








The emergence of graphene research topics through interactions within and beyond

Ai Linh Nguyen¹ , Wenyuan Liu¹ , Khiam Aik Khor² ,
Andrea Nanetti³ , and Siew Ann Cheong¹ 

¹Division of Physics and Applied Physics, School of Physical and Mathematical Sciences, Nanyang Technological University, 21 Nanyang Link, Singapore 637371

²School of Mechanical & Aerospace Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798

³School of Art, Design and Media, Nanyang Technological University, 81 Nanyang Dr, Singapore 637458

an open access  journal



Citation: Nguyen, A. L., Liu, W., Khor, K. A., Nanetti, A., & Cheong, S. A. (2022). The emergence of graphene research topics through interactions within and beyond. *Quantitative Science Studies*, 3(2), 457–484. https://doi.org/10.1162/qss_a_00193

DOI:
https://doi.org/10.1162/qss_a_00193

Peer Review:
https://publons.com/publon/10.1162/qss_a_00193

Supporting Information:
https://doi.org/10.1162/qss_a_00193

Received: 6 May 2021
Accepted: 15 March 2022

Corresponding Author:
Ai Linh Nguyen
S180009@e.ntu.edu.sg

Handling Editor:
Ludo Waltman

Copyright: © 2022 Ai Linh Nguyen, Wenyuan Liu, Khiam Aik Khor, Andrea Nanetti, and Siew Ann Cheong. Published under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.



Keywords: emergences, graphene, interactions, journal publications, topics

ABSTRACT

Scientific research is an essential stage of the innovation process. However, it remains unclear how a scientific idea becomes applied knowledge and, after that, a commercial product. This paper describes a hypothesis of innovation based on the emergence of new research fields from more mature research fields after interactions between the latter. We focus on graphene, a rising field in materials science, as a case study. First, we used a coclustering method on titles and abstracts of graphene papers to organize them into four meaningful and robust topics (*theory and experimental tests*, *synthesis and functionalization*, *sensors*, and *supercapacitors and electrocatalysts*). We also demonstrated that they emerged in the order listed. We then tested all topics against the literature on nanotubes and batteries, and the possible parent fields of *theory and experimental tests*, as well as *supercapacitors and electrocatalysts*. We found incubation signatures for all topics in the nanotube papers collection and weaker incubation signatures for *supercapacitors and electrocatalysts* in the battery papers collection. Surprisingly, we found and confirmed that the 2004 breakthrough in graphene created a stir in both the nanotube and battery fields. Our findings open the door for a better understanding of how and why new research fields coalesce.

1. INTRODUCTION

In the prevailing linear theory of innovation (Turney, 1991), a commercial product first emerges as an idea in pure research before the idea is fleshed out in the field of applied research. Once this idea is mature, innovators would then develop prototypes based on it and go through many trials and optimizations before one or more prototypes become commercially feasible to appear on the market. In this picture of innovation, progress from pure research to applied research to technology is in stages (Kline & Rosenberg, 1986; Mansfield, 1991; Narin, Hamilton, & Olivastro, 1995, 1997; Rosenberg & Birdzell, 1990). After getting interested in a scientific study of graphene research, we started looking into this innovation process using graphene publication data (Nguyen, Liu et al., 2020). Since then, we have carried out additional analyses, and our preliminary results suggest that the linkages between applied graphene research and graphene technology are very complex. Therefore, we started to focus more on understanding the connections between pure and applied graphene research because these appear to be more straightforward. Nevertheless, we realize that describing these linkages in terms of stages oversimplifies the whole process.

Our previous paper looked at macroscopic indicators comprising the number of publications, the number of references per publication, and the number of citations per publication, and how they changed over the years. We have explained that it is challenging to tell pure graphene research at the level of these indicators apart from applied graphene research and suggested that these two stages may be distinguished through a community analysis of the citation network. Furthermore, we believe that the two stages should also be distinguishable through the different technical terms used in pure and applied graphene research papers. Following this intuition, we investigated the titles and abstracts of graphene journal papers and identified four topics based on this linguistic information. Two of them, *theory and experimental tests* and *synthesis and functionalization*, can be considered pure graphene research because they are motivated by curiosity. The other two, *sensors* and *supercapacitors and electrocatalysts*, can be regarded as applied graphene research because they aim to produce usable technologies.

