Placeholder: Delayed Recognition of Co-cited Articles

Wenxi Zhao [1] **and George Chacko** [1,*]

[1] *Netelabs, NET ESolutions (an NTT DATA Company), McLean, VA, USA*

Correspondence*:
George Chacko
netelabs@nete.com

ABSTRACT

For full guidelines regarding your manuscript please refer to Author Guidelines.

As a primary goal, the abstract should render the general significance and conceptual advance of the work clearly accessible to a broad readership. References should not be cited in the abstract. Leave the Abstract empty if your article does not require one, please see Summary Table for details according to article type.

Keywords: bibliometrics, co-citations, X, Y, Z

1 INTRODUCTION

Wenxi- we need to have a discussion about each of the following references and what they have contributed to the understanding of delayed recognition and sleeping beauties (Garfield, 1980; van Raan, 1990; Glänzel et al., 2003; Glänzel and Garfield, 2004; van Raan, 2004; Wang et al., 2013; Li, 2014; Ke et al., 2015; Li and Ye, 2016; Song et al., 2018; van Raan and Winnink, 2019). I will prepare some notes for Monday too. *Need placeholder for Devarakonda et al. (2020)*

Garfield-1980: Clarity in writing. Premature discoveries or Resisted Discoveries. Position in the hierarchy of science. Cole thought it was the result of content. Mendel 1866-1900.

Sleeping Beauty in scientific community refers to a single well-cited article which receives delayed recognition because of being ahead of time or overlooked for periods. Anthony F.J. van Raan and other researchers have proposed different methods to identify Sleeping Beauty articles and measure their kinetics. This phenomenon also happens for co-cited pairs that a pair of publications have been cited by one or more publications together. The more often a pair of publications has been co-cited, the more related they are assumed to be [1]. The appearance of Sleeping Beauty among co-cited pairs indicates that as time goes on, different disciplines gradually have overlaps or different ideas have been combined and evolved into a new concept. In our experiment, we extend to examine Sleeping Beauty phenomenon for highly co-cited pairs that have been sunk in sleep for a long period before attracting citations, propose an approach to define Sleeping Beauty among co-cited pairs and investigate the kinetics of Sleeping Beauty co-cited pairs and their individual articles.

[1] https://arxiv.org/pdf/1707.03076.pdf

To find possible Sleeping Beauty co-cited pairs, we constructed a dataset of articles from Scopus (Elsevier BV, 2019) by extracting co-cited pairs which were composed of highly co-cited individual articles and have received at least 10 co-citations. We assembled 4.12 million co-cited pairs as a result.

We developed an initial method to find possible Sleeping Beauty co-cited pairs within our dataset, and in order to further investigate kinetics of those possible Sleeping Beauty co-cited pairs, we built another dataset of individual articles from filtered co-cited pairs. To determine the number of Sleeping Beauty individual articles found in this dataset, we applied modified van Raan's procedures to filter out articles by setting the most stringent constraints and then calculated the beauty coefficient proposed by Ke et al. Then by calculating the number of Sleeping Beauty individual articles that each filtered co-cited pairs has and adjusting the constraints we applied, we answered the question that does it take at least one Sleeping Beauty individual article to generate a Sleeping Beauty co-cited pair.

Lastly, we also investigated edge cases that a co-cited pair awaked after a long period of dormancy and then fell asleep again. We proposed a new method to define such Sleeping Beauty individual articles, since both van Raan's and Ke's method were not applicable in such cases, and defined them as unstable Sleeping Beauties.

