Technological Forecasting & Social Change 164 (2021) 120502

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

journal homepage: www.elsevier.com/locate/techfore

Predicting scientific breakthroughs based on knowledge structure variations

Chao Min [a,*], Yi Bu [b], Jianjun Sun [a]

[a] School of Information Management, Nanjing University, Nanjing, China
[b] Department of Information Management, Peking University, Beijing, China

ARTICLE INFO

Keywords:
Scientific breakthrough
Early citing structure
Knowledge structure
Structure variation
Prediction

ABSTRACT

Breakthrough research plays an essential role in the advancement of the scientific system. The identification and recognition of scientific breakthroughs is thus of extreme importance. We propose a citing-structure perspective for observing the unfolding of breakthrough research from variations in knowledge structure. The hypothesis is empirically validated that scientific breakthroughs show distinctive knowledge structure characteristics, which are further utilized to predict breakthroughs in their early stage of formation. These characteristics include average clustering coefficient, average degree, maximum closeness centrality, and maximum eigenvector centrality in the direct citing networks of a breakthrough publication. Several explanations are provided for the effectiveness of the predictive models. We also show that: (1) the number of direct citation counts is of low predictive power, with even a negative impact on prediction performance; (2) disciplinary differences exist in knowledge structure, and this should be taken into account; (3) breakthrough characteristics are most prominent in the first layer of citing networks; (4) timing is critical, and 2- to 3-year-old citing networks have greater predictive power.

1. Introduction

Tech mining is defined in Porter & Cunningham's pioneering work (2004) as "the application of text mining tools to science and technology information, informed by understanding of technological innovation processes". The original definition of tech mining thus emphasizes two characteristics: the approach of text mining, and the goal of understanding technological innovation processes. As the discussion and impact of tech mining grow (Madani, 2015), the interpretation and application of this concept evolves as well. "Advanced Tech Mining",[1] for example, covers a broad range of methods, including bibliometrics, text analysis, social network analysis (Suominen et al., 2019), and deep learning (Zhang et al., 2018). Specific applications include development of novel indicators for tracking technological evolution, measurement of technology emergence, citation pattern analysis of emerging research topics (Porter et al., 2020), and identification of scientific and technological breakthroughs. In this work, we adopt a network structure perspective, rather than the text analysis method, but we too aim to broaden the understanding of the innovation processes that underlie scientific breakthroughs.

Scientific breakthroughs often lead to scientific revolutions (Kuhn, 1962) that change the way we know the world. Is it possible to forecast scientific breakthroughs in the early stage, based on the radical changes they proceed to bring about following breakthroughs' birth? In Kuhn's time, this seemed difficult, as scientific revolutions were not easily validated in empirical data. However, with the development of digitalization technologies, large-scale scholarly data has become widely available, bringing the possibility to quantitatively study scientific breakthroughs—even scientific revolutions—from the data. In this study, we aim to answer the previous question by exploring the unique features that breakthroughs have left in the space of science, from the perspective of their early citing structures. The logic behind this perspective is the simple and intuitive speculation that breakthrough research leads to dramatic changes (Kuhn, 1962; Koshland, 2007; Chen, 2012; Fortunato et al., 2018) in the structure of the scientific space. To verify this intuition, however, is not so easy, as neither breakthrough research nor structural change in the scientific space can be readily quantified. Since citing publications are in a sense inspired by a source publication, we view the entire set of citing publications and their citation relations as a proxy of the knowledge space that is brought

* Corresponding author.
E-mail address: mc@nju.edu.cn (C. Min).

[1] https://www.journals.elsevier.com/technological-forecasting-and-social-change/call-for-papers/advanced-techmining-measurement-emergence-indicators

https://doi.org/10.1016/j.techfore.2020.120502
Received 30 June 2020; Received in revised form 27 November 2020; Accepted 29 November 2020
Available online 15 December 2020

about by the source publication. Accordingly, we operationalize the structural changes in the scientific space as those in the network of citing publications.

It is not easy to operationalize a precise definition of breakthrough research or a breakthrough article. The concept of a breakthrough is variously defined in dictionaries but is generally considered to be associated with important discoveries and further development in many cases. Following are some representative definitions:

- *Oxford Dictionary*: "a sudden, dramatic, and important discovery or development."
- *Collins Dictionary*: "a significant development or discovery."
- The Free Dictionary: "a major achievement or success that permits further progress."

In addition, breakthroughs are also referred to as transformative research in the literature. The U.S. National Institutes of Health define transformative research as "unconventional research projects that have the potential to create or overturn fundamental paradigms".[2] The National Science Board (NSB) similarly describes transformative research as resulting in "a new paradigm or field of science or engineering" (NSB, 2007).

Despite these various definitions, whether a scientific work is considered breakthrough research is eventually decided by peers (Li et al., 2020; Schneider and Costas, 2017). In this study, we select Nobel Prize-winning papers as a benchmark of breakthrough research. Rather than focusing on the longitudinal depiction of individual scientific breakthroughs, we compare breakthrough papers with their non-breakthrough (or "less ground-breaking") counterparts to find the differences. Based on these differences, we take a step further by exploring the possibility of, and the optimized strategy for, predicting scientific breakthroughs. More than one hundred cases of breakthroughs in science history are collected and analyzed, together with their non-breakthrough counterparts. It is worth noting that, given the lack of consensus on the word "breakthrough", one should be cautious in using it and its variations or related terms. Nobel Prize research covers only a portion of all breakthrough research, and Nobel Prizes are awarded for individual discoveries as well as for a researcher's scientific oeuvre. Within the context of this study, only Nobel Prizes awarded for individual discoveries are relevant.

We adopt a research perspective very distinct from prior studies (e.g., Ponomarev et al., 2014, 2012; Savov et al., 2020; Schneider and Costas, 2017), namely, a citing-structure method. We define citing structure as the network topology of the citation network of the citing publications of a focus paper. In such a network, the nodes are precisely those publications that cite a paper of interest, and the citation relations among these citing publications are also considered. Huang et al. (2018) termed this network a "citing cascade." A prior study (Min et al., 2018) has shown that certain features of citation patterns can differentiate breakthrough and non-breakthrough papers. Our further observation implies that this effect is most significant in the first-order citing structure, since experimental results in Min et al. (2018) show that the difference between breakthrough and non-breakthrough is more significant in the first generation of citations than in further generations. The present study extends that line of investigation by depicting the difference in early ***network topology*** of directly citing publications between papers with different extent of breakthrough and by exploring the possibility of using this information to predict scientific breakthroughs at an early stage. We aim to find potential answers to the following research questions (RQs):

- RQ1: Do scientific breakthroughs impose particular influences on the knowledge structure of science? What are they, if any?
- RQ2: How can the particular influences that scientific breakthroughs have (if any) be explained?
- RQ3: Can we predict scientific breakthroughs based on their specific characteristics, if any? How can this prediction be performed?

2. Related studies

Scientific discoveries, whether theoretical or practical, are usually documented and recorded in formal scientific publications that exert an impact on the relevant research communities and change the whole science system. The sources and forms of scientific discoveries are quite complex. On the one hand, scientific discoveries can be either single—made by a single person—or multiple, i.e., independently made by multiple persons simultaneously (Merton, 1961; Burnham 2008). Contributing factors to multiple discoveries include the genius of scientists, cultural maturation, natural logical evolution, and coincidences (Ogburn and Thomas, 1922; Brannigan and Wanner, 1983a, 1983b; Simonton, 2004). On the other hand, scientific discoveries result not only from the professionalism of scientists, but also from serendipitous interactions between the scientists and the world (Campanario, 1996; Deflem, 2005; Fukawa, 2006; Yaqub, 2018). Van Andel (1994) has listed many examples of serendipitous patterns in which discoveries came from a surprising observation, a wrong hypothesis, a disturbance in an experiment, and so forth.

Despite an exponential increase in scientific publications over the years (Dong et al., 2017), only a small number of discoveries lead to a major achievement or success that permits further progress. These noteworthy discoveries are termed "scientific breakthroughs" per the definition of Free Dictionary. The occurrence of breakthroughs is shaped by institutional or organizational factors. In the setting of biomedical science, for example, Hollingsworth (2006), who held a path-dependent perspective, indicated that such institutional or organizational characteristics as autonomy, flexibility, scientific sensitivity, and communication all exert an influence. Moreover, breakthroughs usually unsettle the status quo of their relevant research areas and even establish new scientific paradigms. Due to their extreme importance, the identification of scientific breakthroughs has been widely studied in various domains, such as innovation studies (e.g., Guo et al., 2019; Petzold et al., 2019), quantitative science studies (e.g., Bettencourt et al., 2009), and library and information science (e.g., Bornmann et al., 2018; Small, 2018).

