Journal of Informetrics 8 (2014) 493–502

Contents lists available at ScienceDirect

Journal of Informetrics

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

A study of the "heartbeat spectra" for "sleeping beauties"

Jiang Li[a], Dongbo Shi[b], Star X. Zhao[a], Fred Y. Ye[c,*]

[a] Department of Information Resource Management, Zhejiang University, Hangzhou 310027, China
[b] School of Public Policy and Management, Tsinghua University, Beijing 100084, China
[c] School of Information Management, Nanjing University, Nanjing 210093, China

ARTICLE INFO

Article history:
Received 23 December 2013
Received in revised form 2 April 2014
Accepted 2 April 2014

Keywords:
Sleeping beauty
Heartbeat spectrum
Citation pattern
G_s index
Gini coefficient

ABSTRACT

We first introduced interesting definitions of "heartbeat" and "heartbeat spectrum" for "sleeping beauties", based on van Raan's variables. Then, we investigated 58,963 papers of Nobel laureates during 1900–2000 and found 758 sleeping beauties. By proposing and using G_s index, an adjustment of Gini coefficient, to measure the inequality of "heartbeat spectrum", we observed that publications which possess "late heartbeats" (most citations were received in the second half of sleeping period) have higher awakening probability than those have "early heartbeats" (most citations were received in the first half of sleeping period). The awakening probability appears the highest if an article's G_s index exists in the interval [0.2, 0.6).

1. Introduction

Mendel (1866) paper had been a classical example of citation phenomenon in science where publications did not achieve recognition until some years after their original publication (Zirkle, 1964). These publications are referred to as "premature discoveries" (Wyatt, 1961), "resisted discoveries" (Barber, 1961), "delayed recognition" (Cole, 1970), and recently "sleeping beauties" (van Raan, 2004).

The name of "sleeping beauty" came from a well-known fairy tale, and brought interesting image to informetrics. A sleeping beauty in science is a princess (an article) which sleeps (goes unnoticed) for a long time and then, almost suddenly, is awakened (receives a lot of citations) by a prince (another article). It is fairly common to find sleeping publications which received few citations in a period after publication, but only a small fraction was awakened and became sleeping beauties. In this research, we defined "heartbeat spectra" of sleeping publications, and investigated what kind of heartbeat spectra produced the most sleeping beauties.

2. Literature review

The prematurity of or resistance to scientific discoveries appeared, when they were not consistent with the accepted knowledge at the time or not verifiable technologically. These publications were referred to as "premature discoveries" (Wyatt, 1961) or "resisted discoveries" (Barber, 1961). The two terminologies have been dominated by "delayed recognition" (Cole, 1970) since the 1970s. In essence, they all depict slow obsolescence of publications. Delayed recognition publications are initially unappreciated or unused but are later recognized as significant, according to "diachronous"

* Corresponding author. Tel.: +86 25 83592959; fax: +86 25 83592959.
E-mail addresses: yye@nju.edu.cn, blueyye@gmail.com (F.Y. Ye).

http://dx.doi.org/10.1016/j.joi.2014.04.002
1751-1577/

(Line & Sandison, 1974; Nakamoto, 1988) or "retrospective" (Glänzel, 2004) measurement of obsolescence. They often have high quality ideas and methods (McCain & Turner, 1989; Levitt & Thelwall, 2009). In their citation records, there is often a sudden of citations at a point in time well beyond a typical paper for that field. The citation curve of a typical paper appears "lognormal" shape, which rises to a citations-peak in a few years after publication and then is gradually less cited with time (Cunningham, 1995; Egghe & Rao, 1992).

