
ELSEVIER
Journal of
INFORMETRICS
An International Journal

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.

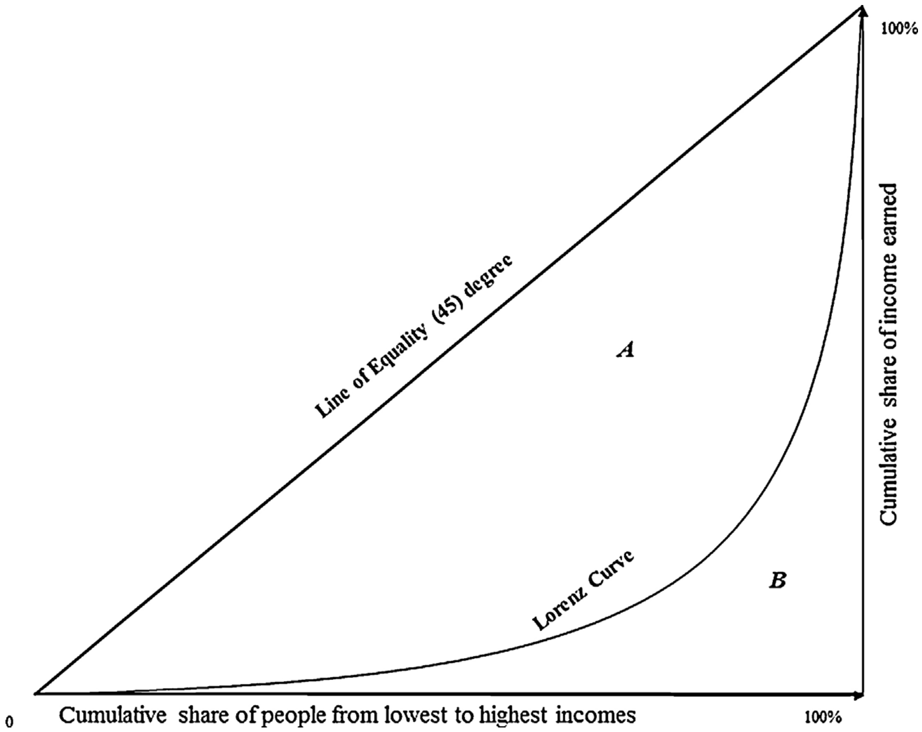

Fig. 1. Graphical representation of the Gini coefficient.

The Science Citation Index (SCI) plays a crucial role in searching publications with delayed recognition or sleeping beauties. It also promoted the visibility of scientific research, which reduced negligence of previous research to certain degree. That is why Garfield (1970) claimed that Mendel's work would not have been neglected if there had been a citation index at that time. The reasons for delayed recognition were concluded as: (1) concepts and theories in the publications were ahead of their time and (2) the hierarchical system of science caused the older resist the younger (Barber, 1961; Cole, 1970; Hook, 2002; Price, 1976; Stent, 1972).

3. Methodology

3.1. Definitions

We consider a publication "sleeping" if it received at most 2 citations on average per year in at least five years since publication. This period is named "sleeping period". We consider the publication "awakened" if it received more than 20 citations in the four years following the sleeping period. Hence, we have the following definitions.

Definition 1: The "heartbeat" of a sleeping beauty is the annual citation(s) it received in the sleeping period. Let $c_i \geqq 0$ denote the number of citation(s) it received in the ith year in the sleeping period, then the sleeping beauty's heartbeat in the ith year is c_i.
Definition 2: A "heartbeat spectrum" is a vector of a sleeping beauty's heartbeat, i.e., $H = (c_1, \ldots, c_i \ldots, c_n)$, where n indicates the duration of sleeping period.
Definition 3: The "length of heartbeat spectrum" is the duration of sleeping period, i.e., n in the vector H. Given t_1 (publication year) and t_n for the beginning and ending year of sleeping period, we have $n = t_n - t_1 + 1 \geqq 5$.
Definition 4: The "strength of heartbeat spectrum" is the average citations in the sleeping period, i.e., $C_s = \sum_1^n c_i/n = C/n$, where C is the number of citations received in the sleeping period, $0 \leqq C_s < 1$ indicates deep sleep and $1 < C_s \leqq 2$ indicates less deep sleep.

We consider a paper "unawakened" if it has a sleeping period but has not been awakened. Unawakened publications are also investigated in order to contrast sleeping beauties.