Although both qualitative and quantitative strategies have been adopted to identify breakthroughs in business (e.g., Hüsig et al., 2005; Sainio and Puumalainen, 2007), scientific breakthroughs have most often been investigated in a quantitative manner (Wang et al., 2013). Studies of such breakthroughs often employ the bibliographic data of scientific publications and examine their citations (e.g., Garfield, 1977; Shibata et al., 2007; Small et al., 2017; Schneider and Costas, 2017). Bettencourt et al. (2009), for example, held that the creation and spread of scientific discoveries lead to measurable changes in the social structure of a scientific community; they found evidence for this claim in topological transitions in researchers' collaboration networks. Shi et al. (2010) investigated the citation relations among the papers that a certain paper refers to, which they termed the "citation projection graph." They found that the way in which scientific papers draw previous knowledge together correlates with the papers' forward citation impacts. Shibata et al. (2007) investigated the correlation between citation counts and such centrality measures as clustering centrality, closeness centrality, and betweenness centrality, among which betweenness centrality was found to be correlated with future citations. Wolcott et al. (2016) considered both time-dependent and -independent characteristics and utilized citation- and coauthorship-wise metrics.

Yet purely considering citation counts may not allow for accurate identification of scientific breakthroughs (Wuestman et al., 2020; Min et al., forthcoming). One potential issue has been raised by Wuestman et al. (2020), who demonstrated that not all scientific breakthroughs are the same; thus, leveraging citation counts for detection introduces biases

[2] https://factor.niehs.nih.gov/2009/october/spotlight-2010grants.cfm

for breakthroughs driven by research objects (versus research questions). Hence, some scholars have explored other angles in breakthrough detection. For example, Small and Klavant (2011) investigated the role of citation contexts; specifically, they assigned a citation context into multiple categories corresponding to varying degrees of knowledge modality. Afterwards, they showed that the match rate between different modality categories and subsets of citation contexts could be indicative of breakthroughs. Winnink et al. (2019) designed and implemented five algorithms that leveraged not only citation counts but also other citation features (e.g., number of citing disciplines and new researchers citing the paper). These algorithms could categorize papers into five types corresponding to varying degrees of likelihood of being breakthroughs. Some other prior studies seek to make empirical use of citation networks instead of the number of citations itself. For instance, Wu et al. (2019) applied a previously proposed metric (Funk and Owen-Smith, 2017) to characterize disruptiveness. They defined the disruptiveness of a scientific publication by considering the number of publications that cite the focal publication but not the focal publication's reference(s), the number that cite the focal publication and its reference(s) simultaneously, and the number that do not cite the focal publication but do cite its reference(s). Min et al. (2021) proposed temporal and structural measurements to compare scientific breakthrough and non-breakthrough publications and observed that the structural dimension of citing literature networks shows particular promise for early identification of scientific breakthroughs.

While the problem of breakthrough identification has been approached from many angles, in this study we focus on an underexplored one: the structural properties of citing patterns. This approach is inspired by the observation that citation growth has a structural dimension (e.g., Garfield, 1977, 2006; Hu and Rousseau, 2016, 2017). Garfield (2006) offered the example of Albert Einstein, who published relatively few highly cited papers, but whose work was cited by other super-cited Nobel-class scientists. Another illustration of the second-generation citation effect appears in Francis Crick's work, which has been cited by 50 super-cited papers.

The publication of a particular paper, therefore, not only releases a new piece of knowledge but also changes the intellectual structure in its field of knowledge (Chen, 2012; Chen et al., 2009; Lv et al., 2018). Leydesdorff (2001) shows that newly published scientific papers may make fundamental changes to the existing body of knowledge. This is vividly illustrated by the diagrams he devised to display how a particular paper reduces the uncertainty in the current state of knowledge. These changes in network structure may in turn influence the spread of information over the networks (Lahiri et al., 2008). In citation networks specifically, Takeda and Kajikawa (2010) reported three stages of clustering: first the formation of core clusters, then the emergence of peripheral clusters, then finally the predominance of core clusters. Upham et al. (2009) called these cohesive intellectual communities in knowledge networks "schools of thought." Their analysis of scientific papers reveals how "schools of thought" both promote and constrain knowledge creation. They concluded that inclusion in a school of thought is particularly advantageous for new knowledge and that the most impactful position within a school of thought is in the semi-periphery.

Based on the above intuitions, we propose a hypothesis: the more radical the breakthrough research contained in a paper, the more drastic the changes it causes to the knowledge structure. We test the hypothesis by scrutinizing the citing structure of breakthrough and non-breakthrough papers. We try to determine whether there exists more evidence of difference in the topology of the citing structure and explore the possibility of predicting scientific breakthroughs based on empirical findings.

3. Methods and data

3.1. Methods

To quantify the structure of a citation network G with a set of nodes (papers) V and a set of edges (citation relations) E, we adopt the following network metrics:

(1) Number of nodes: $|V|$, the number of nodes in G;
(2) Number of edges: $|E|$, the number of edges in G;
(3) Average degree:

$$\text{average degree} = \frac{\sum_{v\in V}\deg(v)}{|V|}$$

where $v \in V$ and $\deg(v)$ refers to the degree of v;

(4) Network density:

$$network\ density = \frac{2|E|}{|V|(|V|-1)}$$

(5) Average clustering coefficient:

$$ACC = \frac{1}{|V|}\sum_{v\in V} C(v)$$

where $C(v)$ is the clustering coefficient of node v and is expressed as:

$$C(v) = \frac{number\ of\ closed\ triads\ connected\ to\ v}{number\ of\ triples\ of\ vertices\ centered\ on\ v}$$

(6) Maximum betweenness centrality: the highest betweenness centrality of nodes in G. The betweenness of a node v_i is:

$$\mathrm{B}(v_i) = \sum_{j<k} g_{jk}(v_i) \Big/ g_{jk}$$

where g_{jk} is the number of shortest paths from node v_j to node v_k, and $g_{jk}(v_i)$ is the number of shortest paths that pass through v_i;

(7) Maximum closeness centrality: the highest closeness centrality of nodes in G. The closeness of a node v is:

$$\mathrm{C}(v) = \frac{|V|-1}{\sum_{u=1}^{|V|-1} d(v,u)},$$

where $d(v,u)$ is the shortest-path distance between nodes v and u;

(8) Maximum eigenvector centrality: the highest eigenvector centrality of nodes in G. A node has more influential neighbours if it has a higher eigenvector centrality. The eigenvector centrality score of a node v is:

$$x_v = \frac{1}{\lambda}\sum_{t\in M(v)} x_t = \frac{1}{\lambda}\sum_{t\in G} a_{v,t}x_t$$

where $M(v)$ is a set of the neighbors of v, $A = (a_{v,t})$ is the adjacency matrix, and λ is a constant that can be derived from a small rearrangement of the above equation:

$$\mathbf{Ax} = \lambda \mathbf{x}$$

(9) Number of components: the number of connected components in a network.

As the source paper always has the highest degree in its citing network, we exclude it from the network to more objectively measure the network structure. The citing network t years after a focal paper's publication, $CNet_t$, is defined as: a network G in which the nodes V are publications that cite a focal paper of interest until year t, and the edges E are citation relations among these citing publications.

$$CNet_t = G(V, E)$$

Since we are interested in the knowledge structures of breakthrough papers relative to non-breakthroughs, we use logistic regression models to predict the presence or absence of an outcome in the Nobel group or the control group. In this model, the explained variable (prediction outcome) is a binary variable that indicates whether a paper is from the Nobel or the control group. The explanatory variables (predictors) are selected from the structural metrics of the citing network triggered by the paper of interest.