In general, unless it appears de novo, we expect a research topic to be an offshoot from a more established research topic. For example, there is a more extended history of research progressing from graphite (Kelly, 1981) to fullerenes (Bethune, Johnson et al., 1993; Kroto, Heath et al., 1985) to nanotubes (Ajayan, 1999; Iijima, 1991) before graphene became recognized as a research field by itself (Geim, 2009; Novoselov, Geim et al., 2004). Also, research into graphite did not stop when scientists started studying fullerenes. The same is true for nanotubes after graphene was discovered. Therefore, a more accurate way to describe the continuation of an old research topic and the emergence of a new research topic is in terms of research streams. In this stream-based picture, we developed indicators to show explicitly the times at which new research streams emerge from their parent streams. Then, we let the data speak for itself: whether an emergent research topic is completely novel or born from parent streams and what these parent streams might be.

There is an implicit suggestion that the applied research stage/stream emerges from the standard innovation model's pure research stage/stream. We found it unreasonable to think more carefully about this implication because most scientists are specialists in different research topics. For example, we do not expect graphene *theory and experimental tests* specialists to switch after some time to fabricating graphene *sensors* and graphene *supercapacitors and electrocatalysts*. Instead, it is more reasonable to assume that the first scientists working on graphene *supercapacitors and electrocatalysts* have previously worked on batteries and other types of energy storage devices. If this is true, then the battery stream is the parent of the graphene *supercapacitors and electrocatalysts* stream. We also believe that this could not have happened spontaneously, but only after the scientists in the battery stream became aware of the progress made by scientists working on graphene *synthesis and functionalization*. We think of this information flow as arising from interactions between research streams. For example, in pure graphene research, the *theory and experimental tests* stream frequently interacts with the *synthesis and functionalization* stream. Some of these interactions do not bear fruit, but others can lead to breakthroughs at various scales. Naturally, a more careful analysis is necessary to determine whether traditional battery research scientists learn graphene synthesis on their own, or graphene *synthesis and functionalization* scientists learn the science and technology of batteries on their own, or the two groups collaborate.

To put it simply, while everyone agrees that new knowledge is created based on old knowledge, it is unclear how this happens. Our hypothesis can be stated in the stream picture: A new stream has to emerge from an old stream after an incubation period, following the interactions between the old stream and other existing streams. To test this hypothesis, we ask the following scientific questions: (a) which topics emerge first (and when), and which topics emerge later (and when), and are there logical reasons for the sequence of emergence; (b) which are the parent topics for the various graphene topics; and (c) what are the interactions between topics that led to the creation of the graphene topics?

To answer these questions, we outline our theoretical framework in Section 2. We will also describe the hypothesis and indicators in detail before surveying literature relevant to our hypothesis and scientific questions. This paper will limit the investigation to two case studies: (1) *theory and experimental tests* and (2) *supercapacitors and electrocatalysts*, for reasons we will explain in Section 4.2. After that, we describe in Section 3.1 the data used in this paper, which includes a collection of graphene publications and collections of nanotubes and batteries publications required by our two case studies. We then describe in Section 3.2 the coclustering method that we use to partition graphene publications into clusters, each with a distinct word-use pattern. In Section 4.1, we show that the graphene publications can be organized into four robust and validated research topics: (0) *synthesis and functionalization*, (1) *supercapacitors and electrocatalysts*, (2) *sensors*, (3) *theory and experimental tests* with the aid of coclustering method. We then show in Section 4.2 how the numbers and proportions of papers in the four topics and their interest curves change with time to answer our first scientific question. We see that *theory and experimental tests* were the first topic to emerge from the interest curves, followed by *synthesis and functionalization*, and then *sensors*, and finally *supercapacitors and electrocatalysts*. In Section 4.3, we answer our second scientific question through a series of incubation analyses. We expected the four topics to have different parents: in particular for *supercapacitors and electrocatalysts* to have emerged from batteries. Still, we were surprised to find that the nanotubes field is the parent of all of them. We proceed to answer our third scientific question in Section 4.4 by analyzing interactions between research streams. We had expected the interaction signatures between graphene topics and the more mature streams to be weak. Instead, we were surprised to find powerful interactions from graphene to nanotubes and batteries. Finally, we conclude in Section 5.