2 MATERIAL AND METHODS

[process of getting 4.12 million data]

To find possible sleeping beauty pairs, we set four stringent conditions on 4.12 million co-cited pairs: (1) The total number of co-citation frequency that a co-cited pair received is at least 100; (2) The highest co-citation frequency that a co-cited pair received is at least 20; (3) Both individual papers for a co-cited pair should have published no earlier than 1970; (4) A co-cited paper should have slept for at least 10 years and received no more than 2 co-citations in each year during sleeping period. The sleeping duration for a co-cited pair is defined as the number of years from the first possible co-cited year, which is the publication year of the newer individual paper, to the first year that the pair receives more than 2 co-citations. 1380 co-cited pairs satisfy have been extracted from 4.12 million pairs data and have been labeled as possible sleeping beauty co-citation pairs for further experiment. Next, we used a parameter-free approach, proposed by (Ke et al., 2015), and a bibliometric approach, proposed by (van Raan and Winnink, 2019), to investigate whether the individual papers of 1380 co-cited pairs are sleeping beauties or not.

Ke et al proposed Beauty Coefficient B to quantify how much a given paper can be considered as an SB (Ke et al., 2015). Given a publication, Ke defines c_0 as the number of citations received in the year of publication and c_t as the number of citations received in each year after the year of publication, where t indicates the number of years. The reference line of a publication is defined as a straight line l_t that connects the start point $(0, c_0)$ and peak point (t_m, c_{t_m}) where m indicates that at the age of t_m the paper receives its maximum number of citations c_{t_m}. Then this line l_t can be described by the equation:

$$l_t = \frac{c_{t_m} - c_0}{t_m} \times t + c_0 \tag{1}$$

where $t \in [t_0, t_m]$ and $c_{t_m} - c_0/t_m$ is the slope of the reference line. Then for each t, first compute difference between the reference line l_t and the citation history of the paper c_t, and second compute the ratio between this difference and $max\{1, c_t\}$. By summing up this ratio for each t, the Beauty Coefficient B is defined as:

$$B = \sum_{t=0}^{t_m} \frac{\frac{c_{t_m} - c_0}{t_m} \times t + c_0 - c_t}{max\{1, c_t\}} \tag{2}$$

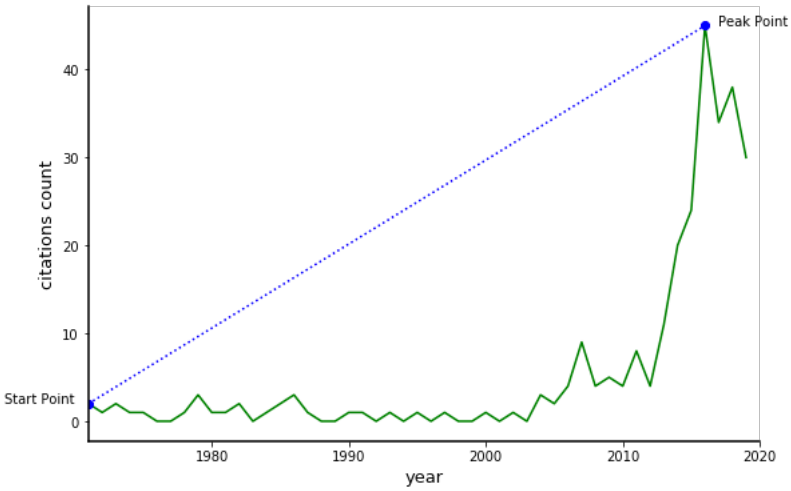

Figure 1. Demonstration of Ke's Beauty Coefficient (Eq. 2)

(We can discuss whether we should add anything to this plot based on the defects I stated below, or cite it directly from Ke, or just remove it)

On **Figure 1**, The green curve represents the number of citations $c_t$that a paper received at each year and t is the number of years between two years. The blue dotted line represents the reference line, which connects the peak point (t_m, c_{t_m}) and start point $(0, c_0)$. This Beauty Coefficient B can be computed for any paper and does not rely on arbitrary selections of parameters such as length of sleeping period and awakening intensity. A paper who have a linear kinetics with time $c_t = l_t$ will have Beauty Coefficient $B = 0$, and thus if a paper can be considered as Sleeping Beauty, it must have a positive Beauty Coefficient, which means its kinetics is a concave function of time. Further, this equation penalizes earlier citations since the summation of ratio in earlier years will be much less than the one in later years when the number of citations received is the same, and so with longer sleeping period, larger maximum citations received, and more abrupt changes happened before the maximum number of citations reached, a publication will have a larger Beauty Coefficient B.

