Scientometrics (2017) 113:1645–1663
DOI 10.1007/s11192-017-2543-3

Sleeping beauties in Computer Science: characterization and early identification

Ratnadeep Dey[1] · Anurag Roy[1] · Tanmoy Chakraborty[2] · Saptarshi Ghosh[3]

Received: 24 May 2017 / Published online: 16 October 2017

Abstract While a large majority of scientific publications get most of their citations within the initial few years after publication, there is an interesting number of papers—termed as *sleeping beauties*—which do not get much cited for several years after being published, but then suddenly start getting cited heavily. In this work, we focus on sleeping beauties (SBs) in the domain of Computer Science. We identify more than 5,000 sleeping beauties in Computer Science, and characterise them based on their sub-field and their citation profile after awakening. We also reveal some interesting factors which led to their awakening long after publication. Furthermore, we also propose a methodology for early identification of sleeping beauties, and develop a machine learning-based classification approach that attempts to classify publications based on whether they are likely to be SBs. The classifier achieves a precision of 0.73 and a recall of 0.45 in identifying SBs immediately after their year of publications, and the performance significantly improves with time. To our knowledge, this is the first study on sleeping beauties in Computer Science.

Keywords Citation networks · Sleeping beauties · Prince of sleeping beauty · Classification

✉ Tanmoy Chakraborty
tanmoy@iiitd.ac.in

[1] Department of Computer Science and Technology, Indian Institute of Engineering Science and Technology, Shibpur, Howrah 711103, India

[2] Department of Computer Science and Engineering, Indraprastha Institute of Information Technology (IIIT-D), Delhi, India

[3] Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur, Kharagpur 721302, India

Introduction

Several prior works have studied how the citation patterns of scientific papers vary over time (Chakraborty et al. 2015; Garfield 2001). It has been seen that, while a large majority of papers get most of their citations within the initial few years after publication, followed by an exponential decay, there are few exceptionally popular papers which steadily accumulate citations with time. Yet another interesting class of papers has been observed—those which do not get much cited for several years after being published, but then suddenly start getting cited. This work focuses on such papers, which are typically termed as *sleeping beauties* (abbreviated as SBs) (Raan 2004).

There have been prior studies on the phenomenon of 'late awakening' or 'delayed recognition' of papers; see "Related work" section for a literature survey. However, almost all the prior studies have focused on papers in basic sciences (e.g., Physics), and, to our knowledge, there has not been any prior study of SBs in the Computer Science domain.

In this work, using a large dataset of papers crawled from Microsoft Academic Search, we identify and characterize more than 5,000 SBs in Computer Science. Different from the prior works, we find various sub-classes within SBs, which behave differently along various aspects. For instance, while some SBs continue to get increasingly cited after awakening, many other SBs get cited for few years and then the citations decline again. Again, we show that many characteristics of SBs vary depending on the related sub-fields of the Computer Science domain. For instance, most SBs are from the sub-fields of 'Algorithms and theory' and 'Scientific Computing'. Many of these SBs awaken after longer durations of time (as compared to SBs from other sub-fields); however, once they awaken, they often get cited from many distinct sub-fields of Computer Science. In contrast, the SBs from the sub-fields 'Natural Language and speech' and 'Hardware and Architecture' are generally cited from only their own sub-field. We also analyze potential factors which lead to the awakening of a SB long after publication, and find that these factors can also be different for SBs related to different sub-fields of Computer Science.

We further attempt to develop a methodology for identifying SBs as early as possible after their publication. To this end, we propose a set of novel characteristic features derived from the meta-data of each paper, and then use a machine learning-based classification setting to identify SBs. In particular, the feature set constitutes information related to author, publication venue, keywords and citations of each paper. Among three predictive models used for this identification, SVM turns out to be the best model, achieving a precision of 0.73, a recall of 0.45 immediately after the publication of the examined papers. With the progress of time, as more evidences are accumulated, the performance of the model increases up to 10% in terms of precision and 35% in terms of recall.

The rest of the paper is organised as follows. "Related work" section contains a brief literature survey on sleeping beauties in other domains. "Dataset and identification of sleeping beauties" section details the dataset and the methodology we used to identify SBs, while "Characterizing sleeping beauties" section characterises the SBs that were identified. "What leads to awakening of SBs?" section investigates the various factors that lead to the awakening of SBs. Finally, a methodology for early identification of SBs is discussed in "Early identification of sleeping beauties" section, and the paper is concluded in "Conclusion" section.

Related work

Understanding the dynamics of citation growth of scientific articles has always been an interesting problem in bibliometrics (Wallace et al. 2009; Solomon et al. 2013). While studying citation dynamics of Computer Science articles over its effective lifetime, a generalized observation (Chakraborty et al. 2014, 2015; Garfield 1999) reveals that, following publication there is an initial growth (growing phase) in the frequency of citations collected within the first two to three years, followed by a constant peak, i.e., the frequency of incoming citations becomes stagnant for next one to two years (saturation phase), and then, there is a final decline over rest of the lifetime of article (decline phase) and gradually, at some point no further activity is observed (obsolete phase). However, the motivation of our present paper stems from a fundamental question raised by Ruiz-Castillo (2013) in connection with scientometrics that goes as follows: "Are citation distributions for different sciences very similar or rather different?"

In our earlier work (Chakraborty et al. 2015), we introduced the idea of various citation profiles of scientific articles in Computer Science domain. We further showed that none of the existing growth models such as Preferential Attachment models (Barabasi and Albert 1999) capture these profiles, and hence we proposed a new citation growth model to mimic these diverse citation profiles. In the following work, we showed how one can use this profile information to predict the future citation count of an article at the time of its publication (Chakraborty et al. 2014). We proposed a two-stage stratified learning framework which in the first stage uses a rule-based approach to map the citation profile of the examined paper to one of the categories; then in second stage the model is trained on papers belonging to only the mapped category, to predict the future citation count of the examined paper. We also quantified the interdisciplinarity of a paper (vis-a-vis a domain) by analysing the citation distribution and contextual properties of papers such as keywords, topics, etc. (Chakraborty et al. 2013).