3.2. Data

As an outstanding example, the Nobel Prizes are acknowledgedly given to important breakthroughs in science. We operationalize scientific breakthroughs as papers winning the Nobel Prize in a given field (Physics, Chemistry, Physiology or Medicine, and Economic Sciences). Following Shen and Barabási (2014) and a prior study (Min et al., 2018), we collect 116 Nobel Prize breakthroughs, construct the prize-winning papers as breakthrough papers (Nobel group hereafter), and assign to a control group counterpart papers that were published in the same journal, in the same year, having received approximately equivalent citation counts. Next, the citing network of each of these papers is extracted from an in-house version of the Web of Science database. These citing networks are simplified to undirected graphs when we calculate the network metrics. As early citing structure is more useful in prediction and identification of breakthroughs, we focus on the citing networks within four years of the target paper's publication. This approach helps us obtain about 270 K citation relations between the focal papers and their citing papers.

4. Results

4.1. Disciplinary differences in knowledge structure

One could speculate that patterns of scientific knowledge accumulation might vary among scientific disciplines, and so might the formation and manifestation of scientific breakthroughs. If this is the case, disciplinary differences should be taken into account when identifying and predicting scientific breakthroughs. Therefore, we first of all seek to determine whether there do exist such differences in knowledge accumulation among the four fields of Nobel Prizes (Physics, Chemistry, Physiology or Medicine, and Economic Sciences) and what, if any, these differences are.

We analyze the structural metrics of citing networks that are initiated by all papers in both the Nobel group and the control group, grouped by disciplines. We find that the four disciplines indeed exhibit significant differences, given the statistics of the citing networks from all four years (Table 1).

The average publication year indicates that Economics papers are the oldest, followed by Physics and then Chemistry, with Physiology or Medicine papers the most recent. However, Physiology or Medicine papers receive many more direct citations (**NoN**) while Economic Sciences papers receive far fewer, with Physiology or Medicine > Chemistry > Physics > Economic Sciences. This is in stark contrast to the trend in publication date. Because the citation data end at the same time point for all papers, we can make the brief observation that Physiology or Medicine research accumulates citations the fastest and Economic Sciences the slowest, with Chemistry and Physics in between.

Table 1
Mean values of network metrics for papers from four disciplines (all years).

Mean	NoN	NoE	ND	ACC	MBC	MCC
Medicine	295.846	1619.723	0.092	0.460	0.007	0.194
Chemistry	148.092	630.992	0.138	0.442	0.017	0.200
Physics	115.528	574.545	0.172	0.474	0.019	0.235
Economics	19.750	31.324	0.354	0.190	0.003	0.048

Mean	**MEC**	**NoC**	**AD**	**Pub year**	**Prize year**	**Prize lag**
Medicine	0.089	45.846	5.000	1985.800	2009.785	23.990
Chemistry	0.143	31.042	3.407	1981.450	2005.980	24.530
Physics	0.149	17.900	3.753	1980.236	2003.472	23.240
Economics	0.365	10.951	0.433	1975.020	2004.950	29.930

Note: **NoN** = number of nodes, **NoE** = number of edges, **ND** = network density, **ACC** = average clustering coefficient, **MBC** = maximum betweenness centrality, **MCC** = maximum closeness centrality, **MEC** = maximum eigenvector centrality, **NoC** = number of components, **AD** = average degree.

In terms of the number of citation relations (**NoE**) among the citing publications, a similar finding is that Physiology or Medicine > Chemistry > Physics > Economic Sciences. In fact, this metric widens the gap among the four disciplines. But this order is reversed when we consider the ratio between actual citation relations and the possible maximum number (in a complete graph), that is, network density (**ND**): Economic Sciences shows the highest density, Physics the second-highest, Chemistry the third, and Physiology or Medicine the lowest. In other words, although physiologists and medical scientists create and cite knowledge more frequently in absolute quantity than economists do, the knowledge networks they build are not as dense as those constructed by economists.

Unlike network density, which measures a network from a global perspective, average clustering coefficient (**ACC**) measures the neighborhood of an "average" node in the network. This metric distinctly separates the natural-science groups from Economic Sciences: Chemistry, Physiology or Medicine, and Physics are very close in **ACC**, but Economic Sciences shows a much lower value. This suggests that the citation and flow of natural-sciences knowledge occurs more frequently in the local neighborhood of a paper, while the knowledge flow of the economic sciences is relatively slow in a paper's neighborhood. A reasonable speculation is that Economic Sciences papers tend to absorb knowledge from relatively distinct fields rather than their own fields. A related metric, average degree (**AD**), measures the number of nodes an "average node" links to in a network. Economic Sciences again shows the lowest value, while the three natural sciences are much higher: Physiology or Medicine has an extremely high value of 5, and Chemistry (3.407) and Physics (3.753) are fairly close. This further confirms the interdisciplinary variation in patterns of knowledge growth that was previously shown by **ACC**.

MBC, MCC, and **MEC** are relatively complicated metrics that consider the maximum value of node degree in a network from three different perspectives. Again, the three metrics consistently indicate the difference among the four disciplines. Economic Sciences shows lower values in both **MBC** and **MCC**, but higher values in **MEC**, than the three natural sciences. The three metrics further reveal a distinction within the natural sciences: Chemistry and Physics exhibit close values, but Physiology or Medicine seems to be a little different.

Physiology or Medicine has the largest number of components (**NoC**) in its citing networks. Two factors have played a role here: on the one hand, there are many nodes (**NoN**) in those networks; on the other hand,

Physiology or Medicine also has the lowest network density. Economic Sciences has the lowest **NoC**, with Chemistry and Physics in between, consistent with the situation for **NoN**.

With regard to award lag, Economic Sciences papers were on average the latest to receive a Nobel Prize after initial publication (29.93 years), while papers from Chemistry (24.53), Medicine (23.99), and Physics (23.24) received the award slightly faster.

4.2. Comparison between two groups in the timeline

We next compare the structure of citing networks initiated by papers from the Nobel group and the control group, respectively, at the resolution of each year after publication. Since our data are paired but not normally distributed, we adopt a nonparametric test, the Wilcoxon signed-rank test, for the comparison. As indicated in Table 2, papers from the two groups have persistently shown significant differences in four structural metrics (**NoE, ACC, MCC**, and **AD**) in each of the first four years after publication. However, these papers consistently show no significant between-group difference in terms of other metrics (**ND, MBC, MEC**, and **NoC**). Moreover, the difference in **NoN** is statistically significant only in the first year; it fades away through the following three years. In fact, the two groups of papers are relatively close in average citation counts in the first four years, with the Nobel group only slightly higher. This reveals that direct citation counts cannot effectively distinguish the Nobel from the control group in the early stage of breakthrough formation.

However, other metrics are able to depict the difference between the two groups in more detail. Generally speaking, the Nobel group shows higher values in most metrics, including **NoN, NoE, ACC, MBC, MCC** and **AD**; the exceptions are **ND, MEC** and **NoC**.

The Nobel group has an overall larger **NoE** than the control group. The effect is statistically significant and appears consistently over the four years. This phenomenon is interesting, especially given that the two groups have a similar number of direct citations (**NoN**) and that **NoN** is not significantly different in the second through fourth years. This indicates that a breakthrough paper has more citations in the network of its citing publications. Furthermore, a citing publication in such a network has on average more citations from (or references to) a publication that is also citing the breakthrough paper. This is revealed by the fact that the Nobel group has significantly higher average degree (**AD**) than does the control group.

The Nobel group also shows a significantly higher **ACC** than the control group, suggesting that the neighborhood of an average node in the former's citing network is more connected than in the latter's citing network. However, it is worth noting that the control group has a higher network density (**ND**) than the Nobel group from the second to the fourth years. This is somewhat strange, as both average clustering coefficient and network density measure the density of a network to an extent, but they seem to contradict each other in this case. The reason lies in the fact that the Nobel group has increasingly more citations (**NoN**) than the control group from the second year on, leading to exponentially decreasing network density by definition. Moreover, the difference in network density is not significant through the second to the fourth years. Therefore, we argue that **ACC** can more properly and effectively describe breakthrough papers than **ND** can.

Furthermore, the Nobel group shows higher values in terms of both **MCC** and **MBC. MCC** reveals that in the citing network of a breakthrough paper, there is a publication that is very close to all other publications. **MBC** suggests that there is also a publication that excels at connecting publications that are otherwise separated. The difference is statistically significant for **MCC**, though not for **MBC**.

The control group consistently displays larger **NoC** and **MEC** from the first year to the fourth year, although the difference is not strictly statistically significant. The metric **NoC** shows there are, on average, more separated subnetworks in a non-breakthrough paper's direct citing network. This again suggests a more disconnected network for the control group than for the Nobel group, a result previously indicated by **NoE, ACC**, and **AD**. The metric **MEC**, on the other hand, seems to indicate that the direct citing networks of the control group are characterized by an influential node that is linked to many other important nodes.

