

Chapter 12

Identifying Sleeping Beauties in the Lore of Regional Science

Rajendra Kulkarni and Roger R. Stough

12.1 Introduction

The Scientific world has seen an unparalleled growth in new knowledge creation over the last several decades. Growth rates from 1% just a couple of centuries ago to over 8 and 9% per year by 2012 have been estimated (Bomann and Mutz 2014). Compared to natural sciences, the field of regional science is relatively young, and yet the new knowledge creation rates are comparable to other scientific disciplines. Given this extraordinary growth, some researchers have been investigating a not-so-rare phenomenon of what Van Raan (2004) refers to as “sleeping beauty” for research that goes unnoticed for a long duration.

If “sleeping beauties” are not uncommon in the scientific literature, then, regional science is likely to have its own share of such research. In this paper, we present a novel method that combines network analysis with text and bibliometric analytics and that does not explicitly depend on citation patterns to help identify “sleeping beauties” in regional science. Additionally, we explore how changes in published scholarly work may change over the next few decades and how some of those changes are likely to affect future sleeping beauties.

R. Kulkarni (✉)
George Mason University, Fairfax, VA, USA
e-mail: rkulkarn@gmu.edu

R.R. Stough
Schar School of Policy and Government, George Mason University, Fairfax, VA, USA
e-mail: rstough@gmu.edu

12.2 Evolution of Scientific Literature and the Concept of Sleeping Beauties

Collective scientific knowledge—when measured in terms of publications, appears to be incremental, i.e., new scientific ideas are built on top of the existing knowledge base, which in turn traces its roots to older research. Mostly a linear view of knowledge building is accepted as the norm with rare exceptions where truly path breaking research emerges that takes the discipline or the field of research by storm and leads the field in a different direction. And then there are those—apparently not-so-rare—examples of research that lay dormant or were ignored by the scientific community at the time they were published but were discovered much later for their true significance and contribution to the field of research. Such works of published scientific work with delayed recognition have been referred to as “sleeping beauties” of science (Crassey 2015).

This delayed recognition phenomenon raises such questions as: Why does this happen?; Was the research premature for its time?; and How to quantify research with delayed recognition. Since the mid-1970s, there has been growing interest in the fields of science of sciences, scientometrics, infometrics, etc. that is devoted to researching all aspects of scientific research, including citation and co-citation analysis, bibliometric coupling, co-occurrence and co-authorship network analysis. Garfield (1970) and Price (1976) tried to answer the questions about why it happens and pointed to possible causes such as poor communication skills on the part of authors. In the past, Garfield (1980) and others have discussed issues related to premature research, or research before its time.

As mentioned in the introduction section, van Raan in 2004 developed a methodology to quantify “occurrence of sleeping beauties” in the science literature (Raan 2004). Using the bibliometric approach, he identified different phases of the phenomenon, such as deep sleep, duration of sleep and awakening duration and was able to identify case of sleeping beauty and awakening prince. Qing K et al. (2015) analyzed over 22 million pieces of published literature and developed a “sleeping beauty” coefficient that can define and help identify such publications. The main thrust of both these studies was about gathering quantifiable information for very large number of published materials and analyzing these based on temporal citation distributions. In other words, if a published scientific work has no citations, then it would not be part of sleeping beauty analyses, which filters out very large portions of published knowledge. As the number of avenues for publishing has increased, the citation rates have not, according to Larivière and Gingras (2015).

A vast majority of published research receives fewer than five citations and often no citations. Remler, in her blog (2014), observes that depending on the field or discipline, the number of articles that get zero citations varies from 12% in medicine, 27% in natural science, 60% in social science and a whopping 82% in humanities. Thus, on average almost 45% of the published articles are never cited. In that case, any analyses of identifying sleeping beauties based only on citation frequencies is likely to miss a large number of sleeping beauties.

In the following sections, we explore a novel method that does not depend on citation counts and can be used to identify potential sleeping beauties within a field or discipline using just the bibliometric data for that field or discipline. Specifically, we will explore how one could find potential sleeping beauties in the field of regional science.

In the next section, we briefly describe bibliometric data used for this exploratory analysis, followed by a section on the specifics of the methodology, which will be followed by couple of case studies that help identify sleeping beauties in the field of regional science.

12.3 Citation Indexing Databases and Data Extraction

The data for the analysis was extracted from Thomson Reuter’s Web of Science (WoS). The initial attempt to conduct a topic (key word or phrase) search using “regional science” and “regional analysis” had to be abandoned since we discovered that those phrases are not unique to the field of regional science but have been used by life sciences; especially in biomedical fields such as radiology, nuclear medicine, neuroscience, neurology, MRI analysis, image analysis, etc.