2. THEORETICAL FRAMEWORK AND LITERATURE SURVEY

2.1. Theoretical Framework

The standard model of innovation consists of four stages (Figure 1(a)), namely (a) pure research, (b) applied research, (c) technology, and (d) commercialization (Turney, 1991). The activity starts

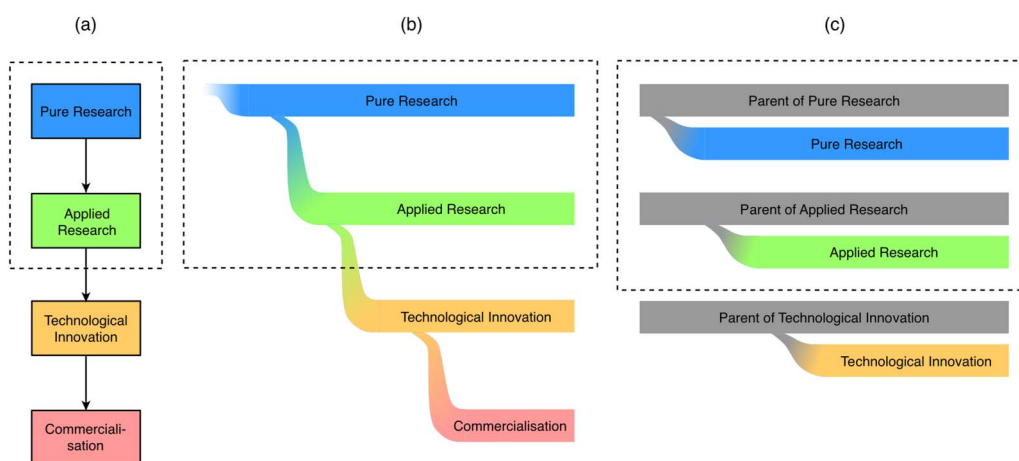


Figure 1. Schematic diagrams for (a) the linear model of innovation, consisting of a pure research stage, followed by an applied research stage, then a technological innovation stage, and finally the commercialization stage; (b) a stream-based visualization of the innovation processes, showing the emergence of pure research, followed by the emergence of applied research, then that of technological innovation, and finally commercialization; and (c) visualization of the emergence of a new field from existing fields, where pure research, applied research, and technological innovation grew out of their respective parent fields.

with the pure research stage, and when the fruits of pure research are ripe, the action moves on to the applied research stage. When applied research is mature, innovators will develop insights from this stage into technology. Eventually, some of the most promising technologies become commercial products on the market. This is essentially a linear model, but feedback loops have been identified between different stages (Liu, Nanetti, & Cheong, 2017). In our previous paper (Nguyen et al., 2020), we argued that this description in stages leaves out the time dimension. Therefore, instead of steps, we go to a stream-based description of innovation processes (Figure 1(b)). Each stream would represent an independent research topic that is persistent in time. In this paper, we focus on the first two stages, which can be considered pure and applied scientific research, and therefore the appropriate data sets are scientific publications.

When we think more deeply about the stream-based picture, and also from our experiences in Liu et al. (2017) and Nguyen et al. (2020), we realize that Figure 1(b) is also overly simplistic. First, the applied research stream could not have emerged from the pure research stream. In the same sense, experimental research could not have emerged as an offshoot from theoretical analysis. Instead, an emergent field's pure and applied research streams must have emerged from different parent streams, as shown in Figure 1(c). In this revised picture, we hypothesize that streams interact episodically with each other, and after an interaction episode, new streams can emerge from old streams after a period of incubation. The critical theoretical concepts we introduce here are the parent streams, interactions between streams, and the new streams' emergence conditions. An embryonic topic may die during the incubation stage if certain conditions are not met. Figure 1(c) shows a pure research stream emerging before a corresponding applied research stream emerges, followed by the corresponding technological innovation stream from their respective parent streams. This is the standard time order in the linear innovation model, but other time orderings of the emergences may also be possible. These would then correspond to feedback loops in the linear innovation model.