There have been few studies on the phenomenon of 'late awakening' or 'delayed recognition' of papers. Garfield (1989) was the first to provide examples of such papers. Later, Glänzel et al. (2003) estimated such delayed recognition and unfolded interesting characteristics of this phenomenon. Raan (2004) first coined the term 'sleeping beauty' to refer to papers with delayed recognition. Redner (2005) analyzed a large dataset of papers in Physics, and discovered articles with delayed recognition. More recently, Ke et al. (2015) introduced a parameter-free methodology to identify SBs in science.

Braun et al. (2010) proposed a process of 'induced citations' (i.e., citations adopted from the reference list of a subsequent paper) to study sleeping beauties and their 'princes'. Sun et al. (2015, 2016) argued against thresholding techniques used to identify sleeping and awakening period of a paper, and proposed an *obsolescence vector* for a paper to measure the drastic fluctuation in the citation profile of the paper. However, the obsolescence vectors cannot differentiate between two citation profiles if there is multiplier relationship between their annual citation counts. For example, both the vectors (0, 8, 0, 8, 0, 8, 0, 8, 0, 8) and (0, 4, 0, 4, 0, 4, 0, 4, 0, 4) have the same obsolescence vector. Li and Shi (2015) proposed a set of new criteria based on the increasing rate of the citation profiles to detect genius articles from the articles of Nobel Prize laureates. However, the criteria they proposed also have some ad-hoc selection. For instances, the criteria are not applicable to rarely cited or never cited articles. An article should have received at least 9 citations and at least 90 citations after 10 years and 50 of its publication

respectively to satisfy the criterion. Their criteria are not suitable for publications less than 10 years old.

Li et al. (2014) further quantified two new notions for a sleeping beauty paper—'heartbeat' (annual citations it receives during the sleeping time) and 'heartbeat spectrum' (a vector representing the heartbeat of the paper). They concluded that the papers that possess later heartbeats have higher awakening probability than than those have early heartbeats. Li (2014) and Li and Ye (2012) studied four special cases where sleeping beauties seem to be injured by spindles so that they fall into sleep, and are then awakened by princes. They also chose ad-hoc criteria to identify sleeping beauties—less than two average citations for a certain time period (at least 5 years) and more than 20 citations in the next 4 years. In another study, they Li and Ye (2016) proposed three criteria—average-based criteria, quartile-based criteria and parameter-free criteria, based on which they distinguished sleeping beauties from others.

van Raan (2015) further studied two important properties of sleeping beauties—(1) the time-dependent distribution, author characteristics, journals and fields, and (2) the cognitive environment of sleeping beauties. He studied Physics, Chemistry and Engineering Science papers and observed that half of the sleeping beauty papers are application-oriented. Min et al. (2016) studied delayed reorganization of individual papers and suggested that the principles of delayed citation timelines and final citation numbers should both be considered in order to depict delayed recognition. Again, Sun et al. (2016) and Li et al. (2014) attempted to identify sleeping beauties early after their publications using Gini co-efficient.

It is evident from the above discussion that different studies have used different criteria for identifying SBs. In this paper too, we initially use a set of threshold-based criteria to identify SBs. However, later we design a machine learning model for early identification of SBs based on various features of the papers. To the best of our knowledge, no one attempted to design *machine learning models* to identify SBs early after their publication.

Dataset and identification of sleeping beauties

This section describes the dataset of Computer Science papers, and how we identify SBs from this dataset.

Dataset of Computer Science papers

We use a large dataset of Computer science papers crawled from Microsoft Academic Search (MAS). Specifically, we collected all the papers published in the Computer Science domain indexed by MAS, as of 2012. The dataset contains data for more than 2 million papers. For each paper, the dataset contains the details of the paper (e.g., title, authors, venue and year of publication, keywords), and the names of the other papers that this paper cites. Also, each paper is mapped to one or more sub-fields of Computer Science. In total, there are 24 sub-fields of Computer Science, such as 'Algorithms & Theory', 'Scientific Computing', 'Artificial Intelligence', 'Networks & Communications', and so on, and one or more sub-fields are mentioned for each paper. The reader is referred to Chakraborty et al. (2014) for details of the dataset.

For this study, we focused on the citations during the period 1950 to 2011, for which we have near-complete data. Also, deciding to focus on popular papers, we considered only those 178,383 papers which received at least 20 citations (till 2011).

Computing normalized citation profiles

For each paper p published in the year y_p, we computed the *citation profile* which is a time series; each entry of the time series corresponds to a year t ($t \geq y_p$) indicating how many other papers cited p in year t.

It has been observed that papers in different sub-fields of Computer Science generally get very different number of citations. Hence, in order to meaningfully compare the trends in citation profiles of papers from different sub-fields, we pre-process the citation profiles to make them comparable. First, we smoothen the time-series data points of a citation profile using five-years moving average filtering. Then, we scale the data points by normalizing them with the maximum value present in the time series (i.e., the maximum citations received by the paper in a particular year), so that all values in the normalized citation profile are in the range [0, 1].

Identifying sleeping beauties

Next, we identified SBs from the nomalized citation profiles of the papers. Raan (2004) has proposed three dimensions along which SBs can be identified—(1) the duration of the sleeping period, (2) the depth of the sleep, i.e., the average number of citations during the sleeping period, and (3) the awakening intensity, i.e., the number of citations accumulated during 4 years after the sleeping period. Out of the above three dimensions, we consider only the first two to identify SBs. We do not consider the third dimension since, as we show in the next section, SBs can have very different citation profiles in the years after awakening.

Specifically, we consider a paper to be a SB if all the data points in its normalized citation profile are less than 0.20 during the first 10 years after the publication of the paper. In other words, we focus on papers for which the sleeping period is at least 10 years, and the average number of citations per year during the sleeping period is at most 20% of its maximum peak. Note that we adapted these criteria from a series of our past work (Chakraborty et al. 2014, 2015; Chakraborty and Nandi 2017). By this process, we identified 5,086 papers as SBs (which is 2.85% of all papers in our dataset, that have at least 20 citations).