3.2. Gini coefficient

Gini coefficient (Gini, 1912), as a measure of statistical dispersion, results from a concept to measure inequality of income. It is applied to a collection whose elements are arranged in non-decreasing order. This leads to a convex Lorenz curve between (0, 0) and (1, 1), as shown in Fig. 1, which depicts the proportion of the total income of the population (y axis) that

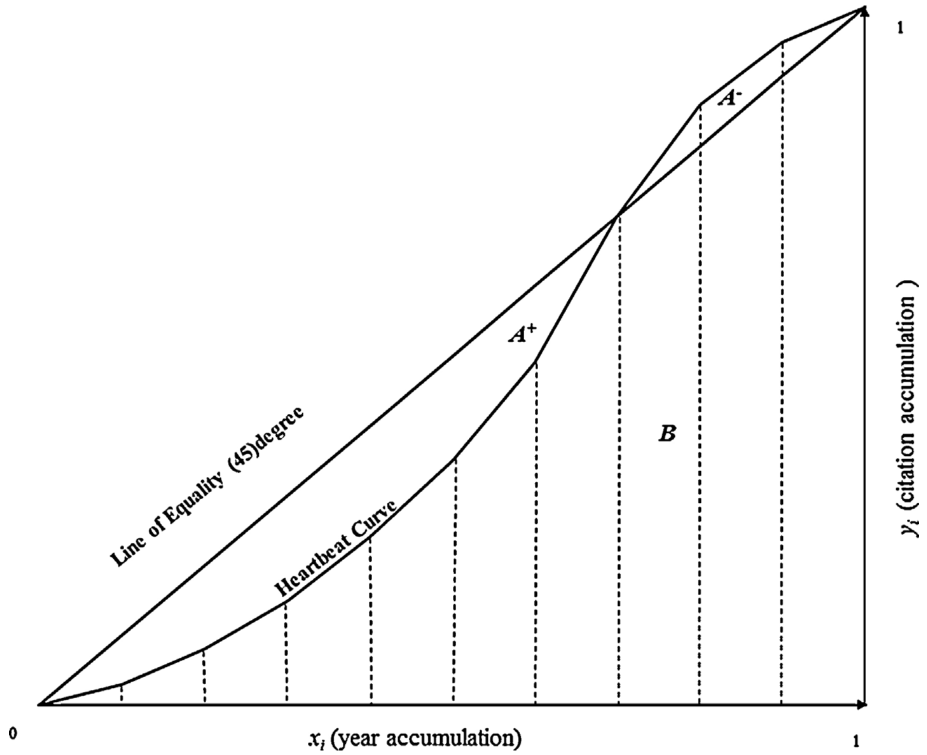

Fig. 2. Graphical representation of the G_s index for measuring the inequality of citation distribution in sleeping period.

is cumulatively earned by the bottom x of the population. For example, in the bottom 80% of population possesses 20% of the income of the total population. The line of equality (45°) represents perfect equality of incomes. Then, in a graphical representation of Gini coefficient, it is the ratio of the area that lies between the line of equality and the Lorenz curve (marked A in Fig. 1) over the total area under the line of equality (marked A and B in Fig. 1), i.e.,

$$G = \frac{A}{A+B} \tag{1}$$

Gini coefficient must lie between 0 and 1, in terms of Eq. (1) and the graphical representation in Fig. 1. Lower Gini coefficient means more equal distribution, with 0 corresponding to complete equality, whereas higher Gini coefficient indicates more unequal distribution, with 1 corresponding to complete inequality.

In bibliometrics, Pratt (1977) proposed an index of concentration for rank-frequency distribution (Pratt's measure), in order to compare subject and journal concentration. The subject field is divided into n categories by any convenient classification and each paper is assigned to one and only one categories. The number of papers in each category is then counted and ranked in decreasing frequency of assignment. However, Carpenter (1979) argued that this measure is nearly identical to the Gini coefficient. Egghe and Rousseau (1990) conducted a comprehensive review of concentration measures, including Pratt's measure and Gini coefficient. They also presented a set of principles that good concentration measures must satisfy, such as "strictly Schur-convex" and "scale invariant".

3.3. *G_s index*

We propose G_s index, an adjustment of the Gini coefficient, for measuring the inequality of heartbeat spectrum. The non-decreasing order of the elements arranged for calculating Gini coefficient of income, is adjusted to a natural time-order of citations for calculating G_s index. It is then feasible to differentiate "early heartbeats" from "late heartbeats" by arranging citations in a natural time-order rather than a non-decreasing order. The cumulative curve hence changes from a Lorenz curve to (we call it) a "heartbeat curve", as shown in Fig. 2. The curve shows what percentage of sleeping period possesses what percentage of the period's citations. For example, in the first 10% of the sleeping period possesses 5% of the period's citations. Accordingly, G_s index measures an overall difference between the heartbeat curve and the uniform distribution curve, (the line of equality), i.e.,

$$G_s = \frac{A}{A+B} \tag{2}$$

As a result, G_s index does not match Egghe and Rousseau's principles mainly because the heartbeat curve is not necessarily convex. In Fig. 2, the horizontal axis is the proportion of year accumulation:

$$x_i = \frac{i}{n}, \tag{3}$$

where i and n retains the meaning in the above definitions. The vertical axis is the proportion of citation accumulation:

$$y_i = \frac{\sum_1^i c_i}{\sum_1^n c_i} = \frac{\sum_1^i c_i}{C} \tag{4}$$

since $x_i, y_i \in (0,1)$ form an isosceles triangle in Fig. 2, so we have

$$A + B = \frac{1}{2}. \tag{5}$$

The area under the cumulative curve approximately equals to the sum of the area of the n trapeziums marked by the dotted lines, so we have