To sum up, we hold that the four metrics **NoE, ACC, MCC** and **AD** reveal relatively strong distinguishing power in a continuous timeline; thus, they have potential indicative value for breakthrough prediction and related tasks.

4.3. Regression analysis

Based on these knowledge structure characteristics, we further quantify scientific breakthroughs with the help of predictive models. We run a series of logistic regression models in which the dependent variable is whether a paper is from the Nobel group or the control group (1 = Nobel group, 0 = control group), and the independent variables are the structure metrics of the paper's citing network. The regressions lead to the elucidation of a series of metrics that can effectively characterize the breakthrough potential of a paper in various conditions.

4.3.1. Economic sciences vs. natural sciences

To start, we run regressions separately for Economic Sciences and the combined natural sciences in view of the overall disciplinary difference. As an exploration, we put all structure metrics into the models as independent variables, not distinguishing year information.

Table 3 displays the overall performance of the logistic regression models, with statistical significance of all variables included. The two models are overall statistically significant because the *p*-values are much less than 0.01. They also adequately fit the data, as the Hosmer-Lemeshow goodness-of-fit has significance values both greater than 0.05.[3] For Economic Sciences, none of the variables has a statistically significant coefficient. For the natural sciences, however, coefficients of both **ACC** and **AD** are statistically significant at the level of 0.01, while the coefficient of **MEC** is statistically significant at the level of 0.1. Moreover, all three variables have positive coefficients consistently for both Economic Sciences and the natural sciences. The Cragg-Uhler (Nagelkerke) R-squared statistic suggests that the models explain 18% and 9.9% of the variability of breakthrough papers in Economic Sciences and the natural sciences, respectively.

The results further suggest that we need to proceed with our analysis while taking this disciplinary difference into account. In addition, as citing networks with different ages could exhibit breakthrough characteristics to varying extents, we may also need to investigate the networks year-by-year after their formation instead of considering all years together.

4.3.2. Timing of breakthrough observation

Following the evolutionary process of breakthrough formation, we conduct further regression modeling in a year-by-year manner separately for Economic Sciences (Table 4) and the aggregated natural sciences (Table 5).

In terms of the overall performance of the models, the second year after the publication of breakthrough papers seems to be a breakpoint, both for Economic Sciences and for the natural sciences. The value of Prob $> \chi^2$ is rather high in the first year, indicating a poor model estimation. It begins to be lower than 0.05 or 0.01 in the second year, after which time the value increases again. Therefore, the statistical significance of the model is best in the second year. Likewise, the value of pseudo R-square reaches its maximum in the second year, suggesting that the model best explains the variability at that time (32.5% for Economic Sciences and 8.7% for the natural sciences).

[3] This statistic indicates a poor fit if the significance value is less than 0.05 (Hosmer et al., 2013).

Table 2
Wilcoxon signed-rank test result for the two groups of citing networks by year.

Indicator	NoN	NoE	ND	ACC	MBC	MCC	MEC	NoC	AD
First year									
Sig.	***0.024***	***0.003***	0.585	***0.001***	0.516	***0.013***	0.459	0.986	***0.001***
Nobel Avg.	43.052	110.809	0.270	0.333	0.011	0.149	0.256	14.435	1.662
Control Avg.	40.983	92.139	0.244	0.237	0.012	0.095	0.275	16.609	1.057
Second year									
Sig.	0.114	***0.030***	0.633	***0.001***	0.614	***0.042***	0.501	0.450	***0.001***
Nobel Avg.	93.313	379.957	0.194	0.401	0.018	0.179	0.196	19.278	3.007
Control Avg.	90.574	307.696	0.213	0.330	0.015	0.139	0.233	23.974	2.035
Third year									
Sig.	0.274	***0.081***	0.422	***0.000***	0.392	***0.017***	0.775	0.230	***0.000***
Nobel Avg.	149.548	799.887	0.139	0.457	0.015	0.217	0.141	22.313	4.236
Control Avg.	145.052	632.348	0.175	0.377	0.006	0.167	0.177	28.661	2.977
Fourth year									
Sig.	0.278	***0.060***	0.865	***0.000***	0.819	***0.001***	0.938	0.131	***0.000***
Nobel Avg.	213.852	1364.078	0.120	0.485	0.016	0.247	0.118	25.243	5.172
Control Avg.	198.235	989.478	0.130	0.422	0.009	0.184	0.138	32.409	3.738

Note: Significance values less than 0.1 are in bold text.

Table 3
Regression results for Economic Sciences and natural sciences (all years).

	Economic Sciences		Natural Sciences	
	Coef.	*p*	Coef.	*p*
NoN	0.006	0.968	–0.002	0.201
NoE	0.009	0.874	0.0002	0.245
ND	–1.126	0.365	0.483	0.572
ACC	0.420	0.876	2.553	**0.002**
MBC	–4.145	0.722	–0.408	0.753
MCC	–2.791	0.334	–1.100	0.237
MEC	0.175	0.888	1.388	**0.084**
NoC	–0.084	0.457	–0.001	0.918
AD	1.559	0.157	0.146	**0.003**
Cons	0.391	0.515	–1.542	0.000
Observations	204		671	
LR χ^2(9)	29.550		51.780	
Prob > χ^2	0.001		0.000	
Hosmer and Lemeshow χ^2 test	0.844		0.533	
Cragg-Uhler (Nagelkerke) R^2	0.180		0.099	

Note: Significance values less than 0.1 are in bold text.

In terms of variable significance, **MEC** is statistically significant ($p = 0.070$) in the second year, and **ND** is statistically significant ($p = 0.035$) in the third year, for Economic Sciences. With regard to the natural sciences, **AD** is the only statistically significant variable ($p = 0.040$) in the first year. The significant effects of **ACC, MCC**, and **AD** last from the second to the third year, while **MEC** is statistically significant ($p = 0.017$) only in the second year. Among these variables, **ACC, MEC** and **AD** show positive effects on the probability of being a Nobel Prize-winning paper, while **MCC** shows a negative effect.

With all these observations borne in mind, we next predict whether a scientific paper will be a breakthrough in the future based on its early citing network structure.

4.4. Prediction

In this section, we attempt to predict the potential of a paper to become a breakthrough in science based only on its early citing knowledge structures. We operationalize the prediction by choosing the structural metrics of citing networks as the predictors, and whether or not the paper comes from the Nobel group as the output. Structural metrics those have shown significant distinguishing power in previous sections are selected. Other structural metrics are also included to see their real effects in the prediction task. Previous observations indicate that 2–3 years after publication will offer the best predicting timing, and that Economic Sciences and the natural sciences are best separated for this task. A 5-fold cross-validation procedure is applied in which the data is randomly split into 5 groups. 80% of the cases are selected as training data to estimate the logistic model, and the remaining 20% are used as test data to validate the predictive power and generalizability of the model. This procedure is implemented 5 times with a bootstrap algorithm to obtain a reliable average performance (AUC) of the predictive model.

4.4.1. Predicting breakthroughs in the natural sciences

First, we run the prediction model for the natural sciences at year = 2 (Table 6). Various combinations of predictors are selected based on

Table 4
Regression results for breakthroughs in Economic Sciences by year.

	Year = 1		**Year = 2**		**Year = 3**		**Year = 4**	
	Coef.	*p*	Coef.	*p*	Coef.	*p*	Coef.	*p*
NoN	–37.773	0.975	–4.134	0.128	0.290	0.496	–0.035	0.845
NoE	17.316	0.977	1.945	0.117	–0.034	0.824	0.009	0.870
ND	–14.569	0.696	15.107	0.103	–6.329	**0.035**	–4.027	0.218
ACC	66.436	0.648	54.853	0.111	4.616	0.465	–4.614	0.317
MBC	0.000		403.989	0.172	–74.729	0.166	–13.999	0.398
MCC	0.000		–39.608	0.163	–6.297	0.376	4.794	0.402
MEC	18.009	0.690	–21.806	**0.070**	2.404	0.394	0.278	0.896
NoC	20.483	0.973	1.901	0.185	–0.484	0.212	–0.009	0.960
AD	–1.745	0.987	–24.004	0.147	0.612	0.810	2.683	0.108
Cons	34.378	0.977	10.221	0.047	1.605	0.373	0.753	0.639
Observations	39		53		56		55	
LR χ^2(9)	8.370		23.870		21.470		15.850	
Prob > χ^2	0.301		**0.005**		**0.011**		**0.070**	
Pseudo R^2	0.156		0.325		0.277		0.208	

Note: Significance values less than 0.1 are in bold text.