2.2. Literature Survey

2.2.1. How do new research topics emerge, and why do scientists choose to work on them?

Different aspects of the stream-based picture we outlined in Section 2.1 have been noted and discussed separately in the Philosophy of Science literature. For example, on the emergence of a new scientific topic, we find the perennial debate between Popper and Kuhn. According to Popper, the scientific method consisted of first formulating a hypothesis, then designing an experiment to test the hypothesis (Popper, 1959). If the experimental results do not contradict the hypothesis, it survives to be tested another day. In this sense, scientific knowledge progresses incrementally. The most significant step is the first step when we go from no understanding of a phenomenon to tentative knowledge in the form of a simple hypothesis. Subsequent steps to refine this hypothesis are assumed to be smaller and smaller. On the other hand, Kuhn realized that in some cases that we have a more and more pronounced discrepancy between theory and experiment, no matter how we refine the hypotheses. He then favored a hypothesis of scientific revolutions, in which new theories with very different structures displace the old theories as an explanation of the phenomena (Kuhn, 1962). In a recent analysis of publications by the American Physical Society (APS), we found that science progresses incrementally as Popper believed. Still, now and then, we found abrupt changes to the organization of scientific knowledge (Liu et al., 2017). Kuhn called the most dramatic of these *scientific revolutions*, but we think of these as scientific breakthroughs of different scales.

In some of these breakthroughs, we find new theories displacing old theories, whereas, in other breakthroughs, we see the emergence of new topics. Kuhn was the earliest to consider

the problem of the emergence of new scientific topics (Kuhn, 1977). In his 1977 book, Kuhn discussed the essential tension that scientists work within. In other words, they are devoted both to the maintenance of the paradigm they work in and making discoveries that might undermine the paradigm. This book discusses the creative destruction of the paradigm, innovations, and divergent and convergent phases on the path to scientific consensus. However, Kuhn did not mention whether the innovative, divergent step involves collaboration or cross-fertilization. Others believed that scientific fields evolve through divergent processes, like branching caused by growth and discoveries (Mulkay, 1975; Price, 1986), specialization, and fragmentation (Dogan & Pahre, 1990). More recently, we also have scientists who believe in the role of concurrent processes. For example, Herrera, Roberts, and Gulbahce (2010) quantified the extents of cross-fertilization between fields in physics from the network of PACS numbers in the APS data set, while Bettencourt, Kaiser, and Kaur (2009) realized that universal cocitation signatures were marking the emergence of new fields and concluded that collaborations in the communities of physicists are meaningful. Sun, Kaur et al. (2013) examined how the modularity of the network of physicists publishing in APS journals evolved and observed that the modularity jumped every time a new journal is created (this is a specialization process and frequently associated with the creation of a new field). After that, the modularity started decreasing again, and Sun et al. believed that this was due to the cross-fertilization between physicists from different areas. They then built a purely social agent-based model of scientists, who follow a few rules to work together on publications. They compared their simulation results against six stylized facts (authors per paper, papers per researcher, researchers per discipline, disciplines per researcher, papers per discipline, and disciplines per paper) and found reasonable agreement. Finally, Salatino, Osborne, and Motta (2017) investigated 75 newly emerged topics and 100 well-established topics, randomly selected from a collection of 3 million computer science papers on 2,000 topics, and found roughly eight out of nine newly emerged topics were preceded by periods of intense collaborations.