$$G_S = 1 - \frac{\sum_1^n ((1/2) \times (y_i + y_{i-1}) \times (1/n))}{1/2}, \tag{6}$$

where $y_0 = 0$. Then we have

$$G_S = 1 - \frac{2 \times \sum_1^n y_i - y_n}{n}. \tag{7}$$

when $C > 0$, we have $y_n = 1$ and

$$G_S = 1 - \frac{2 \times \sum_1^n y_i - 1}{n}, \tag{8}$$

putting Eq. (4) into Eq. (8), we have

$$G_S = 1 - \frac{2 \times [n \times c_1 + (n-1) \times c_2 + \cdots + c_n] - C}{C \times n} \tag{9}$$

when $C = 0$, we have $y_i = 0$ and

$$G_S = 1 \tag{10}$$

therefore, the segmented function of G_s index is

$$G_S = \begin{cases} 1 - \dfrac{2 \times [n \times c_1 + (n-1) \times c_2 + \cdots + c_n] - C}{C \times n}, & C > 0 \\ 1, & C = 0 \end{cases}. \tag{11}$$

when $C = c_n = 1$, we have $c_1 + c_2 + \ldots + c_{n-1} = 0$, and G_s reaches the maximum

$$\max\ (G_S) = 1 - \frac{1}{n}. \tag{12}$$

when $C = c_1$, we have $c_2 + c_3 + \ldots + c_n = 0$, and G_s reaches the minimum

$$\min\ (G_S) = \frac{1}{n} - 1. \tag{13}$$

so we get the range $G_s \in \left[(1/n) - 1, \quad 1 - (1/n)\right]$, where $n \geqq 5$. When $n \to +\infty$ and $C \geqq 0$, we get $Gs \in (-1.1]$. We also have $G_s = 0$ when the area $A = A^+ + A^- = 0$. For example, when the heartbeat curve completely overlaps with the line of equality in Fig. 2, we get $G_s = 0$, which means the citations evenly distribute, e.g., $H = (1, 1, 1, 1, 1, 1)$ for a sleeping period of six years. In addition, when the positive and negative area between the heartbeat curve and the line of equality (A^+ and A^- respectively) offset, e.g., $H = (0, 1, 2, 2, 1, 0)$, we also get $G_s = 0$. The value of G_s index here depends on n, the duration of sleeping period. In order to make comparisons among different length of heartbeat spectrum, we can define a normalized version (Carpenter, 1979; Egghe & Rousseau, 1990) as follows,

$$\hat{G}_s = \frac{n}{n-1} G_s \tag{14}$$

It is then immediate that $\hat{G}_s \in [-1, 1]$ which does not depend on n.

Let's consider a supposed sleeping beauty as an example: an article received 6 citations in a sleeping period of six years, and then was awakened. In Table 1 lists eight possible heartbeat spectra, and the calculation of their G_s indices. H_1 and H_8 are cases satisfying Eqs. (13) and (12) and their G_s indices respectively reach minimum and maximum when $n = 6$. H_1 is an "all-elements-sleeping-beauty", according to Li and Ye (2012)'s definition.

Table 1
Calculation of the G_s coefficients of eight supposed heartbeat spectra.

i	c_i								y_i							
	H_1	H_2	H_3	H_4	H_5	H_6	H_7	H_8	H_1	H_2	H_3	H_4	H_5	H_6	H_7	H_8
1	6	3	2	1	0	0	0	0	1	1/2	1/3	1/6	0	0	0	0
2	0	2	0	1	1	2	0	0	1	5/6	1/3	1/3	1/6	1/3	0	0
3	0	1	2	1	2	0	0	0	1	1	2/3	1/2	1/2	1/3	0	0
4	0	0	0	1	2	2	1	0	1	1	2/3	2/3	5/6	2/3	1/6	0
5	0	0	2	1	1	0	2	0	1	1	1	5/6	1	2/3	1/2	0
6	0	0	0	1	0	2	3	6	1	1	1	1	1	1	1	1
G_s									−0.833	−0.611	−0.167	0	0	0.167	0.611	0.833
$\hat{G}_s$									−1	−0.733	−0.200	0	0	0.200	0.733	1

The heartbeats of H_4 evenly distribute among the six years, so its heartbeat curve completely overlaps with the line of equality as shown in Fig. 3, and its $G_s = 0$. The G_s indices of H_1, H_2 and H_3 are negative, since they have heartbeats in the first half period. We call these cases "early heartbeats", i.e., most citations were received in the first half of sleeping period. The G_s indices of H_6, H_7 and H_8 are positive and their heartbeats appear in the second half. We call these cases "late heartbeats", i.e., most citations were received in the second half. Table 1 shows that negative G_s indicates "early heartbeats", positive G_s means "late heartbeats", and $G_s = 0$ in H_4 and H_5 presents symmetrical citation distribution or positive and negative area offset. In addition, High heartbeat (high citations) in the first year of sleeping period assures low G_s index, and high heartbeat in the last year assures high G_s index.

