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
Gini 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 $G_i$ 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 $G_i$ index exists in the interval [0.2, 0.6].

© 2014 Elsevier Ltd. All rights reserved.

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”
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 criteria, 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: $N = \{f(s, c_s, c_w) \sim s^{-0.25} \}$, 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; Obba & 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 beenleeping 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 awakening period (leaping), a sleeping period, an awakening period and a happy ending (the princess and the prince received high co-citations). 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 ≤ “Year 50%” ≤ 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.
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 $c_i \geq 0$ denote the number of citation(s) it received in the i-th year in the sleeping period, then the sleeping beauty’s heartbeat in the i-th 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 $n = t_n - t_1 + 1 \geq 5$.

Definition 4: The “strength of heartbeat spectrum” is the average citations in the sleeping period, i.e., $C_s = \sum_{i=1}^n c_i / n = C/n$, where $C$ is the number of citations received in the sleeping period, $0 \leq C_s < 1$ indicates deep sleep and $1 < C_s \leq 2$ 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
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.,

$$G = \frac{A}{A + B}$$  \hspace{1cm} (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. $G_2$ index

We propose $G_2$ 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 $G_2$ 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, $G_2$ index measures an overall difference between the heartbeat curve and the uniform distribution curve, (the line of equality), i.e.,

$$G_2 = \frac{A}{A + B}$$  \hspace{1cm} (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:

\[
x_i = \frac{i}{n},
\]

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

\[
y_i = \frac{\sum_{1}^{i} c_i}{\sum_{1}^{n} c_i} = \frac{\sum_{1}^{i} c_i}{C}
\]

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

\[
A + B = \frac{1}{2}.
\]

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

\[
G_s = 1 - \frac{\sum_{y_i}^{n} (1/2) \times (y_i + y_{i-1}) \times (1/n))}{1/2}.
\]

where \( y_0 = 0 \). Then we have

\[
G_s = 1 - \frac{2 \times \sum_{1}^{n} y_i - y_n}{n}.
\]

when \( C > 0 \), we have \( y_n = 1 \) and

\[
G_s = 1 - \frac{2 \times \sum_{1}^{n} y_i - 1}{n},
\]

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

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

when \( C = 0 \), we have \( y_i = 0 \) and

\[
G_s = 1
\]

therefore, the segmented function of \( G_s \) index is

\[
G_s = \left\{ \begin{array}{ll}
1 - \frac{2 \times [n \times c_1 + (n - 1) \times c_2 + \cdots + c_n] - C}{C \times n}, & C > 0 \\
1, & C = 0
\end{array} \right.
\]

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

\[
\max (G_s) = 1 - \frac{1}{n}.
\]

when \( C = c_1 \), we have \( c_2 + c_3 + \cdots + c_n = 0 \), and \( G_s \) reaches the minimum

\[
\min (G_s) = \frac{1}{n} - 1.
\]

so we get the range \( G_s \in [1/n, 1 - (1/n)] \), where \( n \geq 5 \). When \( n \rightarrow +\infty \) and \( C \geq 0 \), we get \( G_s \in (-1,1] \). 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 \( (A^+ \text{ and } A^- \text{ respectively}) \) 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,

\[
\hat{G}_s = \frac{n}{n-1} G_s
\]

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 \( H_6 \) 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 $G_4$ coefficients of eight supposed heartbeat spectra.

<table>
<thead>
<tr>
<th>$i$</th>
<th>$c_i$</th>
<th>$y_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$H_1$</td>
<td>$H_2$</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

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 $G_4 = 0$. The $G_4$ 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 $G_4$ 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 $G_4$ indicates “early heartbeats”, positive $G_4$ means “late heartbeats”, and $G_4 = 0$ 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 $G_4$ index, and high heartbeat in the last year assures high $G_4$ 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.
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 \(H_3\) 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%.
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).