In the literature focusing on the emergence of new research topics, Boyack, Klavans et al. (2014) and Boyack and Klavans (2019) are most relevant to us because they discussed the emergence of graphene as a research area. In Boyack et al. (2014), they relaxed the criteria for community detection in the cocitation networks from 2000 to 2010 to obtain a large number of *microcommunities*. By matching microcommunities from successive years based on how their references overlap, they organized the microcommunities into *threads*. They then found 15% to 16% of all threads starting within the 11-year time window and identified these as *emergent topics*. Finally, they described their analysis of about 50 graphene-related threads that contain at least 35% of their graphene collection or contain the three largest microcommunities in any given year. Their research tells the story of the explosive growth of graphene as a field starting in 2006. In their 2019 review paper (Boyack & Klavans, 2019), after introducing readers to various large bibliographic databases, such as Web of Science, Scopus, Microsoft Academic Graph, large full-text databases, such as ScienceDirect, arXiv, and IEEE Xplore, and large patent databases, such as USPTO and Derwent, as well as project-level funding databases, such as UberResearch, Boyack et al. (2014) went on to describe different similarity measures and various community detection and clustering methods used to explore and analyze these large data sets. After that, they surveyed the literature on analyzing such large data sets. The state of the art is a dynamic picture of science as a system of evolving and interacting clusters that change from year to year. Their main finding is high rates of birth and death events. We have done something similar in Liu et al. (2017). However, we considered no deaths of topical clusters at the granularity level, and births were also rare. One of the most prominent birth events we found was high-temperature superconductivity, but even so,

we have to be careful because these papers do have references. These references were all very old, leading to weak similarity with the topical cluster from the previous year. Boyack et al. then presented the emergence of graphene research as a topical case study. They identified four graphene topics as the most important based on their methods. Still, they highlighted that these topics all have references from 1990, thus concluding that the graphene topics are not entirely novel according to their definition.

Related to how new topics emerge is the question of how frequently scientists switch topics. Working on the Association for Computing Machinery (ACM) and Institute of Electrical and Electronics Engineers (IEEE) data sets of publications, Hoonlor, Szymanski, and Zaki (2013) found that, on average, computer scientists switched topics every 10 years. Later, in their perspective paper, Battiston, Musciotto et al. (2019) examined the APS data set and found that many physicists moved away from the fields that their first papers are in. Depending on the fields, these moves can be as early as 3–4 years after the first publication or as late as 6–7 years after publication. They did not investigate subsequent transitions. Zeng, Shen et al. (2019) found that modern physicists and computer scientists switch topics more frequently than their predecessors. Furthermore, the probability of hitting topics is higher for the earlier part of the career than the latter part. They then developed a model to explain the negative correlation between average citations per paper and switching probability. Aleta, Meloni et al. (2019) also tested the APS data set and used PACS as a proxy for different physics topics. The PACS is hierarchical, so we can tell that the two PACSs are more closely related. They found that most physicists switch topics every 4–5 years, but the new topics remain in broader areas.

Finally, we ask how scientists choose the new topics to switch into, given the myriad choices available. For example, do they decide to go into a subject closely related to what they are working on, or do they seek out more distant topics? In an early personal reflection published in *Science*, Reif (1961) lamented how much grant funding dictated the choices of research topics made by practicing scientists. Gieryn (1978) summarized quantitative studies of this problem of options of research topics and wrote down a typology of changes and continuities. More recently, Hoonlor et al. (2013) compared keywords extracted from NSF funded grants to those extracted from the ACM and IEEE publications. They found that changes in topics of interest in the publications were preceded by changes in such topics in the grants. Osborne, Scavo, and Motta (2014) studied papers published between 2000 and 2010 on the World Wide Web and Semantic Web. Their results supported the connection between multi-disciplinary collaboration among mature fields and the emergence of new research areas. Using 30 years of APS publications, Jia, Wang, and Szymanski (2017) found that the overlap between a physicist's current research interests and those early in their career (measured in terms of PACS numbers) decays exponentially with time. They then built the seashore-walk model to explain how physicists switch from one topic to the next, based on how rewarding the topic is perceived to be. While studying how frequently scientists change topics, Zeng et al. (2019) also observed from their APS publication data set that most scientists have narrow distributions of research topics. Going beyond an analysis of bibliometric data, Foster, Rzhetsky, and Evans (2015) also developed a typology of five research strategies: (a) jump, (b) new consolidation, (c) new bridge, (d) repeat consolidation, and (e) repeat bridge. The jump and new bridge strategies are associated with scientists switching their research topics.