3.4. Data

For practical verification, we try to find real data. During the period 1901–2012, the Nobel Prize in Chemistry, Physics, Physiology or Medicine, and the Prize in Economic Sciences were awarded to 163, 194, 201 and 71 laureates, respectively. (http://www.nobelprize.org/). We searched their publications from 1900 to 2000 in the Web of Science of Thomson Reuters, and reduced the duplication of names from the results, by manually scrutinizing the education and research background of each laureate. As a result, we obtained 19,938, 12,862, 22,418 and 3745 papers and their citations till 2011, respectively

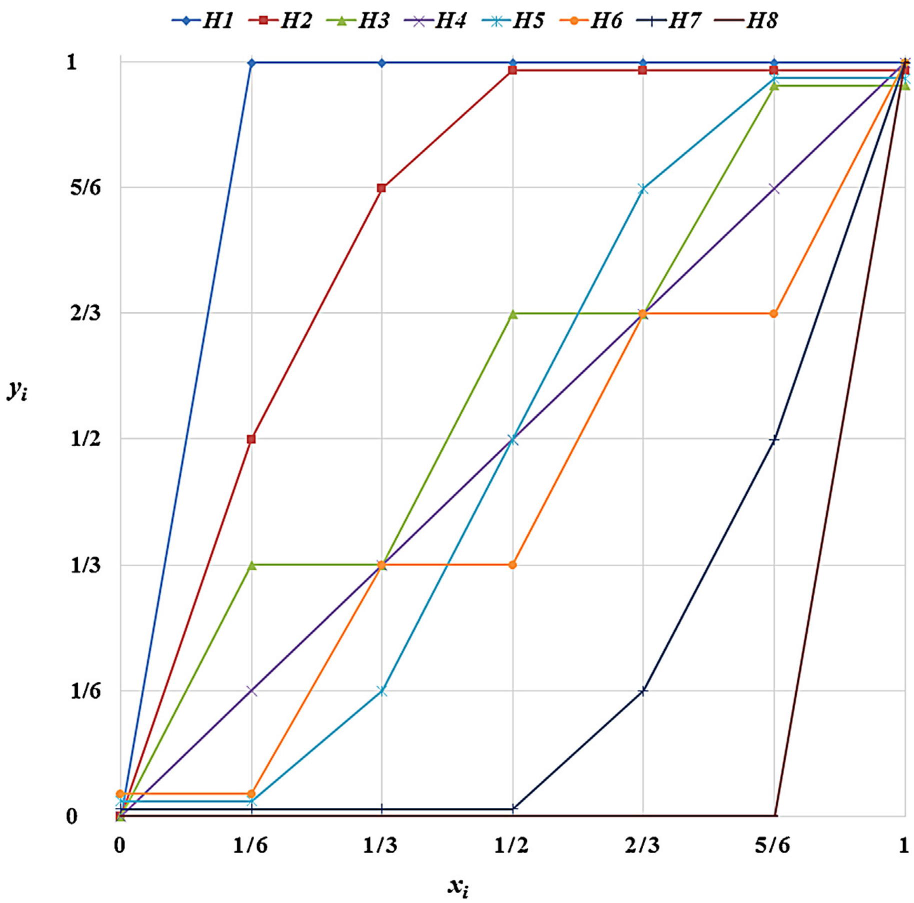

Fig. 3. Heartbeat curves of H_1–H_8 in Table 1.

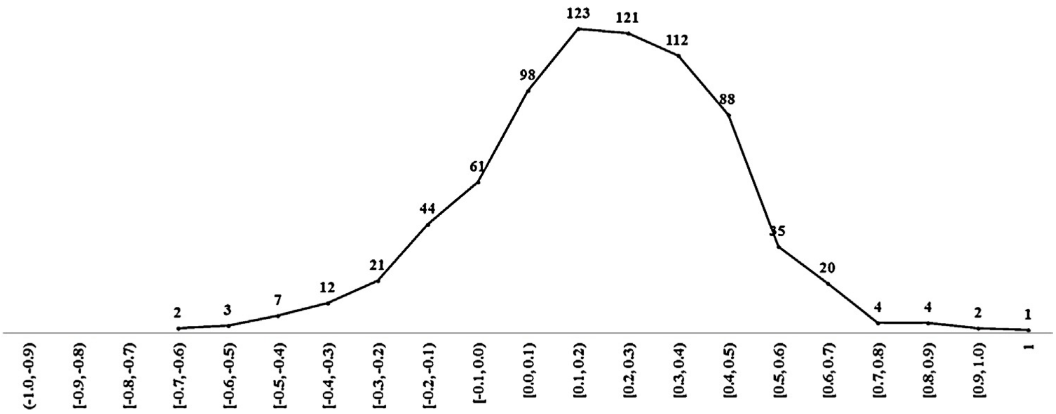

Fig. 4. G_s distribution of sleeping beauties in the intervals between −1 and 1.

for the four subsets of laureates. Among the 58,963 publications, 50,789 received at least one citation in total, left 8174 never-cited. The proportion of never-cited items is large, but it does not indicate exceptional among top scientists (Burrell, 2012; Egghe, Guns, & Rousseau, 2011).

