (1998). How popular is your paper? an empirical study of the citation distribution. *The European Physical Journal B-Condensed Matter and Complex Systems*, 4(2), 131–134. 9.

Redner S. (2005). Citation statistics from 110 years of physical review. *Physics today*, 58(6), 49–54.

Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. *BMJ: British Medical Journal*, *314*(7079), 498.

Thelwall M., Wilson P. (2014), Distributions for cited articles from individual subjects and years. *Journal of Informetrics*, 8(4), 824–839.

Thelwall M. (2016). Citation count distributions for large monodisciplinary journals. *Journal of Informetrics*, 10(3), 863–874.

Wallace M.L., Lariviére V., Gingras Y. (2009). Modeling a century of citation distributions. *Journal of Informetrics*, 3(4) 296–303.