Garfield (1980) proposed that parameters should be set for what truly qualifies as delayed recognition, although he called for examples of delayed recognition from some research fields (Garfield, 1989a, 1990). The criteria that Garfield (1989b) set are as follows: (1) highly cited papers that had low citation frequencies for the first 5 or more years, with more than 10 years being preferred, and (2) low initial citation frequency was defined as being near the average of one cite per year for a typical paper. As a result, he found five examples from 1800 papers. Glänzel, Schlemmer, and Thijs (2003) considered a paper published in 1980 having delayed reception, if it has received (a) only one citation in an initial 3-year period or (b) at most two citations in an initial 5-year period and it is highly cited if it has received at least 100 citations in the remaining period till 2000. They found 77 papers out of the almost 450,000 publications under the weak condition (a) and 29 papers under the stronger condition (b). After revising the "received at least 100 citations" into "received at least 50 citations and 10 times the journal impact", the selection resulted in a set of 60 (weak condition) and 16 papers (strong condition), respectively. The 3- or 5-year citation window was defined by the fact that in general more than 80% are cited in an initial 3-year window and more than 90% in an initial 5-year citation window in terms of first-citation statistics (Glänzel et al., 2003). Later, delayed recognition papers were defined (Glänzel & Garfield, 2004) as those which, during a period of five years, were initially rarely cited but then became highly cited (at least 50 citations or 10 times the journal's 20-year cumulative impact factor) during the next 15 years. Following these criterions, van Raan (2004) termed delayed recognition papers "sleeping beauties" and suggested three variables for such papers: (1) depth of sleep (C_s), they receive at most 1 citation per year on average (deep sleep), or between 1 and 2 citations per year on average (less deep sleep) for a few years after publication; (2) length of sleep (s), i.e., duration of the sleeping period; and (3) awakening intensity (C_w), number of citations per year, during four years following the sleeping period. In addition, he derived a general Grand Sleeping Beauty Equation: $N = f\{s, c_s, c_w\} \sim s^{-2.7}\ {c_s}^{2.5}\ {c_w}^{-6.6}$, where N is the number of sleeping beauties.

Then the understanding of sleeping beauties has been extended. The three variables enable automatically search for sleeping beauties from citation databases (Braun, Glänzel, & Schubert, 2010; Lange, 2005; Ohba & Nakao, 2012). Moderately aroused sleeping beauties might very well be expected (Burrell, 2005). Li and Ye (2012) found four special sleeping beauties in *Nature* which had leaping before sleeping in citations, and named them "all-elements-sleeping-beauties". Braun et al. (2010) proposed that a candidate prince should be among the first citing articles which are highly cited and have a number of co-citations with the sleeping beauty. Li (2014) suggested in a recent study that an "all-elements-sleeping-beauty" should include an awaking period (leaping), a sleeping period, an awakening period and a happy ending (the princess and the prince received high co-citaitons). van Clester (2012) provided an extreme example of a sleeping beauty, i.e., Peirce (1884) note in *Science* was rarely cited until 2000. This example revealed a limitation of the modalities of sleeping beauties: the beginning year of the awakening period is ambiguous. The note received 21 citations during 2006–2009, prior to which, it received less than 1 citation per year. The two periods of the note qualify for a sleeping beauty. However, it received less than 2 citations per year in the whole period till 2012, which indicates the note has not been awakened. The reason for the ambiguity is that the quantitative definitions used averages.

Using averages in bibliometrics is criticized (Glänzel, 2008). Costas, van Leeuwen, and van Raan (2010) proposed using quantiles as an alternative to determine delayed recognition publications. First of all, they identified the year after publication in which the document received for the first time at least 50% of its citations ("Year 50%"). Then, they calculated, for all documents of the same year of publication, the quantiles 25 and 75 of the distribution function of the value of "Year 50%", and recorded them as "$P25$" and "$P75$". At last, the general criterion for the classification of documents in a specific field was as follows: (1) flashes in the pan: "Year 50%" $< P25$; (2) delayed documents: "Year 50%" $> P75$; and (3) normal documents: $P25 \leq$ "Year 50%" $\leq P75$. They observed that the percentages of the three types of durability were 9.4%, 20.2% and 70.4%, respectively, in a dataset of 8,162,537 publications. Using quantiles is a relative method. Hence, it is difficult to identify individual delayed recognition paper without calculating the citations of its whole field. Furthermore, the status determined by quantiles is variable. For example, a flash in the pan can evolve into delayed recognition if the article suddenly receives massive citations in the future.

Citation patterns have been summarized from the citation history of papers. Price (1965) observed that 25% of the papers were cited at a constant rate without declining over the years, 25% gradually increased in citedness and then declined at a similar rate, and 50% were cited at a constant rate for several years. Based on Price's findings, Aversa (1985) proposed two citation patterns: "early rise, rapid decline" and "delayed rise, no decline". Similarly, Lange (2005) termed "hits" for works noticed by the scientific community soon after their publication, and "missed signals" for works that went unnoticed until much later, which were also named "shooting stars" and "sleeping beauties" (Mingers, 2007), respectively. Aksnes (2003) supplemented the third citation pattern: "medium rise-slow decline", to Aversa's patterns. van Dalen and Henkens (2005) categorized four citation patterns based on their citations: early ("flash in the pan"), late ("sleeping beauty"), little and many. Costas et al. (2010) proposed a general "technical" definition of different types of durability of documents regardless of publication year or total number of citations: "flashes in the pan", "delayed" and "normal" documents, corresponding to Aversa's and Aksnes's "early rise, rapid decline", "delayed rise, no decline" and "medium rise-slow decline", respectively.