4. Results

4.1. G_s index of heartbeat spectrum

We found 758 sleeping beauties from the 58,963 papers. The distribution of their G_s indices between −1 and 1 presents an approximately symmetric shape in Fig. 4, where most of the sleeping beauties have positive G_s indices and the average value of G_s is 0.084. During the sleeping period of the 758 sleeping beauties, "late heartbeats" is four times that "early heartbeats".

The curve in Fig. 4 peaks in the interval of [0.1, 0.2). There are 79.6% sleeping beauties existing in the interval of [−0.1, 0.5), and 95.4% in the interval of [−0.3, 0.7). So, the cases H_1, H_2 and H_8 are rare ones. There are five sleeping beauties which have $G_s = 0$. All of them have positive and negative area offset rather than uniform citation distribution. There is only one extreme sleeping beauty which has $G_s = 1$, which means the princess had no heartbeat during the sleeping period at all. It is Reichstein and Shoppee (1949) article, which had no heartbeats in the sleeping period of 10 years from 1949 to 1958, and was suddenly awakened by receiving 35 citations from 1959 to 1962.

The heartbeat spectra of publications which slept for at least five years but have not been awakened, is significantly different from those of sleeping beauties, as shown in Fig. 5. Most of the unawakened papers received a few citations in the following years after publication, but quickly declined, like Costas et al. (2010)'s "normal" publications. An extreme example is Wien (1900) paper. It received one citation as soon as it published but was never cited during the following 111 years till 2011. Hence, its G_s index, the minimum among the 45,018 unawakened papers, equals to −0.991 in terms of Eq. (14). The 8174 never-cited papers, whose G_s indices equal to 1, results in leaping in the tail of the curve in Fig. 5.

By contrast, most of the heartbeat spectra of unawakened publications appear "early heartbeats", whereas the heartbeat spectra of sleeping beauties mainly present "late heartbeat". The heartbeat spectrum whose G_s indices lie in the intervals [0.2, 0.3), [0.3, 0.4), [0.4, 0.5) and [0.5, 0.6) has the highest percentages to be awakened, i.e., 11.5%, 15.8%, 19.2% and 14.1%, respectively. The percentages in other intervals are lower than 10.0%. The lowest one is from never-cited papers, i.e., 0.012%.

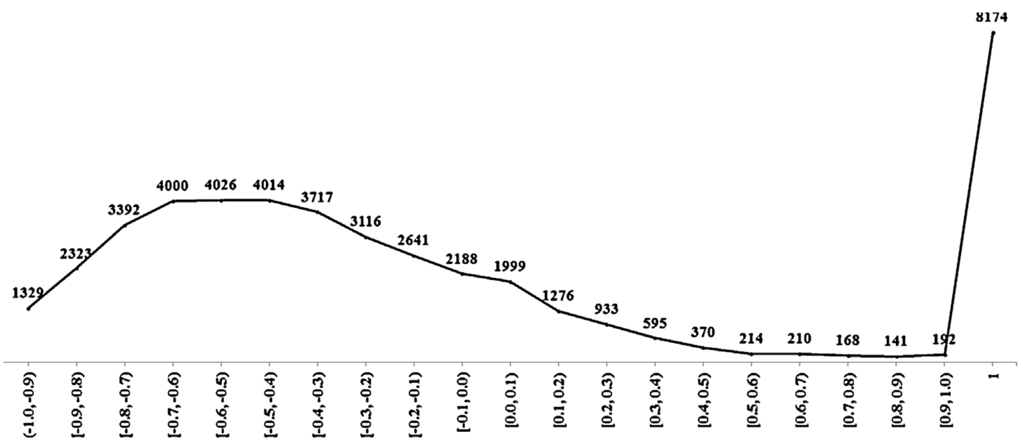

Fig. 5. G_s distribution of unawakened papers in the intervals between −1 and 1.

Table 2
Percentages of being awakened for different length and strength of heartbeat spectra (i = length of heartbeat spectrum, N = number of papers, sb = sleeping beauty, MC = median citations).

i	Deep sleep				Less deep sleep			
	N	N, sb	%	MC, sb	N	N, sb	%	MC, sb
5	29,667	32	0.108	150.5	7409	98	1.323	120.5
6	29,757	13	0.044	159	7493	36	0.480	117
7	29,996	8	0.027	160	7541	20	0.265	122
8	30,240	3	0.010	161	7678	26	0.339	132
9	30,558	4	0.013	164.5	7783	20	0.257	128.5
10	30,930	5	0.016	162	7815	17	0.218	137
15	30,270	4	0.013	165	6708	14	0.209	163
20	29,711	1	0.003	149	5339	15	0.281	173

Table 3
Percentages of being awakened if an article received continuous n zero-citations after publication.

i	N, sb	N	sb%	MC, sb
5	14	8828	0.158	117
6	7	8655	0.081	281
7	6	8556	0.070	157.5
8	3	8482	0.035	101
9	2	8421	0.024	149.5
10	3	8405	0.036	136
>10	15	19,595	0.076	142