Instead of topic searches we generated a list of journals that publish regional science research and extracted articles based on journal names. (See Table 12.1 for a sample list of journals.) This process generated a total of over 17,000 records (WoS July 17, 2015). Of these, about 300 records were discarded due to missing information in some of the fields. Of the remaining 16,700, just over 55% (9200) of records representing published scientific research had one or more citations since the publication date (Fig. 12.1). The flip side of this is that nearly 45% of articles (7495) had zero citations (Fig. 12.2). In theory, one could use Google Scholar as an alternate source, and in some fields Google Scholar does have better citation counts

Table 12.1 Some of the journals used to extract data from Web of Science

Regional Science Journals	
Papers in Regional Science	International Regional Science Review
Journal of Regional Science	Journal of Regional Analysis and Policy
Regional Science Policy and Practice	Journal of Urban Economics
Applied Spatial Analysis and Policy	Papers in Regional Science
The Annals of Regional Science	Regional Science and Urban Economics
Regional Studies, Regional Science	Regional Studies
Spatial Economic Analysis	Review of Regional Studies
Annals of Regional Science	Growth and Change
Canadian Journal of Regional Science	Review of Urban and Regional Development Studies

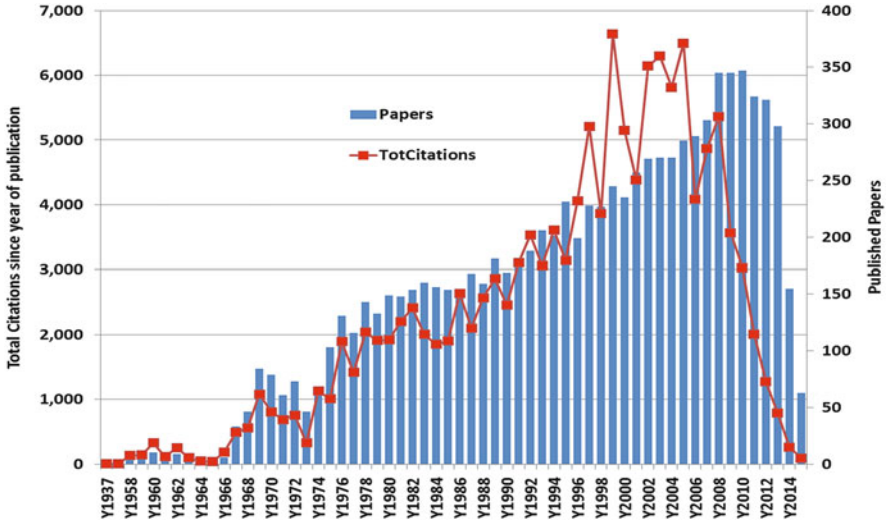


Fig. 12.1 Regional science papers from WoS: Articles and total citations by year of publication

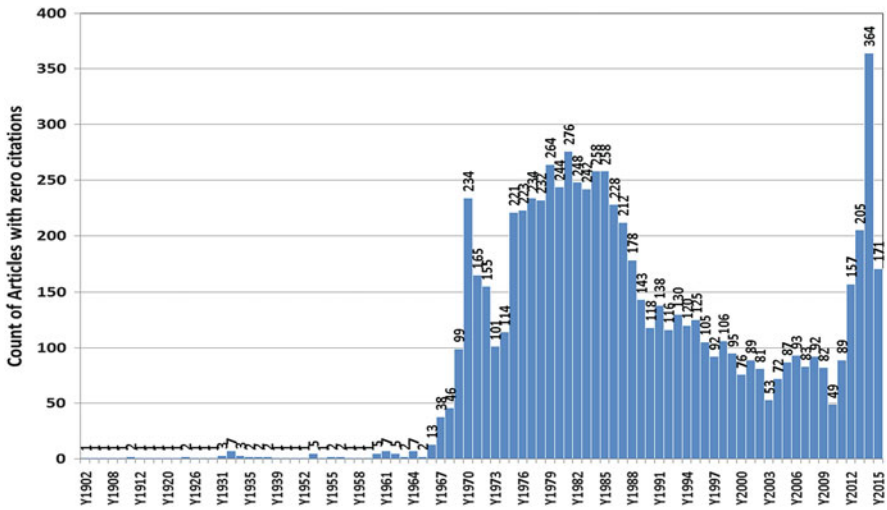


Fig. 12.2 Papers with no citations by year of publication

(Anne-Wil 2016); however, the percent of published papers with zero citations is likely to remain high.

In the following section, we briefly discuss the burst detection analysis that may be useful to identify sleeping beauties.

12.3.1 Using Burst Analysis to Identify Potential Sleeping Beauties

Kleinberg (2002) developed a burst detection and analysis algorithm to “extract meaningful structure from document streams that arrive continuously over time”. By analogy, one could think of citations received by published articles over time as streams of information and apply burst analysis to identify structural changes in the citation patterns, such as low or no citations to a sudden burst of citations for a time period.