It is worth mentioning that we actually considered a flexible criteria for most of the cases. For instance, the normalized citation count for first 10 years was considered as 0.20 ± 0.05. The time window after the publication was consider as 10 ± 2 years. The flexibility in the criteria ended up producing more or less the same set of SBs as stated above.

One may argue against our normalization procedure, that if a paper keeps getting growing attention right from its time of publication and it gets so many citations subsequently that the proportion in the early years becomes relatively small, it may mistakenly be identified as a sleeping beauty by our approach (such papers are usually known as 'evergreen' papers, and not sleeping beauties). To cross-check if such papers exist among the identified 5, 086 SBs, we further measured raw citation count of each identified SB during its sleeping time (first 10 years after publication). We did not find any SB receiving more

than 50 citations during its sleeping time. Therefore, we conclude that our normalization method did not mistakenly detect even-green papers as SBs.

Comparing our method for identifying SBs with prior methods

As described in "Related work" section, several prior studies have tried different methods for identifying SBs. We investigate how well the set of 5, 086 SBs identified by our method matches with SBs identified by the methods from Sun et al. (2016) and Li et al. (2014). To this end, we implemented the methods proposed in Sun et al. (2016) and Li et al. (2014) and applied them on our dataset. Table 1 shows a comparative analysis of the set of SBs identified by the different methods, in the form of confusion matrices.

Using the method in Sun et al. (2016), we obtained 169, 209 SBs from our dataset, out of which 4, 671 SBs are common with those identified by our method. On the other hand, we obtain 38, 570 SBs using the method in Li et al. (2014), out of which 465 SBs are common with our set.

These results essentially indicate that our criteria for discovering SBs are more strict than the prior methods. Therefore, we obtain less SBs compared to the other methods. We focus on the 5, 086 SBs identified by our method in the rest of this paper.

Characterizing sleeping beauties

In this section, we characterize the sleeping beauties identified by the methodology stated in the previous section.

Which sub-fields of Computer Science lead to most SBs?

As was stated earlier, each paper in the dataset is mapped to one or more sub-fields of Computer Science. Table 2 (2nd column) shows the distribution of the 5,086 identified SBs across the different sub-fields. The sub-fields of 'Algorithms & theory' and 'Scientific computing' account for more than 50% of the SBs, while 'Artificial intelligence' and

Table 1 Confusion matrix showing a comparative analysis of the sleeping beauties obtained from our method, with those obtained by the methods from - Sun et al. (2016) and Li et al. (2014)

		Sun et al. (2016)	
		Yes	No
(a)			
Our	Yes	4671	415
	No	164,538	–
		Li et al. (2014)	
		Yes	No
(b)			
Our	Yes	465	4621
	No	38,105	–

Table 2 Distribution of SBs across different sub-fields of Computer Science

Sub-field	% SBs	% papers which become SBs
Algorithms and theory	31.67	0.69
Scientific computing	19.01	0.60
Artificial intelligence	12.64	0.21
Natural language andspeech	9.83	0.27
Networks and communications	6.51	0.15
ML and pattern recognition	5.13	0.23
Hardware and architecture	3.85	0.13
Software engineering	3.70	0.12
Data mining	3.13	0.27
Information retrieval	2.91	0.31

The 2nd column states what fraction of the identified SBs come from each sub-field

The 3rd column states what fraction of all papers in a sub-field become SBs

The table is ordered w.r.t. 2nd column, and only top 10 sub-fields are shown

'Natural language & speech' account for another 22%. We also observed that the distribution across the different sub-fields remains almost the same for each of the three sub-classes of SBs (we skip this result in the interest of space).

Table 2 (3rd column) presents what fraction of all papers from a certain sub-field (as included in our dataset) become SBs. Again, a much higher fraction of papers from the sub-fields of 'Algorithms & theory', and 'Scientific computing' become SBs, which is probably because the algorithms/methodologies contributed by these papers later find application in different sub-fields of Computer Science (as detailed in "What leads to awakening of SBs?" section).

Interestingly, though the sub-field of 'Information Retrieval' contributes lesser SBs than many other sub-fields (2nd column of Table 2), the fraction of papers from this sub-field which become SBs is higher than that for many other sub-fields (3rd column of Table 2).

Types of SBs based on citation profiles after awakening

We start by checking whether the citation profiles of different SBs look similar or different after their awakening. For this, we detect *peaks* in the citation profile of a SB by applying the following heuristics: (1) a peak should be a local maxima, with the height on either side being lesser than (or at most equal to) the peak-height, (2) the height of a peak should be at least 70% of the globally maximum peak-height, and (3) two consecutive peaks should be separated by more than 2 years, otherwise they are treated as a single peak.

Interestingly, we observe three different sub-classes of SBs based on the number of peaks in their citation profile after awakening.

1. *Single-peak (SP)* These SBs gradually accumulate citations after awakening, resulting in a peak in the citation profile, which is followed by a deterioration of citation count. This sub-class accounts for 43.8% of all the identified SBs.
2. *Multiple-peak (MP)* The citation profiles of these SBs have multiple peaks separated by few years (37.9% of all SBs).

3. *Monotonically-increasing (MI)* The citation profiles of these SBs continuously rise with time, at least till 2011 (till which we have complete citation data). This sub-class contains 18.3% of the SBs.

Table 3 states few examples of highly-cited SBs from each of the three sub-classes, while the normalized citation profiles of some of these SBs are shown in Fig. 1.

Do SBs eventually get more cited than other types of papers?

We compare the total citation counts of SBs (which were selected from among the papers having at least 20 citations) and that of all 178,383 papers in our dataset, that have at least 20 citations.

Figure 2a shows the distribution (CDF) of total citation counts of the two sets of papers. It is evident that, in general, the SBs eventually get more citations compared to other types of papers. For instance, 25% of the SBs receive 100 or more citations, compared to less than 11% of all papers. Thus, though the SBs get recognized late, many of them eventually become more popular than other types of papers.

We also compare the total citation counts of the three sub-classes of SBs. Figure 2b shows that, among SBs, the monotonically-increasing sub-class generally get the highest number of citations, followed by the multiple-peak, and then the single-peak.

How long do the SBs sleep before awakening?