4.2. Length and strength of heartbeat spectrum

The longer the length of a heartbeat spectrum is (or the longer an article sleeps), the lower its probability of being awakened is, as shown in Table 2. For deep sleep publications, the length of heartbeat spectrum is uncorrelated to the median citations of sleeping beauties (Pearson correlation −0.186), but for less deep sleep ones, they are strongly correlated (Pearson correlation 0.980). It is significantly easier to awaken less deep sleep publications than deep sleep ones. There are still 0.3% of the publications awakened even if they less deeply slept for 20 years, whereas the percentage for deep sleep publications is only 0.003%. The article which slept for the longest time is Sabatier and Senderens (1902) French paper. It deeply slept 106 years from 1902 to 2007 with 93 citations, then was awakened by receiving 22 citations from 2008 to 2011. It is also an example which deeply slept for a long time and received disproportionate citations.

The probability of being awakened is less than 0.2% if an article had no heartbeat for more than five years, as shown in Table 3. It decreased to less than 0.05% for more than ten years. It is not required that the awakening period follows the n continuous zero-citations, so the strength of heartbeat spectra of these sleeping beauties is not necessarily zero. For example, the most zero-citation article is Lippmann (1908) French paper, which has the first heartbeat (received the first citation) in 1957, and was not awakened until 1999. It received 50 continuous zero-citations and 37 citations in the sleeping period. The unique case whose awakening period closely follows the zero-citation sleeping period is Reichstein and Shoppee (1949) article.

4.3. Comparison of heartbeat spectra in different disciplines

The reference preferences of natural sciences are different from those of social sciences. Natural science researchers tend to cite current articles published in English journals, whereas social science researchers cite older literature and rely on books as well as journal articles (Hicks, 1999; Huang & Chang, 2008; Lariviere, Archambault, & Gingras, 2006; Leydesdorff, 2003). In this research, we take Chemistry, Physics and Physiology or Medicine as "natural sciences", and Economic Sciences as

Table 4
Comparisons of the percentages of being awakened between natural sciences and social sciences.

i	Natural sciences			Social Sciences		
	N	N, sb	%	N	N, sb	%
5	33,942	106	0.312	3134	24	0.766
6	34,170	40	0.117	3079	9	0.292
7	34,492	23	0.067	3045	5	0.164
8	34,898	22	0.063	3024	7	0.231
9	35,337	18	0.051	3004	6	0.200
10	35,757	20	0.056	2988	2	0.067
15	34,337	12	0.035	2641	6	0.227
20	32,688	10	0.031	2362	6	0.254

"social sciences". Table 4 shows that social science publications are significantly easier to be awakened than natural science ones, by a Sign Test (Siegel & Castellan, 1988) at the significance level $\alpha = 0.05$. Sleeping beauties appear more commonly in social science publications. There are 864 social science publications and 7310 natural science publications which received no citations, accounting for 23.1% and 13.2% of the total number of publications in each category, respectively.

5. Discussion and conclusions

Under the framework of "heartbeat spectrum", we can speculate the awakening probability of an article. We found 758 sleeping beauties from Nobel laureates' 58,963 publications between 1900 and 2000. By calculating the G_s indices of 758 sleeping beauties and 45,018 unawakened papers, we firstly observed that the publications which have "late heartbeats" have higher probability to be awakened than those having "early heartbeats". The awakening probability is the highest if an article's G_s index exists in the interval [0.2, 0.6). Secondly, the shorter length or the higher strength the heartbeat spectrum has, the higher its awakening probability is. The awakening probability is rather low if an article has no heartbeat (received continuous zero-citations, "vegetable state") in at least five years. Last but not least, social science publications are significantly easier to be awakened than natural science ones, since the former prefer to cite old literature while the latter prefer current articles.

G_s index introduces citation dispersion to characterize heartbeat spectra of sleeping beauties. It is suggested that never-cited or less cited papers should not be neglected (Hu & Wu, 2014), because some of them become sleeping beauties a few years later. G_s index offers a novel way to assess less cited papers' potential of becoming sleeping beauties. Although G_s index is a feasible indicator, it has limitations for measuring the inequality of citation distribution. First, it cannot differentiate two heartbeat spectra if there is multiplier relationship between them. For example, both (0, 2, 0, 2, 0, 2) and (0, 1, 0, 1, 0, 1) have $G_s = 0.167$. Second, it cannot differentiate two heartbeat spectra if they have $A^+ + A^- = 0$ in Fig. 2. For example, both heartbeat spectrum (1, 1, 1, 1, 1, 1) and (0, 1, 2, 2, 1, 0) in Table 1 have $G_s = 0$.