One could also apply a burst algorithm to detect sudden fad or popularity of certain keywords/terms associated with scientific articles. Both of these methods were used to detect bursty articles as well as bursty terms by applying the burst detection algorithm to the regional science articles. This particular analysis raises two issues:

- Use of the burstiness property of citation patterns by definition ignores nearly half (45%) of published articles that have zero or no citations. By definition, burst is a sudden change from few citations to more citations one may safely assume that all articles that received between 1 and 3 citations won't be considered for the burst analysis either. In other words, nearly, two-thirds (62%) of articles are ignored by citation burst analysis and, therefore, detection of sleeping beauties is limited only to articles that have more than a certain number of citations.
- The burstiness property of terms identifies sudden popularity, or a fad, of using certain terms over time, however, that does not address the main objective of identifying specific articles that contributed to a term with high burstiness. In theory, one could develop another set of analyses to identify the contribution of each article towards burstiness and then figure out which article started that trend. That part of the analysis will be explored in future work.

12.3.2 Use of Multimode Network and Text Analytics to Detect Sleeping Beauties

Typically, each full record W_i , where i is the index into number of records, extracted from WoS, has four text attributes (fields): Title T_i , Abstract Ab (if present), a set of Key words KW (if present) and scientific classification C_i (one or more). These

four elements of text fields are converted into a vector X_i of N-gram tokens of length 1–4, such that $X_i = [W_i(t1), W_i(t2), W_i(t3), \dots, W_i(tn)]$. A simplistic example will illustrate how a title of a paper such as “Industry and Firm Location Analysis by Regions,” is turned into N-gram tokens. Each word in the title except common words such as “and” and “by,” etc. is called a uni-gram or 1 gram. A sequence of two words would be bi-gram or 2-gram such as “Industry Firm”, “Firm Location”, etc. In the same vein, a 3-gram consists of 3 words and 4 gram consists of 4-words.

Each full record W_i also has other attributes, such as cited references R_f , year of publication P_y , times cited T_c , authors A_u , affiliation A_f , journal details J_d , unique I_d , etc. For an article W_i , let Y_i refer to a vector whose members are the published article and all of the cited references associated with that article. We used software package Citespace (2015) to build a multi-mode,¹ mixed direction network consisting of co-cited articles and references as well as co-occurrence N-gram tokens. The co-cited connections and co-occurrence N-gram are not directed, while the N-gram tokens associated with articles are connected to articles by directed edges from N-gram tokens to articles.

After generating a network consisting of P articles, R references and T n-gram tokens, we computed network statistics such as node degree, weighted node degree and node centrality measure called “betweenness” index. The node degree refers to number of connections to its immediate neighbors while weighted node degree refers to weighted sum of connections to its immediate neighbors. The main difference between a degree and weighted degree of a node is that if node “A” has two connections to node “B” and one connection to node “C,” then the weighted degree of node “A” is 3, while degree of node “A” is 2. The node betweenness index refers to how often a node is visited for an abstract set of paths from each node to every other node. Thus, a node with high betweenness index is more central to the network than a peripheral (or leaf) node that is either a “start” or an “end” node for a path from (to) that node to (from) other nodes in the network.

12.4 Network Analysis and Sleeping Beauties

In theory, a network consisting of 16,700 WoS articles (P) with over 300 K reference records (R) and nearly 150 K tokens from terms T forms a humongously large multi-mode mixed direction network consisting of nearly a half million nodes and

¹When a network consists of just one type of node and edges connecting these nodes, then such a network is referred to as a uni-modal network. If a network consists of two types of nodes, then it is referred to bi-modal or multi-modal network. Depending on whether edges connecting nodes have directions or a combination of directional and non-directional edges, the network would be called as having directional and mixed directional connectivity respectively.

at least many magnitudes more of edges between these nodes. However, due to computing limitations, we built a network consisting of 33 K nodes and almost 350 K edges with a diameter of 12 hops and an average path length 3.5; and over 600 million shortest paths. The diameter of 12 refers to the longest path from over 600 million shortest paths; in other words, the longest shortest path of 12 refers to 12 connections (also called as “hops”) between the farthest nodes in the network, while on average the shortest path (hops) is of 3.5 connections. Next, we computed a weighted node degree and node centrality betweenness index. The following is a short list of token nodes with the highest centrality betweenness index: *Growth, Regional Study, Regional Policy, Regional Development, Economic Development, Cities, Location, Metropolitan Area, Economic Growth, Regional Science, Employment, Urban Area, Regional Growth, Migration, Agglomeration, Unemployment, Demand, Trade, Local Governments, Competition, Increasing Return*, etc.

Tokens with a high betweenness index are central to the network consisting of articles and tokens. For the current findings, we randomly chose a few of these high betweenness index token nodes and built their 1 to 2 “hops” ego-networks where the token of interest and its immediate 1 and 2 hop neighbors form an ego-network. The concept of ego-networks comes from the field of social network analysis where the ego-network of node A consists of all the nodes that are directly connected to node A and that direct connection indicates that they have a close social relationship with node A. This idea of close relationships among nodes with ego-node will be used to identify potential sleeping beauty candidates.