<table>
<thead>
<tr>
<th>i</th>
<th>Deep sleep</th>
<th>Less deep sleep</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>N, sb</td>
</tr>
<tr>
<td>5</td>
<td>29,667</td>
<td>32</td>
</tr>
<tr>
<td>6</td>
<td>29,757</td>
<td>13</td>
</tr>
<tr>
<td>7</td>
<td>29,996</td>
<td>8</td>
</tr>
<tr>
<td>8</td>
<td>30,240</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>30,558</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>30,930</td>
<td>5</td>
</tr>
<tr>
<td>15</td>
<td>30,270</td>
<td>4</td>
</tr>
<tr>
<td>20</td>
<td>29,711</td>
<td>1</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>i</th>
<th>N, sb</th>
<th>N</th>
<th>sb%</th>
<th>MC, sb</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>14</td>
<td>8828</td>
<td>0.158</td>
<td>117</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>8655</td>
<td>0.081</td>
<td>281</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>8556</td>
<td>0.070</td>
<td>157.5</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>8482</td>
<td>0.035</td>
<td>101</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>8421</td>
<td>0.024</td>
<td>149.5</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>8405</td>
<td>0.036</td>
<td>136</td>
</tr>
<tr>
<td>&gt;10</td>
<td>15</td>
<td>19,595</td>
<td>0.076</td>
<td>142</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>i</th>
<th>Natural sciences</th>
<th>Social Sciences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N, sb</td>
<td>%</td>
</tr>
<tr>
<td>5</td>
<td>33,942</td>
<td>106</td>
</tr>
<tr>
<td>6</td>
<td>34,170</td>
<td>106</td>
</tr>
<tr>
<td>7</td>
<td>34,492</td>
<td>23</td>
</tr>
<tr>
<td>8</td>
<td>34,898</td>
<td>22</td>
</tr>
<tr>
<td>9</td>
<td>35,337</td>
<td>18</td>
</tr>
<tr>
<td>10</td>
<td>35,757</td>
<td>20</td>
</tr>
<tr>
<td>15</td>
<td>34,337</td>
<td>12</td>
</tr>
<tr>
<td>20</td>
<td>32,688</td>
<td>10</td>
</tr>
</tbody>
</table>
“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 \( G_1 \) 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 \( G_1 \) 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.

\( G_1 \) 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. \( G_1 \) index offers a novel way to assess less cited papers’ potential of becoming sleeping beauties. Although \( G_1 \) 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 \( G_1 = 0.167 \). Second, it cannot differentiate two heartbeat spectra if they have \( A^* + A^- = 0 \) in Fig. 2. For example, both heartbeat spectrum \((1, 1, 1, 1, 1)\) and \((0, 1, 2, 2, 1, 0)\) in Table 1 have \( G_1 = 0 \).

Moreover, the framework of heartbeat spectrum is extendable. Let vector \((c_{n1}, c_{n2}, c_{n3}, c_{n4})\) present the citation distribution in the awakening period, and \( C_w = c_{n1} + c_{n2} + c_{n3} + c_{n4} \) denote the total number of citations in the sleeping period where \( C_w > 20 \), then the “sleeping beauty spectrum” appears \((c_1, c_2, \ldots, c_n, c_{n1}, \ldots, c_{n4})\) where \((c_{n1}, \ldots, c_{n4})\) characterizes the “awakening spectrum”. In addition, if the “awakening period” (Li, 2014) requires more than 20 citations within at most four years, the “all-elements-sleeping-beauty spectrum” appears \((c_1', \ldots, c_j', c_1, c_2, \ldots, c_i, c_{n1}, \ldots, c_{n4})\) where \(4 \leq j \leq 1 \) and \((c_1', \ldots, c_j')\) presents the “awakening 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 (13ZJQQN045YB) for financial support. We thank Ms. Mingli Jiang for data collecting and anonymous reviewers for excellent suggestions improving our submission.

References

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.

Author’s personal copy