Table 5
Regression results for breakthroughs in the natural sciences by year.

	Year = 1		**Year = 2**		**Year = 3**		**Year = 4**	
	Coef.	*p*	Coef.	*p*	Coef.	*p*	Coef.	*p*
NoN	–0.004	0.821	–0.003	0.681	–0.001	0.760	–0.002	0.478
NoE	–0.001	0.884	0.000	0.762	0.000	0.737	0.000	0.419
ND	1.072	0.440	–1.888	0.317	0.456	0.838	3.941	0.154
ACC	0.430	0.800	4.088	**0.024**	4.015	**0.037**	2.852	0.192
MBC	–1.674	0.530	0.201	0.931	8.958	0.357	–0.583	0.866
MCC	–0.127	0.948	–3.448	**0.092**	–3.477	**0.074**	–0.362	0.855
MEC	0.254	0.853	4.507	**0.017**	2.238	0.229	–0.072	0.973
NoC	0.008	0.760	–0.002	0.878	–0.005	0.700	0.000	0.987
AD	0.481	**0.040**	0.268	**0.042**	0.200	**0.051**	0.123	0.166
Cons	–1.037	0.106	–2.205	0.013	–2.250	0.029	–2.249	0.078
Observations	163		166		170		172	
LR χ^2(9)	14.200		20.060		19.560		16.860	
Prob > χ^2	0.116		**0.018**		**0.021**		**0.051**	
Pseudo R^2	0.063		0.087		0.083		0.071	

Note: Significance values less than 0.1 are in bold text.

Table 6
Prediction results for natural sciences at year = 2.

AUC Predictors	Model 1 **NoN**	Model 2 **NoN, NoE**	Model 3 **NoN, NoE, NoC**	Model 4 **NoN, NoE, NoC, ND, ACC, AD**	Model 5 **NoN, NoE, NoC, ND, ACC, AD, MBC, MCC, MEC**	Model 6 **NoC, ND, ACC, AD, MBC, MCC, MEC**	Model 7 **ACC, AD, MCC, MEC**
1st fold	0.439	0.514	0.611	0.641	0.536	0.622	0.729
2nd fold	0.388	0.71	0.71	0.735	0.658	0.618	0.643
3rd fold	0.635	0.675	0.663	0.734	0.79	0.778	0.714
4th fold	0.365	0.365	0.421	0.516	0.599	0.714	0.556
5th fold	0.428	0.526	0.543	0.439	0.513	0.509	0.73
Mean AUC	**0.451**	**0.558**	**0.590**	**0.613**	**0.619**	**0.648**	**0.675**
95% CI	0.291, 0.484	0.427, 0.623	0.455, 0.649	0.529, 0.721	0.523, 0.712	0.554, 0.740	0.534, 0.720
SD AUC	0.107	0.139	0.113	0.132	0.111	0.103	0.076

model complexity and the results reported in previous sections. Model 1 selects the most naïve metric, direct citation counts, as the predictor, leading to a performance (AUC = 0.451) even worse than a random guess. Model 2 performs much better (AUC = 0.558) when the number of edges is considered. After other metrics (**NoC, ND, ACC, AD, MBC, MCC**, and **MEC**) are gradually added in Models 3–5, the performance improves accordingly. However, to our surprise, the predictive power becomes even stronger in Model 6 (AUC = 0.648) when two simple metrics (**NoN** and **NoE**) are excluded from prediction. This suggests that it is not the case that the addition of any predictor would increase the predictive power, and that some predictors can even be counterproductive. In Model 7, we choose only the four metrics that are statistically significant in Table 5 (year = 2) as predictors. With this set of metrics, the AUC increases to 0.675, the highest among all models in Table 6.

Next, we implement the prediction at year = 3 (Table 7). Model 8 selects three metrics that are statistically significant in Table 5 (year = 3) as predictors and obtains an AUC of 0.629. We further add **MEC** in Model 9 and obtain an increased AUC = 0.652, but this is still lower than 0.675 at year=2 in Table 6. Therefore, we tend to think that an optimal prediction strategy for natural sciences is to use **ACC, AD, MCC, MEC** two years after a paper is published.

Table 7
Prediction results for natural sciences at year = 3.

AUC Predictors	Model 8 **ACC, AD, MCC**	Model 9 **ACC, AD, MCC, MEC**
1st fold	0.638	0.675
2nd fold	0.592	0.608
3rd fold	0.618	0.643
4th fold	0.606	0.689
5th fold	0.690	0.646
Mean AUC	**0.629**	**0.652**
95% CI	0.488, 0.684	0.520, 0.712
SD AUC	0.038	0.032

4.4.2. *Predicting breakthroughs in economic sciences*

Predictions are also made for Economic Sciences at year = 2 and year = 3 with metrics that are statistically significant in Table 4, as well as other metrics in Table 6 for comparison. The results in Table 8 show that the models with statistically significant metrics in Table 4, namely Model 7 (year = 2, AUC = 0.695) and Model 8 (year = 3, AUC = 0.630), exhibit better performance than other models. Further, the former presents a better predictive power than does the latter, although the performance is relatively unstable (SD of AUC is 0.210). This again suggests that 2 years after a paper's publication is an ideal timing for predicting its potential to represent a scientific breakthrough.

5. Discussion

5.1. *Evidence: knowledge structure as predictive of breakthroughs*

When the science system is considered as an evolving network of knowledge, it can be intuitively speculated that the birth of a scientific breakthrough would cause immense impacts on the structure of this network. Though this phenomenon is easy to imagine, empirical evidence for it is lacking in the existing literature. By operationalizing the knowledge network as citing networks initiated by a scientific publication, we provide empirical evidence that scientific breakthroughs do impose particular influences on the knowledge structure of science, and that this observation is useful for predicting scientific breakthroughs.

From the network structure, a set of metrics is carefully singled out that can distinguish breakthrough papers from their non-breakthrough counterparts. These metrics reveal that breakthrough papers have

Table 8
Prediction results for Economic Sciences.

AUC Predictors	Model 1 **NoN**	Model 2 **NoN, NoE**	Model 3 **NoN, NoE, NoC**	Model 4 **NoN, NoE, NoC, ND, ACC, AD**	Model 5 **NoC, ND, ACC, AD, MBC, MCC, MEC**	Model 6 **ACC, AD, MCC, MEC**	Model 7 **MEC**	Model 8 **ND**
Economic Sciences, year = 2								
Mean AUC	0.5549	0.539	0.5131	0.368	0.5794	0.5998	0.695	0.589
95% CI	0.2622, 0.6096	0.4026, 0.7311	0.3356, 0.6674	0.1991, 0.5110	0.3247, 0.6710	0.4360, 0.7717	0.417, 0.751	0.2652, 0.6039
SD AUC	0.1422	0.1617	0.1319	0.0819	0.1223	0.0502	0.21	0.1832
Economic Sciences, year = 3								
Mean AUC	0.6267	0.565	0.5798	0.6319	0.559	0.572	0.606	0.63
95% CI	0.3433, 0.6825	0.4067, 0.7490	0.4051, 0.7421	0.4345, 0.7823	0.3806, 0.7220	0.4139, 0.7624	0.3678, 0.7138	0.365, 0.718
SD AUC	0.099	0.2088	0.1864	0.1515	0.1583	0.138	0.1954	0.15

more connected citing networks than do papers with less ground-breaking ideas. This is reflected in various aspects of network topology. Compared with that of a less ground-breaking paper, the citing network of a breakthrough paper has more edges globally (**NoE**), while the numbers of nodes (**NoN**) are relatively equivalent. An average node in a breakthrough paper's citing network has significantly more nodes linking to it (**AD**), coming from works that it refers to and works that cite it. Both sources of nodes are also located in the citing network. The neighbours surrounding an average node also cite each other (**ACC**) more frequently, and the network topology is more cohesive (**NoC**), with fewer components disconnected from one another, despite more nodes in the network. Moreover, **MCC** suggests that in the citing network of a breakthrough paper, there is a paper that is very close to all other papers. It is likely that this paper is another important (breakthrough) paper (Min et al., 2020) or a Prince paper (Braun et al., 2010) that assists the breakthrough. In terms of MEC, the Nobel group shows a lower value than the control group, although this effect is not statistically significant. The reason may be that breakthrough papers outshine the brilliance of their citing papers. Altogether, these findings provide empirical evidence for our hypothesis that more radical breakthroughs would cause more drastic changes to the knowledge structure of science, thus answering ***RQ1*** proposed in the Introduction.