When a set of articles that are part of a token’s ego-network are sorted by the date (year) of publication, then the article with the earliest publication date is likely to be a sleeping beauty provided that, several years had lapsed since its publication and that:

1. It has been not been cited
2. Or cited infrequently
3. Or lay dormant for a long duration and then received large number of citations for next several years.

As mentioned before, for a few of these token nodes with very high betweenness centrality; their ego networks were constructed, i.e., the token node of interest and all the nodes that are its immediate neighbors and then searched for a node that happened to have the earliest year of publication (Fig. 12.3).

The Hotelling article is not part of the published papers data (P), however, it has been part of cited references (R). We were able to find citation analysis for the article from the WoS Core database. The chart below shows the output. It is clear

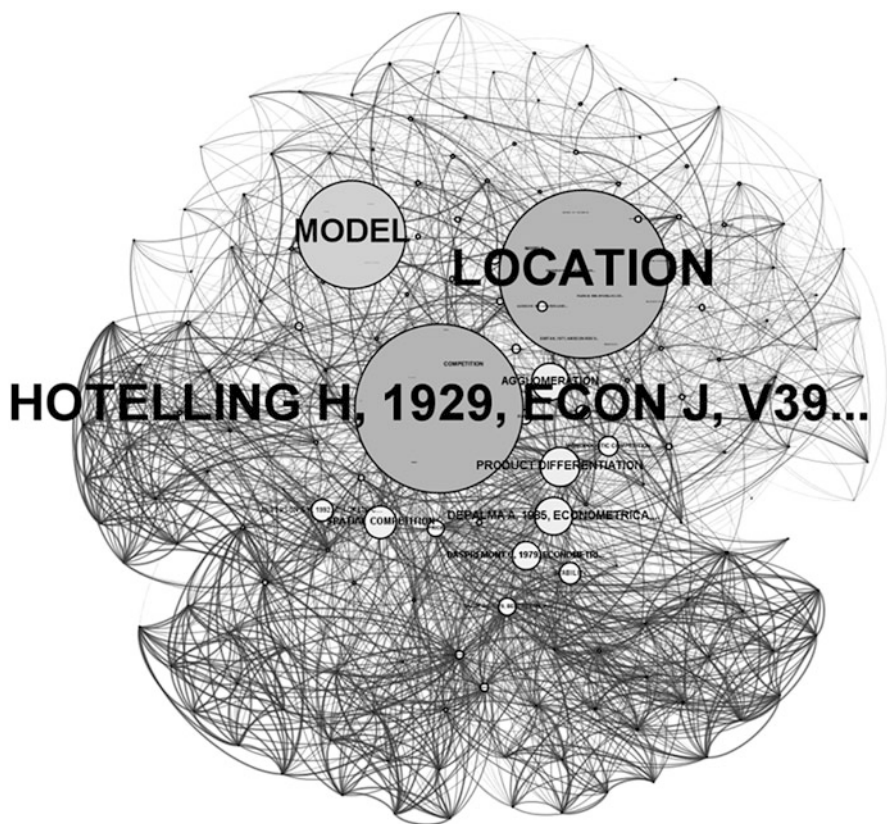


Fig. 12.3 Ego-network of token “Location” and its sleeping beauty, “Hotelling, 1929, Stability in Competition, *Journal of Economics*, V39 N33, DOI 10.2307/2224214”

from the citation analysis shown in Fig. 12.4 that Hotelling’s article was hardly cited for a few dozen years but then garnered a large number of citations much after its publication. This citation pattern fits our definition of sleeping beauty (Figs. 12.5, 12.6, 12.7, 12.8, and 12.9).

We present ego-network analysis for a couple more terms and show associated sleeping beauties. In each of these cases, the sleeping beauties identified were not part of the published paper dataset (P) but were inferred from the ego-network analysis since these were part of the cited reference set (R).

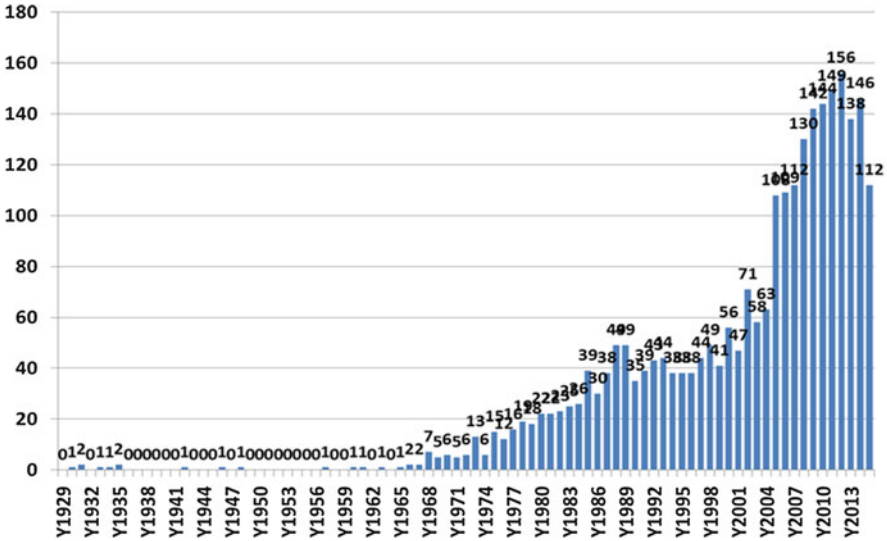


Fig. 12.4 Citation patterns for the Hotelling, 1929, Stability in Competition, *Journal of Economics*, V39 N33, DOI 10.2307/2224214 article

