Physica A 391 (2012) 2129–2134

Contents lists available at SciVerse ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

# A scaling between Impact Factor and uncitedness

Jiann-wien Hsu [a], Ding-wei Huang [b,*]

[a] General Education Center, National Tainan Institute of Nursing, Tainan, Taiwan

[b] Department of Physics, Chung Yuan Christian University, Chung Li, Taiwan

ARTICLE INFO

*Article history:*
Received 14 April 2010
Received in revised form 9 November 2011
Available online 25 November 2011

*Keywords:*
Citation dynamics
Scaling
Impact Factor

ABSTRACT

The Impact Factor has become a well-known measure of the average citation number of articles published in a scientific journal. A journal with a high Impact Factor is assumed to have a low percentage of uncited articles. We show that the scaling relation between the Impact Factor and the uncited percentage can be understood by a simple mechanism. The empirical data can be reproduced by a random mechanism with the cumulative advantage. To further explore the robustness of such a mechanism, we investigate the relation between the average citation number and the uncited percentage from different perspectives. We apply the idea of Impact Factor to the publications of an institute in addition to its general application to the publications of a journal. We find that the same scaling relation can be obtained. We also show that a static relation can be applied to describe the time evolution of a dynamical process. These results provide further justification for the same citation mechanism behind different research fields.



## 1. Introduction

With the flourishing of scientific publications, citation dynamics has become an interesting topic of statistical physics. The rapid accumulation of scientific knowledge has made researchers more aware of where to best publish their research results. The Impact Factor (IF) has been a well-known measure of the citation impact of a science (or social science) journal [1]. Basically, the IF can be taken as the average number of citations to an article recently published in that journal. Specifically, the IF of a journal in a given year is the average number of citations to those articles that were published during the two preceding years. With a wide variety of scientific publications, a wide range of IFs can be expected. Although it is often taken as a direct measure of the quality of a journal, an IF also reflects the popularity of a subject. For example, a journal on biosciences often has a higher IF than a journal on mathematics. Although each journal has its own specialty, and the rankings of journals are quite stable over the years [2,3], there are still wide fluctuations in citation numbers for different articles published in the same journal [4]. Since citation counts are not evenly distributed, an IF calculated by the arithmetic mean is not sufficiently representative of an individual article. The use of a journal's IF to evaluate the citation impact of an article within that journal can be controversial [5,6].

With the advance of statistical physics in the so-called field of sociophysics, various topics have been investigated to reveal the underlying mechanism in citation dynamics. The issue of uncitedness has been raised recently [7]. Intuitively, a journal with a high IF should have a low percentage of uncited articles. An interesting scaling relation in citation dynamics has been revealed [8]. When the average citation number is plotted against the uncited percentage, data from different journals collapse onto a single curve. Obviously the average citation number decreases with the increase of the uncited percentage. Yet the curve changes from a convex decrease to a concave decrease. There have been many discussions on the

* Corresponding author.
*E-mail address:* dwhuang@phys.cycu.edu.tw (D.-w. Huang).

0378-4371/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2011.11.028

mathematical formulation of such a peculiar shape [9–11]. To our knowledge, the underlying mechanism remains unexplored. In statistical physics, scaling of different datasets often indicates a general mechanism behind the process [12,13]. In this work, we show that the scaling curve can be understood by a simple mechanism. The scaling relation implies that the same mechanism works behind a wide variety of citation practices in scientific publications. Such a scaling curve can be used to differentiate among different mechanisms in citation dynamics.

To further explore the robust mechanism of citation dynamics, we study the relation between the average citation number and the uncited percentage from different perspectives. Originally, the idea of IF was applied to the publications of a journal. In this study, we apply the idea of IF to the publications of an institute. In contrast to the uniformity among the many articles collected in a journal, the publication list of an institute consists of many articles from a wide range of scientific topics. Yet we find that the same scaling relation can still be discerned. We also observe that the static relation can be applied to describe the temporal evolution in the dynamical process of citation accumulation. In previous studies, the scaling relation has been portrayed as a static relation for many journals. As in the many cases of equilibrium statistical mechanics, a dynamic relation might become a static relation in the ensemble. These observations can provide further justification for the same mechanism behind different research fields.

## 2. Model

We show that the relation between the average citation number and the uncited percentage can be understood by simple mechanisms in citation dynamics. Consider an ensemble of $N$ articles published within a certain period of time. After publication, these articles begin to accumulate citations from later-published articles. In practice, citation dynamics can be very far from a process of random drawing within the ensemble. As there are too many factors to be included in a detailed theory, however, the random process can be a justified starting point for the citation dynamics. The resultant citation distribution is then a binomial distribution, where different articles receive more or less the same number of citations. With an average citation number $a$ for each article, these $N$ articles have been cited $(Na)$ times in total. The distribution for an article to be cited $n$ times can be written as

$$P(n) = C_n^{Na}\left(\frac{1}{N}\right)^n\left(1-\frac{1}{N}\right)^{Na-n}, \tag{1}$$

where $P$ denotes the probability, and the $N$ articles are treated equally. At the uncited percentage $u$, there are $(Nu)$ articles that remain uncited. An analytic expression can be derived as

$$u = P(0) = \left(1-\frac{1}{N}\right)^{Na} \sim \mathrm{e}^{-a}, \tag{2}$$

where $N \gg 1$. The random mechanism of citation dynamics prescribes an exponential relation between $a$ and $u$. Compared to empirical data, the uncited percentage $u$ at a fixed average citation number $a$ has been underestimated. Such a feature implies that the citations are much more concentrated than the process of random selection. A simple modification to the random mechanism is the cumulative advantage. The same mechanism is known in sociology as the Matthew effect or "*the rich get richer*" effect [14,15]. In the context of citation dynamics, the articles being cited more in a previous stage will attract more citations in the next stage. As a result, most of the citations will be concentrated on a few popular articles. The asymptotic distribution will be changed from a binomial distribution to an exponential distribution, where there is no peak around the average value. Mathematically, the static distribution for the probability density $P(n)$ can be obtained by solving the following equation:

$$\frac{\mathrm{d}P}{\mathrm{d}n} \propto P, \tag{3}$$

where the time dependence has been removed from the asymptotic limit. With proper normalization, we obtain

$$P(n) = \frac{a^n}{(1+a)^{n+1}} = \frac{1}{1+a}\exp(-bn), \tag{4}$$

where $b = \ln[(1+a)/a]$. Thus, the citation distribution $P(n)$ decreases monotonically as $n$ increases, and the highest weight is shifted to uncitedness. An algebraic relation between $a$ and $u$ can be written as

$$u = P(0) = \frac{1}{1+a}. \tag{5}$$

The empirical data can be described well. The difference between the random-selection mechanism and the cumulative-advantage mechanism lies in the large citations $a \gg 1$ (and thus the small uncitedness $u \ll 1$). With small citations $a \ll 1$ (and thus a large uncitedness $u \sim 1$), these two mechanisms provide the same description of $u \sim (1-a)$.

We note that the empirical distributions of citations cannot be described fairly by either of these two mechanisms alone. In practice, various models have been proposed recently to explain citation distributions by combined mechanisms [16,17].