Moreover, the framework of heartbeat spectrum is extendable. Let vector $(c_{n+1}, c_{n+2}, c_{n+3}, c_{n+4})$ present the citation distribution in the awakening period, and $C_w = c_{n+1} + c_{n+2} + c_{n+3} + c_{n+4}$ denote the total number of citations in the sleeping period where $C_w > 20$, then the "sleeping beauty spectrum" appears $(c_1, c_2, \ldots c_i, \ldots c_n, c_{n+1}, \ldots c_{n+4})$ where $(c_{n+1}, \ldots c_{n+4})$ characterizes the "awakening spectrum". In addition, if the "awaking period" (Li, 2014) requires more than 20 citations within at most four years, the "all-elements-sleeping-beauty spectrum" appears $(c_1{}', \ldots, c_j{}', c_1, c_2, \ldots, c_i, \ldots c_n, c_{n+1}, \ldots, c_{n+4})$ where $4 \geqq j \geqq 1$ and $(c_1{}', \ldots, c_j{}')$ presents the "awaking spectrum". The investigation could stimulate interesting studies for "sleeping beauties".

Acknowledgements

We acknowledge the National Social Science Foundation of China Major Key Project (12&ZD221), the National Natural Science Foundation of China (Grant Nos. 71173187 & 71203193) and the Zhijiang Youth Project of Zhejiang province (13ZJQN045YB) for financial support. We thank Ms. Mingli Jiang for data collecting and anonymous reviewers for excellent suggestions improving our submission.

References

Aksnes, D. W. (2003). Characteristics of highly cited papers. *Research Evaluation*, *12*(3), 159–170.
Aversa, E. S. (1985). Citation patterns of highly cited papers and their relationship to literature aging: A study of the working literature. *Scientometrics*, *7*(3–6), 383–389.
Barber, B. (1961). Resistance by scientists to scientific discovery. *Science*, *134*, 596–602.
Braun, T., Glänzel, W., & Schubert, A. (2010). On sleeping beauties, princes and other tales of citation distributions. *Research Evaluation*, *19*(3), 195–202.
Burrell, Q. L. (2005). Are "Sleeping Beauties" to be expected. *Scientometrics*, *65*(3), 381–389.
Burrell, Q. L. (2012). Alternative thoughts on uncitedness. *Journal of the American Society for Information Science and Technology*, *63*(7), 1466–1470.
Carpenter, M. P. (1979). Similarity of Pratt's measure of class concentration to the Gini index. *Journal of the American Society for Information Science*, *30*(2), 108–110.
Cole, S. (1970). Professional standing and the reception of scientific discoveries. *American Journal of Sociology*, *76*, 286–306.
Costas, R., van Leeuwen, T. N., & van Raan, A. F. J. (2010). Is scientific literature subject to a "sell-by-date"? A general methodology to analyze the "durability" of scientific documents. *Journal of the American Society for Information Science and Technology*, *61*(2), 329–339.
Cunningham, S. J. (1995). An empirical investigation of the obsolescence rate for information systems literature. *Library and Information Science Research*,. Retrieved from http://library.fgcu.edu/iclc/lisrissu.htm
Egghe, L., Guns, R., & Rousseau, R. (2011). Thoughts on uncitedness: Nobel laureates and Fields Medalists as case studies. *Journal of the American Society for Information Science and Technology*, *62*(8), 1637–1644.
Egghe, L., & Rao, I. K. R. (1992). Citation age data and the obsolescence function: Fits and explanations. *Information and Processing Management*, *28*(2), 201–217.
Egghe, L., & Rousseau, R. (1990). Elements of concentration theory. In L. Egghe, & R. Rousseau (Eds.), *Informetrics 89/90* (pp. 97–137). Belgium: Diepenbeek.
Garfield, E. (1970). Would Mendel's work have been ignored if the Science Citation Index was available 100 years ago? *Current Contents*, *2*, 5–6.
Garfield, E. (1980). Premature discovery or delayed recognition-why? *Current Contents*, *4*, 488–493.
Garfield, E. (1989a). More delayed recognition. Part 1. Examples from the genetics of color blindness, the entropy of short-term memory, phosphoinositides, and polymer rheology. *Current Contents*, *38*, 3–8.
Garfield, E. (1989b). Delayed recognition in scientific discovery: Citation frequency analysis aids the search for case histories. *Current Contents*, *23*, 3–9.
Garfield, E. (1990). More delayed recognition. Part 2. From inhibin to scanning electron microscopy. *Current Contents*, *9*, 3–9.
Gini, C. (1912–1955). In E. Pizetti, & T. Salvemini (Eds.), *Italian: Variabilità e mutabilità (variability and mutability), C. Cuppini, Bologna, 156 pages. Reprinted in Memorie di metodologica statistica*. Rome: Libreria Eredi Virgilio Veschi.

Glänzel, W., Schlemmer, B., & Thijs, B. (2003). Better late than never? On the chance to become highly cited only beyond the standard bibliometric time horizon. *Scientometrics*, *58*(3), 571–586.

Glänzel, W. (2004). Towards a model for diachronous and synchronous citation analyses. *Scientometrics*, *60*(3), 511–522.

Glänzel, W. (2008). Seven myths in bibliometrics: About facts and fiction in quantitative science studies. *Collnet Journal of Scientometrics and Information Management*, *2*(1), 9–17.

Glänzel, W., & Garfield, E. (2004). The myth of delayed recognition. *Scientist*, *18*(11), 8–9.