2.2.2. Identification of distinct research topics

Even though we were working with topical collections to find subfields within these collections, we were still led to clustering methods. To cluster based on bibliography, we find three main methods: direct citation, cocitation, and bibliographic coupling (BCN, sometimes

referred to as *coreference*). Direct citation (a paper and its references are linked) and cocitation (two papers being cited by the same paper are linked) were first proposed by Small (1973), while BCN (two papers having at least one common reference are linked) was introduced 10 years earlier by Kessler (1963). The cocitation method is the most popular (Janssens, Zhang et al., 2009; Liu, Yu et al., 2010). In fact, in the review by Boyack and Klavans (2010), it was mentioned that “cocitation analysis was adopted as the de facto standard in the 1970s, and has enjoyed that position of preference ever since.” In our work, we appreciate the similarities and differences between the three methods. Most importantly, we realized that the citation and cocitation networks for a given field change over time as more papers are published. On the other hand, once we have decided the collection of papers to use, the BCN constructed will no longer change, even as more papers in the field are published later. This backwards-looking nature of the BCN makes it convenient for doing historical analysis of a field instead of using the forward-looking citation and cocitation networks, whose results might depend on when we end the collection.

Another way to cluster documents is to use the words in them. This idea of text-based clustering can be traced back to Callon, Courtial et al. (1983). After pioneering the co-word methodology as a tool to analyze “the relationships between research activity and the general socio-political context” (Callon et al., 1983), Callon, Courtial, and Lavoie (1991) then followed up and used the tool to understand interactions between innovation steps and to investigate whether basic research or applied research could be the driving force. In this paper, they used polymer science as a case study and found that they could distinguish between pure and applied research in polymer science using co-words. Furthermore, by measuring the centrality, density, and content transformation of the links between co-words, they also found that as the field matures, different parts of the field (characterized by various combinations of co-words) become more closely linked (a phenomenon the authors called *global integration*). At the same time, multiple distinct centers of research activities emerged, a phenomenon the authors referred to as *polycentrism*.

Ultimately, bibliographic and linguistic features of scientific papers are like facets of the more complex objects themselves, which we can better understand by combining information from different aspects. According to Yu, Wang et al. (2017), citation-based and text-based bibliographic clustering offer various advantages and disadvantages over each other. In discussing the disadvantages of both methods, Glänzel and Thijs (2011) realized that the relationships between documents are underestimated in citation-based approaches due to very sparse matrices and overestimated in text-based methods due to the lower discrimination power of highly repetitive vocabularies. Therefore, as early as the 1990s, there were suggestions to combine the strengths of the two methods and overcome their weaknesses by adopting a hybrid approach. In 1991, Braam and coworkers introduced the first hybrid approach to combine cocitation and word analysis in mapping scientific research on the level of research specialties (structural aspects) (Braam, Moed, & Van Raan, 1991a), as well as exploring time-dependent scientific activities (dynamic aspects) (Braam, Moed, & Van Raan, 1991b). In this approach, the clusters were obtained from the citation-based analysis. At the same time, the structural and semantic terms were extracted from their textual content (assuming that documents that share the exact citations will have related word contents). There is also a second approach, proposed by Glänzel and Czerwon (1996) based on a “core document” concept. In a nutshell, a core document in a thematic cluster is the publication with the highest centrality and is considered the representative paper for the topic. As a result, in each thematic cluster obtained using the cocitation methodology, labels were extracted from the core documents’ titles, keywords, and abstracts (Glänzel & Czerwon, 1996; Glänzel & Thijs, 2011). More recently,

in the 2010 and 2011 works by Boyack and his colleagues (Boyack, Newman et al., 2011; Boyack & Klavans, 2010), the performance of text-based, citation-based, and hybrid approaches were compared for a data set of 2.15 million PubMed documents. They concluded that the best citation-based and text-based approaches have similar accuracy, but the hybrid approach outperformed both. In their later work, which considered the relationship between reference similarity and reference proximity (their relative positions in the text) (Gipp & Beel, 2009), Boyack, Small, and Klavans (2013) found an increase in performance accuracy when combining reference proximity into the cocitation model.