The published article dataset (P) that was used for the analysis does not have a record for Alfred Marshall’s work, however, it appears to be in cited references (R) for quite a few articles. Google scholar search reveals an uneven citation history and it appears that the work remained relatively uncited during early 1900s. Even though, this work appears to be part of ego-network of “Growth,” conventional wisdom says that regional science researchers do not associate Alfred Marshall’s work with the literature on “Growth.” It may be that this spurious association between a token term and scholarly work is result of co-citation by scholarly works that are related to “Growth.” In future, we will have to analyze number of different data samples to determine whether an association exists consistently across different samples.

12.5 Conclusions

Based on our current analysis of over 15,000 published papers from 1902 to 2015 in the field of regional science; a multi-modal network analysis combined with extracted ego networks of nodes with high betweenness centrality index scores can be useful to identify potential sleeping beauties in the field. After constructing a multi-modal network from co-cited published papers (P), references (R) and co-occurrence of tokens (T), a nodal betweenness centrality measure is computed. As stated before, the betweenness centrality of a node is a proxy for gauging how

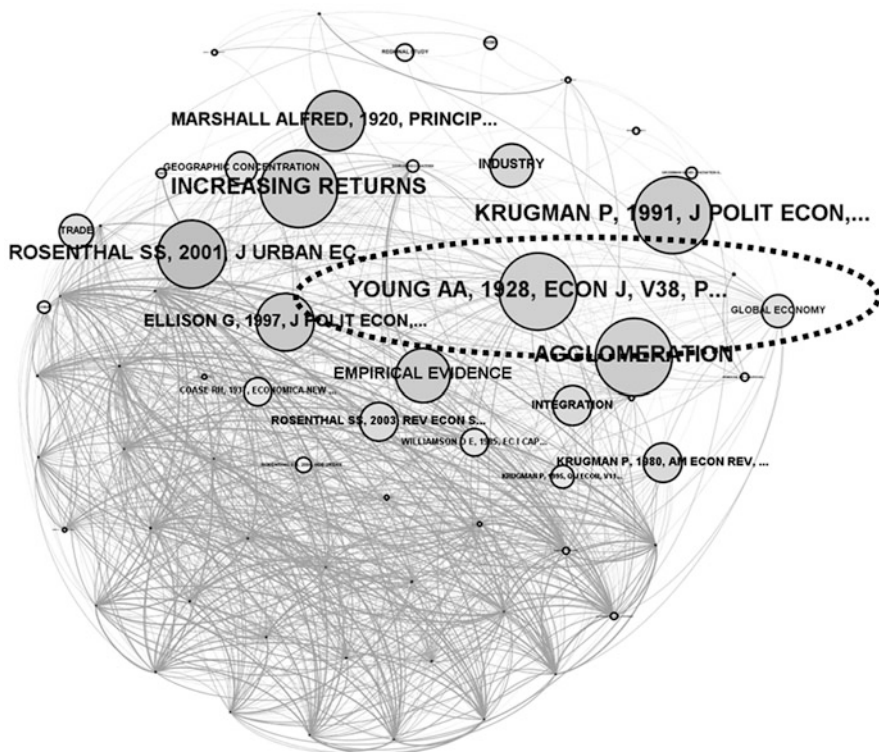


Fig. 12.5 Ego-network of “Agglomeration” and associated sleeping beauty, “Young, Allyn A., Increasing Returns and Economic Progress, *Economic Journal* DEC 1928; DOI: 10.2307/2224097”

central that node is to the network. In terms of a social network, centrality of an actor (node) indicates how powerful or crucial that actor (node) is for relations between all other actors (nodes). For a set of tokens with high betweenness centrality scores, one then can extract their ego-networks to discover the oldest published works associated with that token. If such a published work has zero or very few citations, then that research is a potential candidate for sleeping beauty in that research field. On the other hand, if the published paper does have citations, graphically determine the citation patterns. If the citation pattern shows dormancy of a time period exceeding five or more years, then that paper is a potential candidate for sleeping beauty.