But Ke's method also has three major defects: (i) This approach doesn't put any constraints on any parameters. In general, a paper who has very large Beauty Coefficient B will have a long sleeping period, abrupt changes in citations received before reaching the peak point, and a very high peak point. But in extreme cases, as long as the peak point is high enough, a paper will definitely have a larger B value. For example, if a paper published in 2014 receives 3000 citations in 2018, then even though the sleeping period is only 3 years and each year receives only 2 citations, by (Eq. 2) the Beauty Coefficient B equals to 2998, which is large enough for a paper to be identified as Sleeping Beauty(Ke et al lists top 15 Sleeping Beauties in science and all of those papers have Beauty Coefficient B larger than 2000 (Ke et al., 2015)). The paper in such extreme case apparently shouldn't be considered as Sleeping Beauty, which also proves that a paper must have a large positive Beauty Coefficient to be identified as Sleeping Beauty, but a paper with large positive Beauty Coefficient is not necessarily a Sleeping Beauty. (ii) Ke et al mentions that Sleeping Beauty papers will have a large Beauty Coefficient B, but he doesn't provide a exact threshold to define how large a Beauty Coefficient B should a paper have to be a Sleeping Beauty. When he examines interdisciplinary nature of top Sleeping Beauties, he divides papers into three disjoint subsets with high, medium, and low values of Beauty Coefficient B, which is the group of top 1000 SBs $(B \geq 317.93)$, the $1001st$ to the top 1% SBs $(33.21 \leq B \leq 317.93)$, and the rest $(B \leq 33.21)$ (Ke et al., 2015)). These cutoff values are quite

arbitrary and cannot be directly applied to other datasets. Thus the application of this Beauty Coefficient is quite unclear. (iii) Ke et al does not consider the citations received after the peak point is reached. Since Sleeping Beauty papers are those who have delayed recognization, in most cases they will keep sleeping in earlier years, gradually gain citations and reach the peak point in recent years. But this doesn't mean that we can completely ignore the kinetics after the peak point is reached. The biggest problem it will cause is that when a paper reaches its peak point twice, how should this reference line be drawn? Should we connect the start point to the first peak point or the second peak point? What if there are more than two peaks? Further, if a paper reaches its peak point in its publication year, and goes back to sleep for several years and wake up again, should we ignore all those kinetics since the start point equals to the peak point and the reference line could not be drawn? We'll further explore this kine of edge case later.

van Raan proposed another totally different method to identify Sleeping Beauties. He tuned four main parameters: (1) length of sleeping period; (2) depth of sleep, in terms of the citation rate during sleeping period; (3) awakening period, which is 5 years after sleeping period; (4) awakening citation-intensity in terms of the citation rate during awakening period (van Raan and Winnink, 2019)). van Raan focused on investigating the number of SBs identified with different combinations of parameters, especially with different length of sleeping period, and deriving the general 'General Sleeping Beauty Equation' to gives the number of SBs identified depend on those parameters.

In our experiment, we made minor adjustments to this method. Based on initial limitations we set on our Sleeping Beauty co-citation pairs, we also required individual papers to have slept for at least 10 years. Thus we didn't split sleeping period into several classes as van Raan did, instead we calculated the exact length of sleeping duration for each paper. After adjustments, we also have four main parameters as van Raan proposed and the corresponding limitations are: (1) the length of sleeping period should be at least 10 years; (2) the citation rate during sleeping period should be between 0 and 1; (3) the awakening period is defined as 5 years following the sleeping period; (4) the citation rate during awakening period should be at least 5. Given that t_0 is the year of publication, t_n is the last year of sleeping duration and c_{t_i} is the number of citations received at year t_i, the definition of Sleeping Beauty can be written out as:

$$\text{for } \; n \geq 10, 0 \leq \{c_{\mathrm{t_0}} + ... + c_{\mathrm{t_n}}\}/n \leq 1, \{c_{\mathrm{t_{n+1}}} + ... + c_{\mathrm{t_{n+5}}}\}/5 \geq 5 \tag{3}$$