Hicks, D. (1999). The difficulty of achieving full coverage of international social science literature and the bibliometric consequences. *Scientometrics*, *44*(2), 193–215.

Hook, E. B. (2002). *Prematurity in scientific discovery: On resistance and neglect*. Berkeley: University of California Press.

Hu, Z., & Wu, Y. (2014). Regularity in the time-dependent distribution of the percentage of never-cited papers: An empirical pilot study based on the six journals. *Journal of Informetrics*, *8*(1), 136–146.

Huang, M. H., & Chang, Y. W. (2008). Characteristics of research output in social sciences and humanities: From a research evaluation perspective. *Journal of the American Society for Information Science and Technology*, *59*(11), 1819–1828.

Lariviere, V., Archambault, E., Gingras, Y., et al. (2006). The place of serials in referencing practices: Comparing natural sciences and engineering with social sciences and humanities. *Journal of the American Society for Information Science and Technology*, *57*(8), 997–1004.

Lange, L. L. (2005). Sleeping beauties in psychology: Comparisons of "hits" and "missed signals" in psychological journals. *History of Psychology*, *8*(2), 194–217.

Levitt, J. M., & Thelwall, M. (2009). The most highly cited Library and Information Science articles: Interdisciplinarity, first authors and citation patterns. *Scientometrics*, *78*(1), 45–67.

Leydesdorff, L. (2003). Can networks of journal–journal citation be used as indicators of change in the social sciences? *Journal of Documentation*, *59*(1), 84–104.

Li, J., & Ye, F. Y. (2012). The phenomenon of all-elements-sleeping-beauties in scientific literature. *Scientometrics*, *92*(3), 795–799.

Li, J. (2014). Citation curves of "All-elements-sleeping-beauties": "Flash in the Pan" first and then "Delayed Recognition". *Scientometrics*, http://dx.doi.org/10.1007/s11192-013-1217-z

Line, M. B., & Sandison, A. (1974). "Obsolescence" and changes in the use of literature with time. *Journal of Documentation*, *30*(3), 283–350.

Lippmann, G. (1908). Reversible test prints. Integral photographies. *Comptes rendus hebdomadaires des seances de l'Academie des sciences*, *146*, 446–451.

McCain, K. W., & Turner, K. (1989). Citation content analysis and aging patterns of journal articles in molecular genetics. *Scientometrics*, *17*(1/2), 127–163.

Mendel, G. (1866). Experiments in plant hybridisation. (Versuche über Plflanzenhybriden). In *Verhandlungen des naturforschenden Vereines in Brünn, Bd. IV für das Jahr*. Abhandlungen.

Mingers, J. (2007). Shooting stars and sleeping beauties: The secret life of citations. *EURO XXII*. Prague, 8–11 July, abstract is available at http://kar.kent.ac.uk/13133/

Nakamoto, H. (1988). Synchronous and dyachronous citation distributions. In L. Egghe, & R. Rousseau (Eds.), *Informetrics 87/88* (pp. 157–163). Amsterdam: Elsevier Science Publishers.

Ohba, N., & Nakao, K. (2012). Sleeping beauties in ophthalmology. *Scientometrics*, *93*(2), 253–264.

Peirce, C. S. (1884). The numerical measure of the success of predictions. *Science*, *4*(93), 453–454.

Pratt, A. O. (1977). A measure of class concentration in bibliometrics. *Journal of the American Society for Information Science*, *28*(5), 285–292.

Price, D. J. D. (1965). Networks of scientific papers. *Science*, *149*(3683), 510–515.

Price, D. J. D. (1976). A general theory of bibliometrics and other cumulative advantage processes. *Journal of American Society for Information Science*, *27*(5), 292–306.

Reichstein, T., & Shoppee, C. W. (1949). Chromatography of steroids and other colourless substances by the method of fractional elution. *Discussions of the Faraday Society*, (7), 305–311.

Sabatier, P., & Senderens, J. B. (1902). New methane synthesis. *Comptes rendus hebdomadaires des seances de l'Academie des sciences*, *134*, 514–516.

Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioural sciences* (2nd ed.). New York: McGraw-Hill.

Stent, G. S. (1972). Prematurity and uniqueness in scientific discovery. *Scientific American*, *227*(6), 84–93.

van Raan, A. F. J. (2004). Sleeping beauties in science. *Scientometrics*, *59*(3), 467–472.

van Dalen, H. P., & Henkens, K. (2005). Signals in science – On the importance of signaling in gaining attention in Science. *Scientometrics*, *64*(2), 209–233.

van Clester, B. (2012). It takes time: A remarkable example of delayed recognition. *Journal of the American Society for Information Science and Technology*, *63*(11), 2341–2344.

Wien, W. (1900). Possible ether movement. *Physikalische Zeitschrift*, *2*, 148–150.

Wyatt, H. V. (1961). Knowledge and prematurity-journey from transformation to DNA. *Perspectives in Biology and Medicine*, *18*(2), 149–156.

Zirkle, C. (1964). Some oddities in the delayed discovery of mendelism. *Journal of Heredity*, *55*(2), 65–72.