The advantage of using multi-modal network with betweenness centrality score and ego network approach is that, this methodology is agnostic about citation analysis and yet is able to find potential sleeping beauties. One of the limitations of the current analysis is that the cited references may not have article titles (TI), and thus these references do not contribute to the set of tokens (T) whose ego-networks help identify potential sleeping beauties. Another issue that needs to be dealt in future

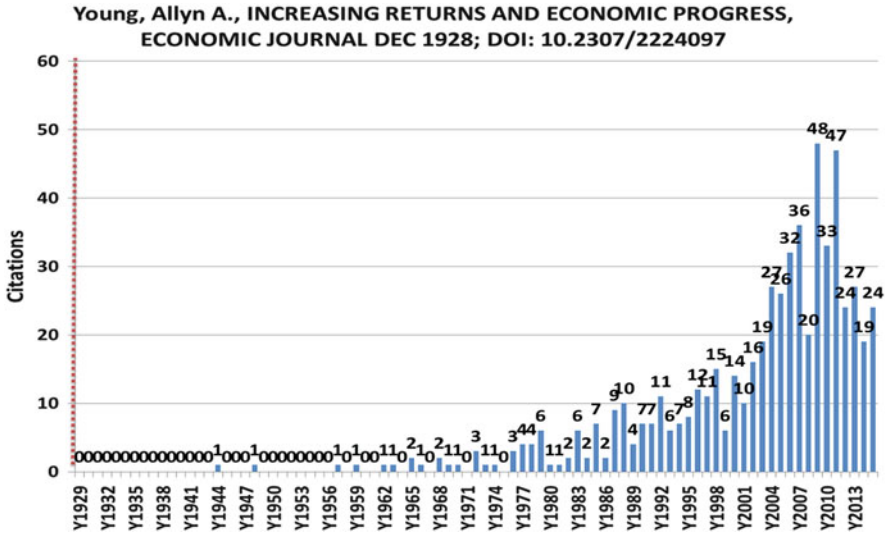


Fig. 12.6 The bar chart shows the citation pattern for the article. Clearly for first few decades this article garnered hardly any citations

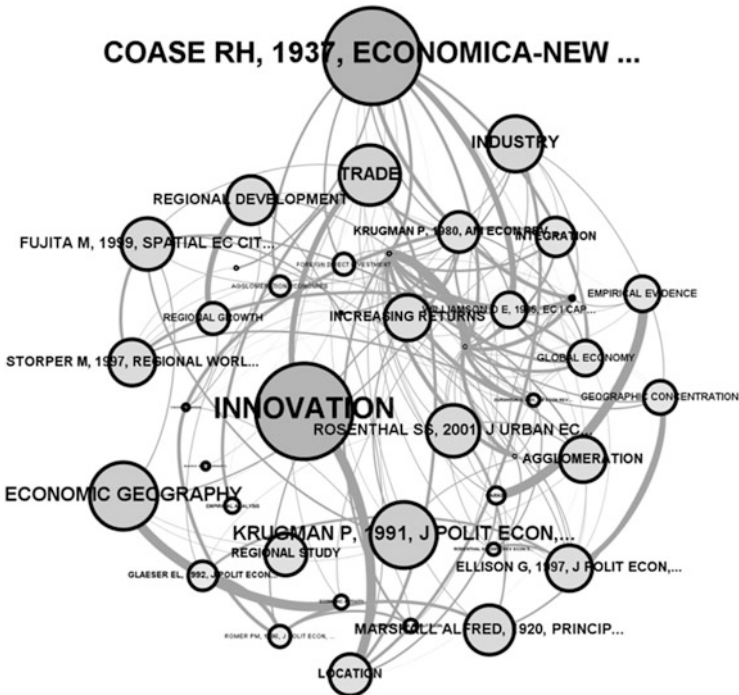


Fig. 12.7 Ego-network for “Innovation” and associated sleeping beauty, “Coase, R., 1937, The Nature of Firm, *Economica-New Series*, V4, N16, DOI:10.1111/j.1468-0335.1937.tb00002.x”

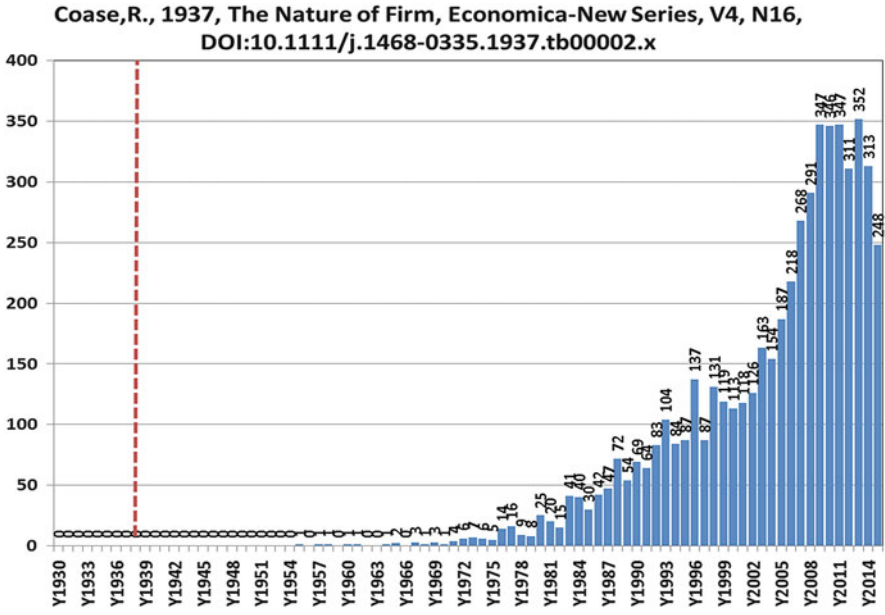


Fig. 12.8 The bar chart shows citation pattern for the article. Clearly for first few decades this article garnered hardly any citations