5.2. Optimal network layer and timing for prediction

Experimental results have shown that the particular influences imposed by scientific breakthroughs on the knowledge structure of science propagate across network layers and time. This poses the problem of choosing the best angle from which to leverage the full potential of this network to identify and predict breakthroughs. Digesting the results from a prior study (Min et al., 2018) in which we extracted four generations of citing networks, we further find that the breakthrough characteristics of those networks fade away generation after generation. That is to say, the first generation of the citing network could be the best network layer in which to observe scientific breakthrough's particular impacts.

In this work, we focus on the first-generation citing network, also called the direct citing network. We further find that breakthrough characteristics vary with time. There are two pieces of evidence: first, for the natural sciences, **AD**'s effect remains significant for the first three years and disappears in the fourth year, while the significant effects of **ACC** and **MCC** last only from the second to the third year. Second, both the statistical significance and the explanatory power of the regression models first increase and then decrease with time, as does the number of significant variables in the models. Therefore, it is reasonable to infer that the predictability of scientific breakthroughs presents an inverted U shape in the timeline, first increasing and then decreasing. Neither too early nor too late a timing is optimal, but 2–3 years after publication is more likely to yield a better prediction.

5.3. Breakthrough prediction: performance, explanation, & improvement

Is it possible to predict which scientific papers will become breakthroughs in science in the future? This is a difficult task (Wang et al., 2017), but we argue that it is not totally impossible. We prove that breakthroughs and non-breakthroughs have shown significant differences in citing network structures in the first few years after publication, not just decades later when the breakthroughs receive well-known Nobel Prizes. These differences were not reflected in early citation counts, but rather in early citation structures. Using the structural characteristics, we are able to predict the potential of a scientific work to become a breakthrough in the future with an AUC of 67.5% or 69.5% only two years after the work's publication. In contrast, prediction with citation counts can only obtain an AUC of 45%, even worse than a random guess. The number of direct citations is thus of very low predictive power, even exerting a negative impact on prediction tasks. Although adding other structure-based predictors improves the model performance, an optimal suite of predictors is demonstrated to be **ACC, AD, MCC, MEC** for the natural sciences and **MEC** for Economic Sciences, in each case two years after publication.

Table 2 shows that, compared with the control group, the Nobel group has on average a higher number of edges (**NoE**), network density (**ND**), average clustering coefficient (**ACC**), and average degree (**AD**), but a lower number of disconnected network components (**NoC**)—a greater network connectivity, in summary. We now answer ***RQ2*** by providing some possible explanations for the greater connectivity of breakthroughs' citing networks than those of non-breakthroughs at the level of an "average" node. First, Kuhn (1962) has suggested that drastic changes in science structure, which he called "paradigm shifts", take place during a "revolutionary" state in science. During such a time, scientific breakthroughs have been achieved and new theories have been proposed to deal with anomalies that cannot be explained by existing theories. We argue that the structural characteristics of breakthroughs' citing networks are a manifestation of "paradigm shifts." Second, when a scientific breakthrough occurs, it will be followed up by more and more subsequent works, as will these subsequent works themselves. This process of breakthrough formation resembles a "gold rush" phenomenon in which once a new lode is discovered by a scientist, an increasing number of scientists turn to this promising field (Suominen et al., 2019). Third, we make an analogy between breakthrough formation and innovation clusters. Innovative entities often emerge in the form of clusters where pools of expertise, talent, and capital accelerate the development of new technologies and new industries (Engel, 2015; Tallman et al., 2004). The observations in the breakthrough papers' citing networks in this study coincide with the mechanism whereby innovations take the shape of clusters.

Obviously, there is space for improvement in terms prediction. Although an AUC of 67.5% or 69.5% represents great progress from random guesses, especially for a social science problem, we have to acknowledge that it is barely satisfactory compared with prediction tasks on other well-defined problems in natural sciences. This is because

there are still many unobservable factors that influence the prediction but have not been found yet. These factors are hard to quantify, but they deserve further exploration in future research.

5.4. Practical notes for breakthrough prediction

Based on the results presented above, we are also inclined to answer ***RQ3*** in the affirmative. From the viewpoint of practical tech mining, we list four points worthy of notice for researchers and practitioners in search of future scientific breakthroughs using the approach proposed in this study.

First, among the structural metrics, only those that have positively predictive powers should be considered in the prediction model. By experiment we find that not all structural metrics are statistically different between breakthroughs and non-breakthroughs. We eliminate those indifferent metrics and are left with some predictive metrics. Moreover, certain metrics, such as **NoN**, may even have a negative effect on prediction. These metrics should surely be excluded from prediction models.

Second, differences among disciplines ought to be taken into account. Different disciplines exhibit multifarious characteristics in terms of knowledge structure. Within the Nobel Prize taxonomy, Economic Sciences and the natural sciences (Physics, Chemistry, Physiology or Medicine) are two basically different classes of scientific fields in terms of almost all structural metrics as well as the time taken to achieve a Prize. It seems more difficult to predict breakthroughs in Economic Sciences than in the natural sciences. Furthermore, the field of Physiology or Medicine is found to be somewhat different from Physics and Chemistry, whose respective knowledge structures share more similar characteristics. Therefore, it is necessary to separate disciplines to attain a better prediction performance.

Third, focusing on the direct citing network is sufficient. The influences of breakthrough ideas are most prominent in the first layer of the citing network. It thus seems unnecessary to go deeper, incurring more costs only to obtain sub-optimal performance.

Fourth, appropriate timing is critical for prediction. Structural metrics are statistically different in some years but not so in other years. This reveals that the knowledge structure variations induced by scientific breakthroughs can only be observed with ease in certain periods. Before that, they are too immature to emerge; afterward, the citing networks have grown too large to reveal significant differences and nuances in some structural features.

6. Conclusion and future research

As early as the 1960s, Margolis (1967) described a dilemma of using citation data: a short-term citation ***count*** is unreliable, but historical evaluation takes too long. He pointed out that an intermediate period must exist when citation networks begin to contain adequate information for "meaningful analysis." We believe that ***tech mining*** research will benefit from the integration of the method of citation analysis and the utilization of such an "intermediate period" during the evolution of citation networks. Echoing Margolis's anticipation, in this study, we leverage the potential of early citation networks to investigate the knowledge structures of scientific breakthroughs and find metrics useful for predicting such breakthroughs, which is one of the many facets of tech mining. Our results so far give affirmative answers to and empirical and theoretical evidence for the three research questions proposed at the beginning of this article, and there is a long way to go in the quest to create practical tools for early identification of scientific breakthroughs. Our next steps include expanding the pool of breakthrough data to include more sources. There are also theoretical problems calling for further research efforts, e.g., (1) why certain metrics are statistically different but others are not, (2) how and why such differences arise and disappear in a longer time window after publication. Lastly, we note that the approach proposed in this study could be employed in real scenarios (e.g., government and funding agencies) to assess its practical value.

CRediT authorship contribution statement

Chao Min: Conceptualization, Methodology, Formal analysis, Investigation, Writing - original draft. **Yi Bu:** Software, Investigation, Writing - original draft, Writing - review & editing. **Jianjun Sun:** Supervision, Project administration.

Acknowledgments

Financial support from the National Science Foundation of China (NSFC No. 71904081, No. 71874077), Chinese Education Department Research Foundation for Humanities and Social Sciences (No. 19YJC870017), Jiangsu Province Social Sciences Foundation (No. 18TQC005) and Fundamental Research Funds for the Central Universities (14380041) is gratefully acknowledged. A short version of this paper was presented at the 17th International Conference on Scientometrics and Informetrics (ISSI 2019) in Rome, Italy, and the authors thank all audiences for offering their invaluable suggestions to improve this paper. The authors are also grateful to several anonymous reviewers for their constructive comments.