We compute the *awakening time* of a SB, which indicates after how many years of publication the paper starts getting cited. For this, we use a variant of the methodology proposed in Ke et al. (2015). Let y_0 be the year of publication of a paper, and y_m the year in which the paper obtained the maximum citations. Let c_0 and c_m be the number of citations obtained by the paper in the year y_0 and y_m respectively. Describing graphically, the methodology considers the straight line joining the points (y_0, c_0) and (y_m, c_m) in the citation profile curve, and identifies that point on the curve from which the distance of this line is maximum, as the point at which the paper awakens. We observed a limitation of this methodology—for many of the multiple-peak SBs, this methodology identifies a point *after* the first peak as the point of awakening. Hence, we used a variant of this

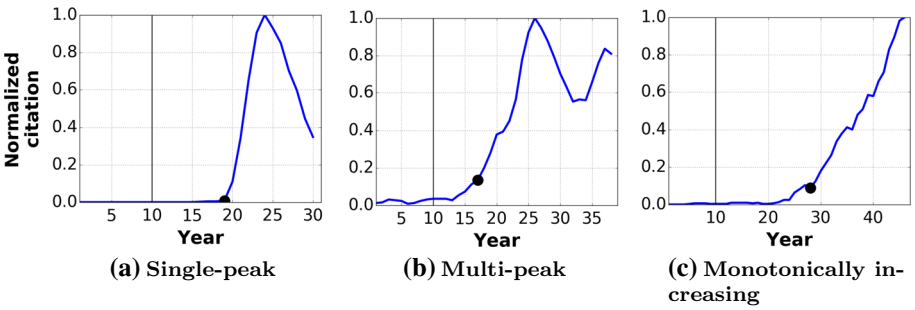

Fig. 1 Normalized citation profiles of some highly-cited SBs, one from each sub-class (for the papers numbered 1, 3, and 5 in Table 3). The black circles denote the year in which each SB 'awakens', computed as described in the text **a** Single-peak. **b** Multi-peak. **c** Monotonically increasing

Table 3 Examples of highly-cited SBs in the three sub-classes

	Paper information	Sub-field
Single-peak (SP)		
1	An Architecture for differentiated dervices (1978)	Networks and communications
2	An intrusion-detection model (1986)	Security and privacy
Multiple-peak (MP)		
3	An efficient heuristic procedure for partitioning graphs (1970)	Hardware and architecture
4	A machine-oriented logic based on the resolution principle (1965)	Algorithms and theory
Monotonically-increasing (MI)		
5	Visual pattern recognition by moment invariants (1962)	Machine learning and pattern recognition
6	A vector space model for automatic indexing (1975)	Data mining, information retrieval

The citation profiles of the papers numbered 1, 3, and 5 are shown in Fig. 1

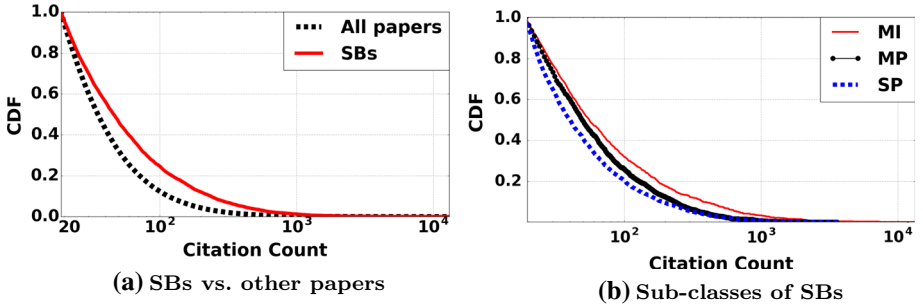

Fig. 2 Comparing the citation-counts of (1) SBs and all papers, (2) the three sub-classes of SBs. Only papers having at least 20 citations are considered in all the sets. In general, SBs eventually achieve more citations than other papers. Among SBs, the MI sub-class achieves the most citations **a** SBs vs. other papers. **b** Sub-classes of SBs

methodology, where we considered the *first* peak (and its year) instead of the point (y_m, c_m). To show the results of this methodology, Fig. 1 shows the point (year) of awakening of the SBs whose citation profiles are shown.

Once the year of awakening of a SB is found, we compute the *awakening time* as the duration between the year of publication and the year of awakening. Figure 3 shows the

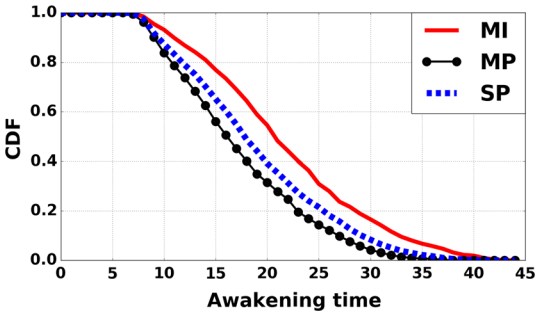

Fig. 3 Distribution of awakening time of SBs (in terms of years after publication after which the paper starts getting cited)

distribution of awakening times. On average, the monotonically-increasing SBs have larger awakening times, compared to the other two sub-classes. Combining Figs. 3 and 2b, we see that though the monotonically-increasing SBs sleep for longer durations, they eventually achieve higher citation counts than the other sub-classes of SBs.

Also note that some of the SBs sleep for durations as long as 40 years or more. An extreme example is a paper by Garfield in 1955, titled "Citation indices for science; a new dimension in documentation through association of ideas". This paper remained in sleep for more than 40 years, after which it suddenly became noticed in 1999 due to a citation from the famous paper by Kleinberg titled "Authoritative Sources in a Hyperlinked Environment". Afterwards, the paper by Garfield received a large number of citations from papers related to impact factor and similar topics.

Finally, out of the 518 SBs which awakened more than 30 years after being published, 50.8% are from just two sub-fields—'Algorithms and theory' and 'Scientific computing'.

In this section, we identified SBs and characterized different sub-classes of SBs based on their citation profiles after awakening. In the next section, we investigate the factors that lead to the awakening of the SBs.

What leads to awakening of SBs?