analysis is to check for consistency of output by using different sets of data. This is essential to filter out spurious association between a scholarly work and a field. In other words if some associations appear randomly between different data sets then those associations will have to filtered out from the analysis. For our future research, we would like to include reference titles which in turn could result in very large networks that will require big data analytics and substantial increases in the required computational resources. However, benefits of adding reference titles could help to convert the multi-mode mixed direction networks into bi-partite networks consisting of articles and tokens. Such bi-partite networks offer new avenues for analysis and identification of sleeping beauties. For the future analysis, we would also like to identify the “waking princes,” those papers that rediscover previously published work, adding a crucial piece of knowledge to the existing knowledge base. The current analysis does show in rudimentary fashion how some of the sleeping beauty articles multiply and are strongly connected to more than one token, and that these multiple tokens are indicators of different disciplines. In our future research, we would like to explicitly identify how certain sleeping beauties may have helped start new research avenues, thereby helping establish new disciplines.

of regions all over the globe. For example, one can envision in the near future, regional scientists and analysts trying to find new answers and solutions to resolve issues related to transportation, warehousing operations of commercial drones and small goods deliveries as fleets of these drones and/or automated vehicles start home and business deliveries over short hauls. Ever changing consumer taste for newer products and services along with demand for faster and efficient deliveries are likely to have ripple effects in regional manufacturing and services sectors, commercial surface transportation sector, labor productivity, regional employment and incomes. A contemporary example is how the local passenger transportation sector is affected by the rise of app-based taxi business (Uber) or how the hospitality sector may have to adjust to the rise of the private homes as temporary rental places (Airbnb). These contemporary examples of new technologies are offering competing or substitutional business opportunities that affect local and regional economies to different degrees, thereby giving researchers new avenues for policy analyses.

Discovering sleeping beauties within huge volumes of scholarly work may not be the most challenging issue, since we already have outlined in this paper an exploratory methodology that can be scaled up to big data. Part of the difficulty in the analysis may be with the compilation of data (published scholarly work) since much of that is accessible only with paid subscription. As it is now, free or public access to published literature is limited to open access journals. Regional science does not have a pubmed (<http://www.ncbi.nlm.nih.gov/pubmed>) equivalent repository of published literature. Google Scholar (<https://scholar.google.com/>) may be the closest to a publically available server of published scholarly work. In theory, one could make use of pre-print servers such as SSRN (<https://www.ssrn.com/>); however, the full access is still limited. There may be a few free access servers such as Sci-Hub on the dark internet (<https://en.wikipedia.org/wiki/Sci-Hub>), but any analyses based on such data source will not be accepted.

Maybe all is not lost. The U.S. Office of Science & Technology Policy (OSTP), via its memo, has put forth proposals that would make access to any published literature free if the underlying research has been supported by U.S. tax payer dollars (OSTP 2013). In response to OSTP's plans, American Publishers Association has embarked upon a public-private partnership initiative called "Clearinghouse for the Open Research of the United States" (CHORUS 2013), and it has been tracking how U.S. agencies are implementing open access to published research (US Agencies—Public Access Plans Details 2016). Recently, European Union's Competitive Council during its May 27–30 meeting adopted conclusions that state, "Member states agreed to common goals on open science and to pursue concerted actions together with the Commission and stakeholders. Delegations committed to open access to scientific publications as the option by default by 2020 and to the best possible reuse of research data as a way to accelerate the transition towards an open science system" (EC Competitive Council 2016; Science 2016). And yet the task of data compilation will not be easy since published scientific works in future are likely to be in the multi-media domain that include blogs, v-blogs (video blogs), self-publishing, microblogging, nanopublishing, simulated participatory games and

in the same vein publications authored by the public at large where different parts of the research are carried out by different teams at different times. Some of these publications would be in the form of living-documents such that the results obtained and conclusions reached are episodic in nature. Not that far out in the future, questions will arise about how to attribute idea origins when the underlying scholarly work is “crowd-sourced” or “crowd-created.” Or it may have been a joint effort by humans and machines—Artificial Intelligence (A.I.)

Who gets credit for original research ideas if they are either the outcome of AI or they are a living-document with episodic output or a multi-team, multi-media effort? Additionally, with fast moving innovative technologies giving rise to variety of new disciplines that involve inter-, cross- and multi-disciplinary research, how does one identify origins of research and ideas? We are seeing beginnings of addressing some of these issues in “altmetrics” (<http://altmetrics.org/manifesto/> 2010).