References

Bettencourt, L.M., Kaiser, D.I., Kaur, J., 2009. Scientific discovery and topological transitions in collaboration networks. J. Informetr. 3 (3), 210–221. https://doi.org/10.1016/j.joi.2009.03.001.

Bornmann, L., Ye, A., Ye, F., 2018. Identifying landmark publications in the long run using field-normalized citation data. J. Doc. 74 (2), 278–288. https://doi.org/10.1108/JD-07-2017-0108.

Brannigan, A., Wanner, R.A., 1983a. Historical distributions of multiple discoveries and theories of scientific change. Soc. Stud. Sci. 13 (3), 417–435. https://doi.org/10.1177/030631283013003004.

Brannigan, A., Wanner, R.A., 1983b. Multiple discoveries in science: a test of the communication theory. Can. J. Sociol.-Cahiers Can. Sociol. 8 (2), 135–151. https://doi.org/10.2307/3340123.

Braun, T., Glänzel, W., Schubert, A., 2010. On sleeping beauties, princes and other tales of citation distributions. Res. Evaluat. 19 (3), 195–202. https://doi.org/10.3152/095820210X514210.

Burnham, J.C., 2008. Accident proneness (Unfallneigung): a classic case of simultaneous discovery/construction in psychology. Sci. Context 21 (1), 99–118. https://doi.org/10.1017/S0269889707001573.

Campanario, J.M., 1996. Using citation classics to study the incidence of serendipity in scientific discovery. Scientometrics 37, 3–24. https://doi.org/10.1007/BF02093482.

Chen, C., 2012. Predictive effects of structural variation on citation counts. J. Am. Soc. Inf. Sci. Technol. 63 (3), 431–449. https://doi.org/10.1002/asi.21694.

Chen, C., Chen, Y., Horowitz, M., Hou, H., Liu, Z., Pellegrino, D., 2009. Towards an explanatory and computational theory of scientific discovery. J. Informetr. 3 (3), 191–209. https://doi.org/10.1016/j.joi.2009.03.004.

Deflem, M., 2005. The travels and adventures of serendipity: a study in sociological semantics and the sociology of science. Soc. Forces 83 (3), 1302–1303. https://doi.org/10.1353/sof.2005.0025.

Dong, Y., Ma, H., Shen, Z., Wang, K., 2017. A century of science: globalization of scientific collaborations, citations, and innovations. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1437–1446. https://doi.org/10.1145/3097983.3098016.

Engel, J.S., 2015. Global clusters of innovation: lessons from Silicon Valley. Calif. Manage. Rev. 57 (2), 36–65. https://doi.org/10.1525/cmr.2015.57.2.36.

Fortunato, S., Bergstrom, C.T., Börner, K., Evans, J.A., Helbing, D., Milojević, S., Vespignani, A., 2018. Science of science. Science 359 (6379). https://doi.org/10.1126/science.aao0185.

Fukawa, I., 2006. Case studies on how to enhance the chance of technical breakthrough and (pseudo) serendipity. In: Proceedings of technology management for the global future (PICMET). Istanbul, Turkey, 2, pp. 668–675. https://doi.org/10.1109/PICMET.2006.296601.

Funk, R.J., Owen-Smith, J., 2017. A dynamic network measure of technological change. Manage. Sci. 63 (3), 791–817. https://doi.org/10.1287/mnsc.2015.2366.

Garfield, E., 1977. Highly cited articles. 39. Biochemistry papers published in the 1950s. Curr. Contents 25, 5–12. http://garfield.library.upenn.edu/essays/v3p147y1977-78.pdf.

Garfield, E., 2006. Identifying Nobel class scientists and the uncertainties thereof. In: European Conference on Scientific Publication In Medicne and Biomedicine & the Third Nordic Conference on Scholarly Communication. http://www.garfield.library.upenn.edu/papers/luND2006.pdf.

Guo, J., Pan, J., Guo, J., Gu, F., Kuusisto, J., 2019. Measurement framework for assessing disruptive innovations. Technol. Forecast. Soc. Chang. 139, 250–265. https://doi.org/10.1016/j.techfore.2018.10.015.