The most immediate factor leading to awakening of a SB is the citation by another paper, which draws the attention of the research community to the SB. The paper which 'awakens' a SB (i.e., cites the SB and draws attention towards the SB) is often described as the 'prince' of the SB (Raan 2004). We start by identifying the prince of each of the SBs that we identified (as described earlier).

Identifying the princes of the SBs

A paper which awakens a SB (i.e., cites the SB and draws attention towards the SB) is often described as the 'prince' of the SB (Raan 2004). Braun et al. (2010) defined the prince of a sleeping beauty as the paper (1) that first cited the sleeping beauty after a long time of its publication, (2) that is highly or at least fairly cited, and (3) that has considerable number of co-citations with the sleeping beauty. However, exact quantification of the citation and co-citation count of the prince was not mentioned. Here we utilize the same definition with proper quantification of the parameters used to detect princes of SBs. For a particular SB x, we identify as the prince, that paper y which (1) cites x within 3 years before or after the year in which x awakens, and (2) is co-cited with x most number of times after the citation of y to x.

Table 4 shows some examples of SBs and their princes. In general, we observe that the SB is usually a paper which proposes a generic methodology / algorithm (e.g., Rough sets, Viterbi algorithm), while the prince identifies an extension or application of the methodology proposed by the SB, thus drawing attention of the research community towards the SB.

We also observe that once the SB is awakened, the SB and the prince often exhibit different types of citation profiles. To demonstrate this, the citation profiles of the SBs and princes stated in Table 4 are shown in Fig. 4. In some cases, the citation profile of the prince closely follows that of the SB (Fig. 4a). In some other cases, the SB continues to get

Table 4 Examples of SBs and their princes. The citation profiles of these papers are shown in Fig. 4 in the same order

	Sleeping beauty	Prince
(a)	Rough sets (1982)	A Generalized definition of rough approximations based on similarity (2000)
(b)	Multidimensional binary search trees used for associative searching (1975)	The R+Tree: A dynamic index for multi dimensional objects (1987)
(c)	The viterbi algorithm (1973)	A tutorial on hidden Markov models and selected applications in speech recognition (1990)

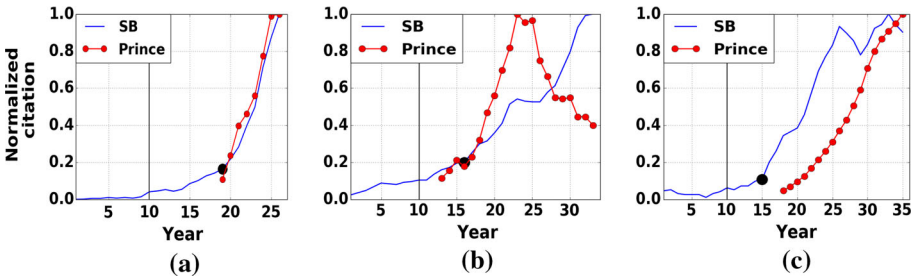

Fig. 4 Examples of citation profiles of some SBs and their princes listed in Table 4

heavily cited even after citations to the prince decay (Fig. 4b), while in a few cases, the citation profile of the prince peaks after that of the SB (Fig. 4c).

Delving deeper into factors leading to awakening of SBs

We now delve deeper into what leads to awakening of a SB. Basically, we attempt to understand why a paper (the prince) would cite another paper published long back which have not been cited much till date (the SB).

As evident from Tables 3 and 4, SBs generally contribute some generic methodology/model/algorithm. Intuitively, there can be two reasons for such a SB getting heavily cited long after publication—(1) the methodology/model proposed by the SB is later extended (e.g., by the prince), after which the extended methodology achieves large popularity, or (2) the methodology/model proposed by the SB is found to be useful in some other sub-field(s) of Computer Science. We now check which of these two reasons can explain the awakening of SBs.

For each SB, we count the number of *distinct sub-fields* of Computer Science, from which it is cited during or after the year it awakens (computed as described in the previous section). Figure 5a shows a wide variation in the number of sub-fields from which a SB is cited. While most of the SBs get cited from 5–7 distinct sub-fields, there are a few SBs which get cited from only one field, and there are some other SBs which get cited from as many as 20 (or more) different sub-fields.

We further quantify the variation of the different sub-fields from which a SB is cited as follows. For each SB s, we compute the distribution c_s of all the citations obtained by s

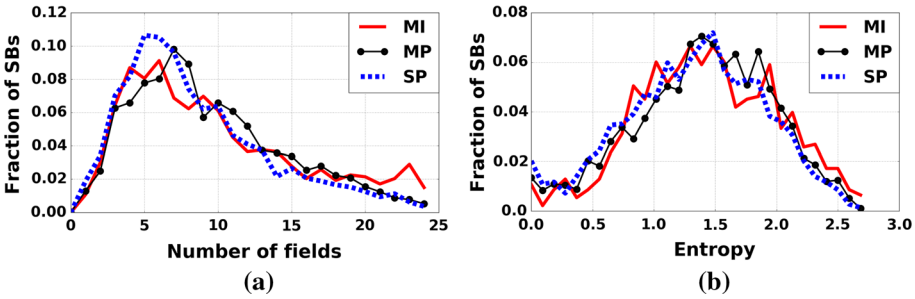

Fig. 5 Papers from how many distinct sub-fields of Computer Science cite a SB after it awakens—**a** variation in the number of sub-fields, **b** variation in entropy of the distribution of the citations from different sub-fields

after awakening, from the 24 sub-fields of Computer Science, and compute the *entropy* $\mathcal{H}(c_s)$ of this distribution.

$$\mathcal{H}(c_s) = -\sum_{i=1}^{24} p(n_i/n) \times log(p(n_i/n)) \tag{1}$$

where n is the total number of citations obtained by s, and n_i is the number of citations obtained by s from the sub-field i (out of the 24 sub-fields available in the dataset). Figure 5b shows the variation of $\mathcal{H}(c_s)$ across all the SBs. We observe that $\mathcal{H}(c_s)$ varies in the range [0.0, 2.76].