Assuming some or all of the above of how regional science research will unfold in the future, then the scientific work of discovering sleeping beauties associated with origins of ideas or new disciplines is going to be quite challenging. We think that some variation of our current approach based on a combination of text mining/analytics and network analysis will play a key role in identifying sleeping beauties in the future of regional science.

Acknowledgement We would like to express our special thanks to GMU’s Office of Research Computing (orc.gmu.edu) for computational resources to conduct largescale network analysis.

References

- Anne-Wil H (2016) Google Scholar as a new data source for citation analysis. <http://www.harzing.com/publications/white-papers/google-scholar-a-new-data-source-for-citation-analysis>. Accessed 3 June 2016
- Bomann L, Mutz R (2014) Growth rates of modern science: a bibliometric analysis based on the number of publications and cited references. <http://arxiv.org/abs/1402.4578>. Accessed 3 June 2016
- Chen C (2015) How to use CiteSpace. <https://leanpub.com/howtousecitespace>
- CHORUS (2013) Clearinghouse for the open research of the United States. <http://publishers.org/news/publishers-introduce-chorus>. Accessed 31 May 2016
- Competitiveness Council (2016) Research and innovation: open science. http://www.consilium.europa.eu/en/meetings/compet/2016/05/st09357_en16_pdf/ence. Accessed 31 May 2016
- Crassey D (2015) <http://www.nature.com/news/sleeping-beauty-papers-slumber-for-decades-1.17615>
- SSRN (2016) <https://www.ssrn.com>. Accessed 3 June 2016
- Garfield E (1970) Would Mendel’s work have been ignored if the ‘Science citation index’ was available 100 years ago? *Curr Contents* 2:69–70
- Garfield E (1980) Premature discovery or delayed recognition available via essays of an information scientist. <http://www.garfield.library.upenn.edu/delayedrecognition.html>. Accessed 3 June 2016
- Google Scholar (2016) <https://scholar.google.com/>. Accessed 3 June 2016

- Ke Q, Ferrara E, Radicchi F, Flammini A (2015) Defining and identifying Sleeping Beauties in science. *Proc Nat Acad Sci U S A* 112(24):7426–7431. www.pnas.org/cgi/doi/10.1074/pnas.1424329112. Accessed 3 June 2016
- Kleinberg J (2002) Bursty and hierarchical structure in streams. Proceedings of the 8th ACM SIGKDD international conference on knowledge discovery and data mining. <https://www.cs.cornell.edu/home/kleinber/bhs.pdf>. Accessed 3 June 2016
- Larivière V, Gingras Y (2015) The decline in the concentration of citations, 1900–2007. <http://arxiv.org/ftp/arxiv/papers/0809/0809.5250.pdf>. Accessed 3 June 2016
- OSTP Memo (2013) Increasing access to the results of federally funded scientific research. https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf. Accessed 31 May 2016
- Price DDS (1976) A general theory of bibliometric and other cumulative advantage processes. *J Am Soc Inf Sci* 27:292–306. doi:10.1002/asi.4630270505
- Priem J, Taraborelli D, Groth P, Neylon C (2010) Altmetrics: a manifesto. <http://altmetrics.org/manifesto>. Accessed 31 May 2016
- Remler D (2014) Are 90% of academic papers really never cited? Reviewing the literature on academic citations, London School of Economics. <http://blogs.lse.ac.uk/impactofsocialsciences/2014/04/23/academic-papers-citation-rates-remler/>. Accessed 4 June 2016
- Science (2016) <http://www.sciencemag.org/news/2016/05/dramatic-statement-european-leaders-call-immediate-open-access-all-scientific-papers>. Accessed 31 May 2016
- Thompson Reuter's Web of Science (2015) https://en.wikipedia.org/wiki/Web_of_Science. Accessed 17 July 2015
- U.S. Agencies Public Access Plans Details (2016) <http://www.chorusaccess.org/resources/us-agency-public-access-plans-details/>. Accessed 31 May 2016
- van Raan AFJ (2004) Sleeping Beauties in Science. *Scientometrics* 59(3):461–466

Rajendra Kulkarni is researcher/analyst at George Mason University (GMU) and affiliated with the Office of Research and School of Public Government and International Affairs. His research interests include nonlinear dynamical systems and their application to social sciences, complexity theory, network analysis, GIS, Transportation and Regional Science. He has BTech in Engineering and MS in Computer Science with extensive experience working with industry, business, government and academia. He holds a patent and has published in variety of disciplines with main focus on interdisciplinary model and algorithm development.

Roger R. Stough is university professor, the Schar School of Policy and Government, Vice President for Research and Economic Development (2008–2013) and currently Vice President, of the Regional Science Academy, an affiliate of the Regional Science International Association. His primary research interests are regional economic development and policy, transport policy, and entrepreneurship and innovation policy. He holds an honorary doctorate of philosophy, University of Jonkoping, Sweden.