Hollingsworth, J.R., 2006. A path-dependent perspective on institutional and organizational factors shaping major scientific discoveries. In: Hage, J., Meeus, M. (Eds.), Innovation, Science, and Institutional Change. Oxford University Press, Oxford, UK, pp. 423–442.
Hosmer Jr., D.W., Lemeshow, S., Sturdivant, R.X., 2013. Applied Logistic Regression, 3rd ed. John Wiley & Sons, Hoboken, NJ.
Hu, X., Rousseau, R., 2016. Scientific influence is not always visible: the phenomenon of under-cited influential publications. J. Informetr. 10 (4), 1079–1091. https://doi.org/10.1016/j.joi.2016.10.002.
Hu, X., Rousseau, R., 2017. Nobel Prize winners 2016: igniting or sparking foundational publications? Scientometrics 110 (2), 1053–1063. https://doi.org/10.1007/s11192-016-2205-x.
Huang, Y., Bu, Y., Ding, Y., Lu, W., 2018. Number versus structure: towards citing cascades. Scientometrics 117 (3), 2177–2193. https://doi.org/10.1007/s11192-018-2952-y.
Hüsig, S., Hipp, C., Dowling, M., 2005. Analysing disruptive potential: the case of wireless local area network and mobile communications network companies. R D Manage 35 (1), 17–35. https://doi.org/10.1111/j.1467-9310.2005.00369.x.
Koshland, D.E., 2007. Philosophy of science - The cha-cha-cha theory of scientific discovery. Science 317 (5839), 761–762. https://doi.org/10.1126/science.1147166.
Kuhn, T.S., 1962. The Structure of Scientific Revolutions. University of Chicago Press, Chicago, IL.
Lahiri, M., Maiya, A.S., Sulo, R., Wolf, T.Y.B., 2008. The impact of structural changes on predictions of diffusion in networks. In: IEEE International Conference on Data Mining Workshops '08. https://doi.org/10.1109/ICDMW.2008.92.
Leydesdorff, L., 2001. The Challenge of scientometrics: The development, measurement, and Self-Organization of Scientific Communications. Universal Publishers, Boca Raton, FL.
Li, J., Yin, Y., Fortunato, S., Wang, D., 2020. Scientific elite revisited: patterns of productivity, collaboration, authorship and impact. J. R. Soc. Interface 17 (165), 20200135. https://doi.org/10.1098/rsif.2020.0135.
Lv, Y., Ding, Y., Song, M., Duan, Z., 2018. Topology-driven trend analysis for drug discovery. J. Informetr. 12 (3), 893–905. https://doi.org/10.1016/j.joi.2018.07.007.
Madani, F., 2015. 'Technology mining' bibliometrics analysis: applying network analysis and cluster analysis. Scientometrics 105 (1), 323–335. https://doi.org/10.1007/s11192-015-1685-4.
Margolis, J., 1967. Citation indexing and evaluation of scientific papers. Science 155 (3767), 1213–1219. https://doi.org/10.1126/science.155.3767.1213.
Merton, R.K., 1961. Singletons and multiples in scientific discovery: a chapter in the sociology of science. Proc Am Philos Soc 105 (5), 470–486.
Min, C., Bu, Y., Sun, J., Ding, Y., 2018. Is scientific novelty reflected in citation patterns?. In: Proceedings of the 81st Annual Meeting of the Association for Information Science and Technology, 55, pp. 875–876. https://doi.org/10.1002/pra2.2018.14505501155.
Min, C., Bu, Y., Wu, D., Ding, Y., Zhang, Y., 2021. Identifying citation patterns of scientific breakthroughs: a perspective of dynamic citation process. Inf. Process. Manage. 58 (1), 102428. https://doi.org/10.1016/j.ipm.2020.102428.
Min, C., Zhang, S., Sun, J., 2020. Citation diffusion in the networks of scientific publications: a case study on the 2011 Nobel Chemistry prize winning paper (in Chinese). J. China Soc. Sci. Tech. Inf. 39 (3), 259–273. http://qbxb.istic.ac.cn/EN/abstract/abstract273.shtml.
NSB (National Science Board), 2007. Enhancing Support of Transformative Research At the National Science Foundation. National Science Foundation, p. 14. https://www.nsf.gov/nsb/documents/2007/tr_report.pdf.
Ogburn, W.F., Thomas, D., 1922. Are inventions inevitable? A note on social evolution. Polit. Sci. Q. 37 (1), 83–98. https://doi.org/10.2307/2142320.
Petzold, N., Landinez, L., Baaken, T., 2019. Disruptive innovation from a process view: a systematic literature review. Creat. Innov. Manag. 28 (2), 157–174. https://doi.org/10.1111/caim.12313.
Ponomarev, I., Williams, D., Hackett, C., Schnell, J., Haak, L., 2014. Predicting highly cited papers: a method for early detection of candidate breakthroughs. Technol. Forecast. Soc. Chang. 81, 49–55. https://doi.org/10.1016/j.techfore.2012.09.017.
Ponomarev, I.V., Williams, D.E., Lawton, B.K., Cross, D.H., Seger, Y., Schnell, J., Haak, L. L., 2012. Breakthrough Paper Indicator: early detection and measurement of groundbreaking research. In: Proceedings of the 11th International Conference on Current Research Information Systems. https://pdfs.semanticscholar.org/5e39/d91f5627d89bb9ab67db65ac9391f1f55904.pdf.
Porter, A.L., Chiavetta, D., Newman, N.C., 2020. Measuring tech emergence: a contest. Technol. Forecast. Soc. Chang. 159, 120176. https://doi.org/10.1016/j.techfore.2020.120176.
Porter, A.L., Cunningham, S.W., 2004. Tech mining: Exploiting new Technologies For Competitive Advantage. John Wiley & Sons, Hoboken, NJ. https://www.readcube.com/articles/10.1002%2F0471698466.
Suominen, A., Peng, H., Ranaei, S., 2019. Examining the dynamics of an emerging research network using the case of triboelectric nanogenerators. Technol. Forecast. Soc. Chang. 146, 820–830. https://doi.org/10.1016/j.techfore.2018.10.008.
Sainio, L.M., Puumalainen, K., 2007. Evaluating technology disruptiveness in a strategic corporate context: a case study. Technol. Forecast. Soc. Chang. 74 (8), 1315–1333. https://doi.org/10.1016/j.techfore.2006.12.004.
Savov, P., Jatowt, A., Nielek, R., 2020. Identifying breakthrough scientific papers. Inf. Process. Manage 57 (2), 102168. https://doi.org/10.1016/j.ipm.2019.102168.
Schneider, J.W., Costas, R., 2017. Identifying potential "breakthrough" publications using refined citation analyses: three related explorative approaches. J. Assoc. Inf. Sci. Technol. 68 (3), 709–723. https://doi.org/10.1002/asi.23695.
Shen, H.W., Barabási, A.L., 2014. Collective credit allocation in science. Proc. Natl. Acad. Sci. U. S. A. 111 (34), 12325–12330. https://doi.org/10.1073/pnas.1401992111.
Shi, X., Leskovec, J., McFarland, D.A., 2010. Citing for high impact. In: Proceedings of the 10th Annual Joint Conference on Digital Libraries. ACM, pp. 49–58. https://doi.org/10.1145/1816123.1816131.
Shibata, N., Kajikawa, Y., Matsushima, K., 2007. Topological analysis of citation networks to discover the future core articles. J. Am. Soc. Inf. Sci. Technol. 58 (6), 872–882. https://doi.org/10.1002/asi.20529.
Simonton, D.K., 2004. Creativity in science: Chance, logic, genius, and Zeitgeist. Cambridge University Press, Cambridge, UK.
Small, H., 2018. Characterizing highly cited method and non-method papers using citation contexts: the role of uncertainty. J. Informetr. 12 (2), 461–480. https://doi.org/10.1016/j.joi.2018.03.007.
Small, H., Klavant, R., 2011. Identifying scientific breakthroughs by combining co-citation analysis and citation context. In: Noyons, E., Ngulube, P., Leta, J. (Eds.), Proceedings of the 13th international conference of the International Society for Scientometrics & Informatics (ISSI), pp. 783–793. In: https://www.issi-society.org/proceedings/issi_2011/ISSI_2011_Proceedings_Vol2_27.pdf.
Small, H., Tseng, H., Patek, M., 2017. Discovering discoveries: identifying biomedical discoveries using citation contexts. J. Informetr. 11 (1), 46–62. https://doi.org/10.1016/j.joi.2016.11.001.
Takeda, Y., Kajikawa, Y., 2010. Tracking modularity in citation networks. Scientometrics 83 (3), 783. https://doi.org/10.1007/s11192-010-0158-z.
Tallman, S., Jenkins, M., Henry, N., Pinch, S., 2004. Knowledge, clusters, and competitive advantage. Acad. Manage. Rev. 29 (2), 258–271. https://doi.org/10.5465/amr.2004.12736089.
Upham, S., Rosenkopf, L., Ungar, L., 2009. Positioning knowledge: schools of thought and new knowledge creation. Scientometrics 83 (2), 555–581. https://doi.org/10.1007/s11192-009-0097-8.
van Andel, P., 1994. Anatomy of the unsought finding. Serendipity: origin, history, domains, traditions, appearances, patterns, and programmability. Br. J. Philos. Sci. 45 (2), 631–648. https://doi.org/10.1093/bjps/45.2.631.
Wang, D., Song, C., Barabási, A.L., 2013. Quantifying long-term scientific impact. Science 342 (6154), 127–132. https://doi.org/10.1126/science.1237825.
Wang, J., Veugelers, R., Stephan, P., 2017. Bias against novelty in science: a cautionary tale for users of bibliometric indicators. Res. Policy 46 (8), 1416–1436. https://doi.org/10.1016/j.respol.2017.06.006.
Winnink, J.J., Tijssen, R.J.W., van Raan, A.F.J., 2019. Searching for new breakthroughs in science: how effective are computerised detection algorithms? Technol. Forecast. Soc. Chang 146, 673–686. https://doi.org/10.1016/j.techfore.2018.05.018.
Wolcott, H.N., Fouch, M.J., Hsu, E.R., DiJoseph, L.G., Bernaciak, C.A., Corrigan, J.G., Williams, D.E., 2016. Modelling time-dependent and-independent indicators to facilitate identification of breakthrough research papers. Scientometrics 107 (2), 807–817. https://doi.org/10.1007/s11192-016-1861-1.
Wu, L., Wang, D., Evans, J.A., 2019. Large teams develop and small teams disrupt science and technology. Nature 566 (7744), 378–382. https://doi.org/10.1038/s41586-019-0941-9.
Wuestman, M., Hoekman, J., Frenken, K., 2020. A typology of scientific breakthroughs. Quant. Sci. Stud. 1 (3), 1023–1222. https://doi.org/10.1162/qss_a_00079.
Yaqub, O., 2018. Serendipity: towards a taxonomy and a theory. Res. Policy 47 (1), 169–179. https://doi.org/10.1016/j.respol.2017.10.007.
Zhang, Y., Lu, J., Liu, F., Liu, Q., Porter, A., Chen, H., Zhang, G., 2018. Does deep learning help topic extraction? A kernel k-means clustering method with word embedding. J. Informetr 12 (4), 1099–1117. https://doi.org/10.1016/j.joi.2018.09.004.

Chao Min is a Research Assistant Professor in the School of Information Management at Nanjing University. His research interests mainly include knowledge and innovation diffusion, science of science and network science. He holds a doctor's degree in management science from Nanjing University. He was a visiting scholar at the School of Informatics, Computing and Engineering in Indiana University Bloomington from 2016 to 2018. He is a member of the International Society for Scientometrics and Informetrics (ISSI) and a member of the Association for Information Science and Technology (ASIS&T).

Yi Bu is an Assistant Professor at the Department of Information Management, Peking University, China. He is doing research in the application aspect of big data analytics, with a particular focus on scholarly data mining. Specifically, he is interested in the process of knowledge dissemination and innovation diffusion, the analysis of scholarly networks and their variants, and bibliometric indicators for research evaluation. He has published over 30 papers and has served as reviewer/PC member for 20+ international journals and conferences.

Jianjun Sun is Professor of Information Management and Dean in the School of Information Management at Nanjing University. He is a distinguished professor of the Cheung Kong Scholar Program of China's Ministry of Education. His researches focus on scientific big data analytics, online information resources management and innovation evaluation. He has published more than 190 academic papers and books. He also serves as Vice Chairman of Chinese Information Society of Social Sciences, executive director of China Society for Scientific and Technical Information, and executive director of China Association for Information Systems.