Since it is difficult to analyze the reason of awakening of each SB on a case-by-case basis, we focus on two extreme sub-sets of the SBs—the SBs with $\mathcal{H}(c_s) = 0.0$, and the SBs with $\mathcal{H}(c_s) \geq 2.5$. In our dataset, both these sub-sets comprise of 77 SBs. Table 5 shows some examples of both types of SBs, and also presents the distribution of the two sets across different sub-fields of Computer Science (only the top 5 sub-fields stated, for interest of space).

SBs with $\mathcal{H}(c_s) = 0.0$: These SBs get all their citations from only one sub-field. 76.6% of these SBs are from the sub-field of 'Natural language and speech'. Clearly, the contributions of these SBs were later extended, after which the contributions achieved popularity within that sub-field.

SBs with $\mathcal{H}(c_s) \geq 2.5$: The examples in Table 5 show that these SBs contribute some generic methodologies—e.g., optimization methodologies for graphs, calculation of Fourier series—which later find applicability in different sub-fields of Computer Science. Hence, once their contribution is recognized, they get cited from many different sub-fields (usually more than 20). Most of these SBs are from the sub-fields 'Scientific computing' and 'Algorithms & theory'. Note from Fig. 5a that many of these SBs fall under the monotonically-increasing sub-class, possibly because new applications of their contributions are regularly found.

Thus, we conclude that the SBs usually contribute some methodologies / models, which are either later extended (e.g., by the prince) so as to achieve higher utility in the same sub-field, or the methodology is found to be applicable in other sub-fields, which lead to the awakening of the SBs. We leave as future work a more detailed investigation of other potential reasons for awakening of SBs.

Table 5 Comparing SBs that are cited, after awakening, from only one sub-field (with $\mathcal{H}(c_s) = 0.0$), and SBs (with $\mathcal{H}(c_s) \geq 2.5$) cited from 20 or more distinct sub-fields of Computer Science

SBs with $\mathcal{H}(c_s) = 0.0$ (cited from only one sub-field)	SBs with $\mathcal{H}(c_s) \geq 2.5$ (cited from 20 or more distinct sub-fields)
Distribution of SBs across sub-fields	
Natural lang and speech: 76.6%	Scientific computing: 25.5%
Hardware and arch: 7.8%	Algorithms & theory: 18.9%
Algorithms and theory: 3.9%	Data mining: 16.0%
Scientific computing: 2.6%	Natural lang & speech: 16.0%
Networks and commn: 2.6%	Databases: 5.7%
Examples of SBs	
Auditory filter shapes in subjects with unilateral and bilateral cochlear impairments (natural lang and speech)	An algorithm for the machine calculation of complex fourier series (scientific computing)
Consonant confusions in noise: a study of perceptual feature (natural lang and speech)	Optimization algorithms for networks and graphs (algorithms and theory)
Functional decomposition and switching circuit design (hardware and arch)	Minimum spanning tree and single linkage cluster analysis (data mining)

Early identification of sleeping beauties

We now turn to the question of early identification of SBs. In this section, we attempt to develop a methodology by which we will be able to predict whether a paper is likely to become a Sleeping Beauty. Ideally, we would like to do this prediction early, i.e., as soon after the paper is published as possible.

We model the problem of identification of SBs as a binary classification problem, where we attempt to distinguish between two classes of papers—SBs and non-SBs—based on a set of features derived from each paper. We start by describing the features used, followed by the classification accuracy.

Features used for distinguishing between SBs and non-SBs

We consider each paper p (published at time t_p) and classify it as SB or non-SB at time t (where $t > t_p$). We leverage the meta-data information of p present in our dataset to extract the following features:

1. *Number of keywords* (Keywd) Total number of keywords assigned to the paper.
2. *Number of research fields of the paper* (PaperField) The number of sub-fields (of Computer Science) associated with p.
3. *Number of authors* (Auth) Total number of authors who wrote the paper.
4. *Average number of publications per author* (PaperAuth) Average number of papers published by an author of p till t.
5. *Average citations received per author* (CiteAuth) Average number of citations received by an author of p till t.
6. *Number of unique research fields of authors* (FieldAuth) Number of distinct sub-fields (of Computer Science) in which the authors of p have published papers.

7. *Number of references* (RefCount) Number of papers that have been cited by p.
8. *Number of research fields of referenced papers* (FieldRef) Number of distinct sub-fields from where papers have been cited by p.
9. *Average citations received by cited papers* (CiteCitedPaper) Average number of citation count of the papers that have been cited by p, at t.
10. *Citations received till the prediction year* (CiteCount) Number of citations received by p till time t.
11. *Entropy of the number of fields from where citations have been received* (EntropyF) Here we consider the set of papers which have cited p (till time t), and compute the distribution of these papers across the 25 sub-fields. Finally, we compute the entropy of the said distribution.
12. *Type of the venue* (VenueType) Type of the venue (conference/journal) where p was published.
13. *Number of papers published in the venue* (PaperVen) Number of papers published in the venue till t, where p was published.
14. *Average citations received by papers published in the venue* (CitePaperVen) Average number of citations received by the papers published in the venues till date, where p was published.

It can be noted that the above features are broadly of the following five types: (1) characteristics of the paper p itself (features 1 and 2), (2) characteristics of the authors of p, e.g., how frequently they publish, in how many research fields they publish, etc. (features 3–6), (3) characteristics of the papers that have been cited by p (features 7–9), (iv) characteristics of the papers that have cited p (if any) till time t (features 10–11), and (v) characteristics of the venue where p was published (features 12–14).

The features that consider the number of distinct sub-fields of Computer Science (or the entropy of the corresponding distribution across the sub-fields) related to p / authors of p / papers cited by p or papers citing p, attempt to estimate how inter-disciplinary p is—these features are motivated by our observation that a significant number of sleeping beauties are inter-disciplinary in nature (as reported in the earlier sections). Again, the features related to the venue where p was published are meant to check whether papers published at more popular venues are more likely to become SBs.