Although van Raan's method is very solid, we still find that one thing has been overlooked: since van Raan didn't consider the maximum number of citations received as an important parameter to be tuned, in cases where a paper slept for ten years, has citation rate during sleeping period between 0 and 1, and then has received 5 citations each year for 5 years, which is the awakening period, it shouldn't be identified as a Sleeping Beauty. In such case, the paper satisfies all conditions but it doesn't show an increasing trend during awaking period. Further, after the awaking period, the paper can go back to sleep again. In such case, this paper will reach its peak point, 5 citations, in its awaking period. van Raan didn't specify either the peak point can't be reached during awaking period or the minimum value of peak point, and this will mistakenly identify some unqualified papers as Sleeping Beauty.

From our 1380 Sleeping Beauty co-cited pairs, we generated the dataset of 1398 individual publications. To identify Sleeping Beauties from this dataset, we used Ke's and van Raan's method to identify SBs independently and compare the result, and then combined them together to avoid some defects mentioned before. Right now we only consider individual publications that have only one peak point. By applying both methods independently, we got the result of number of SBs identified in this 1398 individual publications dataset. Further, we extracted out Sleeping Beauty co-citation pairs which have at least one individual

publication is identified as SB. Then by comparing the result of number of SBs identified by both methods and the number of pairs that have at least one SB, we found the overlaps and difference between two methods, which helped us design the next experiment to combine two methods together.

After setting limitations on individual publications based on van Raan's paper to extract possible SBs, we calculated Beauty Coefficient for those SBs. Then we designed an experiment to calculate the minimum number of Beauty Coefficient that a paper should have and set it as threshold for us to filter out qualified SBs. By setting van Raan's conditions first, papers that have (1) sleeping duration is at least 10 years; (2) citation rate during sleeping duration is between 0 and 1; (3) citation rate during awakening period is at least 5; (4) the maximum number of citations received is at least 20; will be extracted out. Thus we avoided the first defect that Ke's method has and the overlooked point of van Raan's method by setting the peak point. For an extreme case, the minimum conditions that a a paper should have are listed below:

$$n = 10, \{c_{t_0} + ... + c_{t_n}\}/n = 1, \{c_{t_{n+1}} + ... + c_{t_{n+5}}\}/5 = 5, c_{t_m} \geq 20 \quad (4)$$

where c_{t_m} is the maximum number of citations received, which means at year t_m the paper reaches its peak point.

[Implement algorithm to find threshold here]

[Edge case: papers that have more than one peak point. I'm still working on how to measure this kind of edge case. A plot will be added here]

3 RESULT AND DISCUSSION

As we mentioned above, by setting four stringent conditions on our 4.12 million pairs data, we extracted out 1380 Sleeping Beaut y co-citation pairs. To investigate the relationship between Sleeping Beauty co-citation pairs and Sleeping Beauty individual publications, we further explored kinetics of 1398 individual papers which composed of our 1380 Sleeping Beauty co-citation pairs dataset. By using both parameter-free approach and parameter-dependent approach, which are proposed by Ke et al and van Raan respectively, we got 123 Sleeping Beauty individual publications identified by van Raan's method, and since Ke didn't set a clear threshold for Beauty Coefficient, we set different threshold by adjusting the quantile of Beauty Coefficient and extracted out papers that have Beauty Coefficient higher than each threshold. Next, we recorded the number of papers that have been identified as Sleeping Beauty by both Ke and van Raan, and got the number of Sleeping Beauty co-citation pairs that have at least 1 SB, have 2 SBs, and have 0 SB. The results shows in the following table:

Table 1. Result Table

B Threshold	B Value	#SB by Ke	#SB by Both	#Pairs with SB=0	#Pairs with SB=1	#Pairs with SB=2
$\geq$1000	1000	30	12	1358	20	2
0.90 quantile	336.64	140	42	1308	65	7
0.75 quantile	177.79	350	61	1260	105	15
0.50 quantile	82.11	699	114	1229	129	22
0.25 quantile	37.37	1408	123	1220	137	23

4 CONCLUSION

[Add Conclusions Here]

5 ADDITIONAL REQUIREMENTS

For additional requirements for specific article types and further information please refer to Author Guidelines.

CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

The Author Contributions section is mandatory for all articles, including articles by sole authors. If an appropriate statement is not provided on submission, a standard one will be inserted during the production process. The Author Contributions statement must describe the contributions of individual authors referred to by their initials and, in doing so, all authors agree to be accountable for the content of the work. Please see here for full authorship criteria.

FUNDING

Details of all funding sources should be provided, including grant numbers if applicable. Please ensure to add all necessary funding information, as after publication this is no longer possible.

ACKNOWLEDGMENTS

This is a short text to acknowledge the contributions of specific colleagues, institutions, or agencies that aided the efforts of the authors.

SUPPLEMENTAL DATA

Supplementary Material should be uploaded separately on submission, if there are Supplementary Figures, please include the caption in the same file as the figure. LaTeX Supplementary Material templates can be found in the Frontiers LaTeX folder.

DATA AVAILABILITY STATEMENT

The datasets [GENERATED/ANALYZED] for this study can be found in the [NAME OF REPOSITORY] [LINK].

REFERENCES

Garfield, E. (1980). Premature discovery or delayed recognition- why? *Current Contents* 21, 5–10. Also published in Essays of an Information Scientist, Vol:4, p.488-493, 1979-80

Glänzel, W. and Garfield, E. (2004). The myth of delayed recognition. *Scientist* 18, 8

Glänzel, W., Schlemmer, B., and Thijs, B. (2003). Better late than never? on the chance to become highly cited only beyond the standard bibliometric time horizon. *Katholieke Universiteit Leuven, Open Access publications from Katholieke Universiteit Leuven* 58. doi:10.1023/B:SCIE.0000006881.30700.ea

Ke, Q., Ferrara, E., Radicchi, F., and Flammini, A. (2015). Defining and identifying Sleeping Beauties in science. *Proceedings of the National Academy of Sciences* 112, 7426–7431. doi:10.1073/pnas.1424329112

Li, J. (2014). Citation curves of "all-elements-sleeping-beauties": "flash in the pan" first and then "delayed recognition". *Scientometrics* 100, 595–601. doi:10.1007/s11192-013-1217-z

Li, J. and Ye, F. Y. (2016). Distinguishing sleeping beauties in science. *Scientometrics* 108, 821–828. doi:10.1007/s11192-016-1977-3

Song, Y., Situ, F., Zhu, H., and Lei, J. (2018). To be the prince to wake up sleeping beauty: the rediscovery of the delayed recognition studies. *Scientometrics* 117, 9–24

van Raan, A. F. J. (1990). Fractal dimension of co-citations. *Nature* 347, 626–626. doi:10.1038/347626a0

van Raan, A. F. J. (2004). Sleeping Beauties in Science. *Scientometrics* 59, 467–472. doi:10.1023/B:SCIE.0000018543.82441.f1

van Raan, A. F. J. and Winnink, J. J. (2019). The occurrence of 'Sleeping Beauty' publications in medical research: Their scientific impact and technological relevance. *PLOS ONE* 14, 1–34. doi:10.1371/journal.pone.0223373

Wang, D., Song, C., and Barabási, A.-L. (2013). Quantifying Long-Term Scientific Impact. *Science* 342, 127–132. doi:10.1126/science.1237825

FIGURE CAPTIONS

Figure 2. Enter the caption for your figure here. Repeat as necessary for each of your figures

A

B

Figure 3. This is a figure with sub figures, **(A)** is one logo, **(B)** is a different logo.