Classification experiments and results

For the purpose of our experiment, we consider 2300 SBs published during the period 1960-1980 and 2300 randomly selected non-SBs published during the same period. Then

Table 6 Confusion matrices showing the classification performance of three classifiers for the year 1995

	Linear SVM		Decision tree		KNN	
	Non-SB	SB	Non-SB	SB	Non-SB	SB
Non-SB	1913	387	1927	373	1821	479
SB	798	1502	874	1426	880	1420

We observe that SVM performs the best in detecting the SBs (precision of 0.80 and recall of 0.72) compared to other classifiers

The pattern is same for the other years and therefore skipped for the sake of brevity

Table 7 Confusion matrices of SVM classifier for different years

	1980		1983		1985		1987		1990		1995	
	Non-SB	SB	Non-SB	SB	Non-SB	SB	Non-SB	SB	Non-SB	SB	Non-SB	SB
Non-SB	1920	380	1913	387	1905	395	1908	392	1899	401	1913	387
SB	1254	1046	1233	1067	1226	1074	1206	1094	1062	1238	798	1502

Table 8 Accuracy (in terms of precision, recall, and F-score) of the SVM classifier in predicting both Non-SB and SB papers, at different years after their publication

Year	Class	Precision	Recall	F-Score
1980	Non-SB	0.60	0.83	0.70
	SB	0.73	0.45	0.56
	Avg/total	0.67	0.64	0.63
1983	Non-SB	0.61	0.83	0.70
	SB	0.73	0.46	0.57
	Avg/total	0.67	0.65	0.64
1985	Non-SB	0.61	0.83	0.70
	SB	0.73	0.47	0.57
	Avg/total	0.67	0.65	0.64
1987	Non-SB	0.61	0.83	0.70
	SB	0.74	0.48	0.58
	Avg/total	0.67	0.65	0.64
1990	Non-SB	0.64	0.83	0.72
	SB	0.76	0.54	0.63
	Avg/total	0.70	0.68	0.68
1995	Non-SB	0.71	0.83	0.76
	SB	0.80	0.65	0.72
	Avg/total	0.75	0.74	0.74

Table 9 Feature importance for SVM classifier

We arrange features in decreasing order of the Chi Square value

Detailed description of features can be found in "Features used for distinguishing between SBs and non-SBs" section

Feature	Chi Square statistics
EntropyF	3.21299393e+02
PaperVen	2.41498282e+02
VenueType	2.01301782e+02
CiteAuth	3.73830867e+01
FieldAuth	2.25142337e+01
CiteCount	1.75613771e+01
PaperAuth	1.31816098e+01
PaperField	2.50422749e+00
CitePaperVen	1.36658778e+00
Auth	9.60444172e−01
FieldRef	8.72222222e−01
RefCount	7.18300271e−01
Keywd	1.33524534e−01
CiteCitedPaper	5.58120293e−03

we attempt to classify the papers as SB or not, at the years 1980, 1983, 1985, 1987, 1990 and 1995. Essentially, we intend to investigate how early we can predict if a paper would turn out to be a SB. Note that all time-varying features are re-computed at each year of prediction. For each year, we consider a 10-fold cross validation approach.

We measure the accuracy of classification separately for each class (SB and non-SB) using the following metrics:

- *Precision* Out of the papers detected as a certain class by a classifier, what fraction of them are correctly classified.
- *Recall* Out of the papers that originally fall into a certain class, what fraction of them are classified correctly by a classifier.
- *F-Score* Harmonic mean of precision and recall.

We use three classifiers—decision tree, KNN and linear SVM. All hyper-parameters are optimized using grid search. Table 6 reports the confusion matrices of all the classifiers for the year 1995. We notice that SVM outperforms the other classifiers in detecting SBs both in terms of precision and recall. The pattern is same for the other years. Therefore in the rest of the paper, we will report the accuracy of only SVM for different years.

Table 7 shows the detailed classification results using SVM for different years. A summary of the classification accuracy in terms of precision, recall, and F-score is also given in Table 8. We observe that with the increase of time, precision for both SB and Non-SB increases significantly. On the other hand, while recall remains almost same for Non-SB, it increases significantly for SB. The Recall value at 1995 for SB is 0.65 which is 35% higher than the recall at 1980.

This improvement in classification with time is intuitive as with the progress of time, we collect more evidences about the SBs which lead to detect them more accurately. However, predicting SBs immediately around the time of publication yields an F-Score of 0.56 which may be considered significant due to extremely less evidences at the early period.

Feature importance

We further measure the importance of each feature for identifying SBs. Table 9 shows the Chi Square value of each feature for the SVM classifier. The features are ranked in decreasing order of the Chi Square value.

We observe that the entropy of the number of fields from where the target paper has received citations is the most important feature. This result corroborates with our idea mentioned in "Delving deeper into factors leading to awakening of SBs" section that the more the paper has potential to attract attention from multiple fields, the more the probability that it qualifies as an interdisciplinary paper that can become popular eventually. The second-ranked feature is the type of publication venue. Our previous study - Chakraborty and Nandi (2017) showed that it is highly likely that a SB is published in a journal. Here also we notice the similar trend. The third-ranked feature is the average citations received by the authors of the paper. We hypothesize that if a highly-cited author writes a paper, it will eventually be noticed by the research community.

On the other hand, the reference count of the target paper, and the number of keywords turn out to be the least important features. This is also intuitive since, if reference count of a paper was an indicative factor, then all the survey articles would become the sleeping beauties, which is not true in reality. The keyword count is also same across papers, thus can not be an important factor to identify SBs.

Conclusion

We performed an empirical analysis of a massive Computer Science publication dataset to understand and predict delayed recognition *aka* sleeping beauties. We identified more than 5,000 SBs across various sub-fields in Computer Science. We characterized these SBs

based on their citation profiles after awakening, and the number of different sub-fields from which they get cited. The major findings of the paper are as follows:

- A sleeping beauty can be characterized by its awakening time (time to get sufficient citations after publication) and awakening intensity (number of citations received by the paper immediately after it awakens).
- Sleeping beauties are more prominent in the fields like Algorithms & Theory, Scientific Computing, the reason being that the algorithms/methods contributed by these papers are often found to be useful much later after their publications, and in different fields of Computer Science.
- We classified sleeping beauties into three sub-categories based on the peaks in their citation profiles—single-peak, multiple-peak and monotonically-increasing.
- We noticed that although sleeping beauties get delayed recognition, they (mostly the monotonically-increasing category) eventually become more popular compared to the other types of papers.
- We further noticed that monotonically-increasing sleeping beauties tend to sleep more which in turn leads to gain more citations than other types of sleeping beauties.
- We studied the princes of sleeping beauties and showed that although the prince causes the sleeping beauty to awaken, the overall citation profile of both of them is significantly different.
- Deeper investigation revealed that there is no uniform pattern of the number of distinct fields which cite a sleeping beauty when it awakes. Some sleeping beauties receive citations from a concentrated number of fields; while for the other sleeping beauties, the citations are dispersed across many fields.
- Finally, we designed predictive models to identify SBs early after their publication and obtained a F-Score of 0.70 with very little evidences of the citation history of the papers. To our knowledge, this is the first attempt to use machine learning models for predicting SBs in Computer Science.

In future, we would like to develop a model on citation networks to explain the phenomenon of SBs, which cannot be explained by standard models like preferential attachment (Ke et al. 2015). We envisage utilizing the model to develop better methodologies for early identification of SBs. We have made the code publicly available for the reproducibility purpose at https://github.com/ranarag/SleepingBeauties.

References

Barabasi, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509–512.

Braun, T., Glänzel, W., & Schubert, A. (2010). On sleeping beauties, princes and other tales of citation distributions. *Research Evaluation*, *19*(3), 195–202. doi:10.3152/095820210x514210.

Chakraborty, T., Kumar, S., Goyal, P., Ganguly, N., & Mukherjee, A. (2014). Towards a stratified learning approach to predict future citation counts. In *Proceedings of ACM/IEEE-CS joint conference on digital libraries* (pp. 351–360).

Chakraborty, T., Kumar, S., Goyal, P., Ganguly, N., & Mukherjee, A. (2015). On the categorization of scientific citation profiles in computer science. *Communications of the ACM*, *58*(9), 82–90.

Chakraborty, T., Kumar, S., Reddy, M. D., Kumar, S., Ganguly, N., & Mukherjee, A. (2013). Automatic classification and analysis of interdisciplinary fields in computer sciences. In *Proceedings of international conference on social computing (SocialCom)* (pp. 180–187).

Chakraborty, T., & Nandi, S. (2017). Universal trajectories of scientific success. *Knowledge and Information Systems*. doi:10.1007/s10115-017-1080-y.

Garfield, E. (1989). Delayed recognition in scientific discovery: Citation frequency analysis aids the search for case history. *Current Contents*, *23*, 3–9.
Garfield, E. (1999). Journal impact factor: A brief review. *Canadian Medical Association Journal*, *161*(8), 979–980.
Garfield, E. (2001). Impact factors, and why they won't go away. *Nature*, *411*(6837), 522.
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.
Ke, Q., Ferrara, E., Radicchi, F., & Flammini, A. (2015). Defining and identifying sleeping beauties in science. *PNAS*, *112*(24), 7426–7431.
Li, J. (2014). Citation curves of all-elements-sleeping-beauties: Flash in the pan first and then delayed recognition. *Scientometrics*, *100*(2), 595–601. doi:10.1007/s11192-013-1217-z.
Li, J., & Shi, D. (2015). Sleeping beauties in genius work: When were they awakened? *Journal of the Association for Information Science and Technology*, *67*(2), 745–757. http://dblp.uni-trier.de/db/journals/scientometrics/scientometrics107.html#SunML16.
Li, J., Shi, D., Zhao, S. X., & Ye, F. Y. (2014). A study of the heartbeat spectra for sleeping beauties. *Journal of Informetrics*, *8*(3), 493–502. doi:10.1016/j.joi.2014.04.002.
Li, J., & Ye, F. Y. (2012). The phenomenon of all-elements-sleeping-beauties in scientific literature. *Scientometrics*, *92*(3), 795–799. doi:10.1007/s11192-012-0643-7.
Li, J., & Ye, F. Y. (2016). Distinguishing sleeping beauties in science. *Scientometrics*, *108*(2), 821–828. doi:10.1007/s11192-016-1977-3.
Min, C., Sun, J., Pei, L., & Ding, Y. (2016). Measuring delayed recognition for papers: Uneven weighted summation and total citations. *Journal of Informetrics*, *10*(4), 1153–1165. doi:10.1016/j.joi.2016.10.001.
Raan, A. F. J. V. (2004). Sleeping beauties in science. *Scientometrics*, *59*(3), 461–466.
Redner, S. (2005). Citation statistics from 110 years of physical review. *Physics Today*, *58*(6), 49–54.
Ruiz-Castillo, J. (2013). The role of statistics in establishing the similarity of citation distributions in a static and a dynamic context. *Scientometrics*, *96*(1), 173–181. doi:10.1007/s11192-013-0954-3.
Solomon, D. J., Laakso, M., & Bjrk, B. C. (2013). A longitudinal comparison of citation rates and growth among open access journals. *Journal of Informetrics*, *7*(3), 642–650. doi:10.1016/j.joi.2013.03.008. http://www.sciencedirect.com/science/article/pii/S175115771300028X.
Sun, J., Min, C., & Li, J. (2015). A vector for measuring obsolescence of scientific articles. In *Proceedings of international society of scientometrics and informetrics conference*.
Sun, J., Min, C., & Li, J. (2016). A vector for measuring obsolescence of scientific articles. *Scientometrics*, *107*(2), 745–757. http://dblp.uni-trier.de/db/journals/scientometrics/scientometrics107.html#SunML16.
van Raan, A. F. J. (2015). Dormitory of physical and engineering sciences: Sleeping beauties may be sleeping innovations. *PLOS ONE*, *10*(10), 1–38. doi:10.1371/journal.pone.0139786.
Wallace, M. L., Larivire, V., & Gingras, Y. (2009). Modeling a century of citation distributions. *Journal of Informetrics*, *3*(4), 296–303. http://dblp.uni-trier.de/db/journals/joi/joi3.html#WallaceLG09