3. DATA AND METHODS

3.1. Data

To create our data sets, a graphene expert suggested that we used “single-layer carbon” and “graphene” as topic keywords to search for journal papers related to graphene from the Web of Science (Web of Science, n.d.). We found 13,649 papers using “single-layer-carbon,” in contrast to 127,546 papers found using “graphene.” Some 3,882 papers from the “single-layer carbon” collection were also found in the “graphene” collection. As long as the most highly cited papers are included, our collection does not have to be complete. Because of this, we decided to use only the “graphene” collection. As we will be using the topic clustering method to identify the graphene scientific field’s topics, we removed review papers from the collection, because these tend to include keywords associated with multiple topics and interfere with the topic identification process. We also removed conference proceeding papers, books, and other minor categories, because of their small numbers, so that we dealt consistently only with articles. In our analysis, we refer to the remaining article papers as the *G-S collection*. To answer our second scientific question on the parent streams of graphene science, we also collected bibliographic records of articles related to nanotubes and batteries from the Web of Science. We referred to these as the *NT-S* and *B-S collections*, respectively. The numbers of records for these three collections and the periods over which they are collected are shown in Table 1. In some records, we find the occurrences of two or more of the keywords used. The numbers of overlapping entries among the three collections are shown in Table 2.

At this point, let us clarify that we understand the benefits of working with large multidisciplinary data sets and using community detection/clustering methods to identify fields and topics at different levels of granularity. We know the limitations of working with a topical collection of papers obtained through a topic query. We also appreciate how results obtained from this collection cannot be put into the full context of allied topics. We chose to work with the graphene collection because our scientific interests are very focused, and we do not want to have to deal with the whole of science before narrowing it down to graphene. We partially eliminated the lack of context by downloading a nanotubes collection, a 2D materials collection, and later the batteries collection. It remains possible that other fields not included in this study may have significant contributions to the emergence of graphene as a field. However,

Table 1. Bibliographic records downloaded from the Web of Science using three keywords: “graphene,” “nanotube,” and “battery”

	Graphene (G-S)	Nanotube (NT-S)	Battery (B-S)
Number of records	115,988	168,224	119,482
Period	1991–2017	1992–2017	1900–2018

Table 2. Overlapping records in the G-S, NT-S, and B-S collections

Number of overlapping records	G-S	NT-S	B-S
G-S	–	31,214	10,632
NT-S	31,214	–	7,658
B-S	10,632	7,658	–

when they are identified, these fields will enrich our understanding of the abovementioned emergence instead of invalidating the results presented here.

Ultimately, we understand that while there is the potential of using clustering methods to find a cluster that can be unambiguously identified with graphene, the technique is not 100% foolproof. For example, a graphene paper that cites more nanotube papers will likely be clustered together with nanotube papers, even if it explicitly contains graphene in its title.

3.2. Methods

Based on our survey of the literature in Section 2.2.2, we should apply hybrid methods (citation-based and text-based) to identify graphene science research topics from the records we downloaded from the Web of Science. We initially tried Louvain community detection on the BCN in this work, but we were not satisfied with the results. When we broke the data set into yearly BCNs, the modularity values that we obtained ranged from 0.11 to 0.41. These are low compared to 0.40 to 0.55 (Adams & Light, 2014) or 0.48 to 0.85 (Fanelli & Glänzel, 2013), among others in the literature. These are especially low compared to our previous work (Liu et al., 2017), where we obtained modularities between 0.7 and 0.8 for the yearly BCNs. We then tried the text-based coclustering procedure described in Section 3.2.1 on the titles and abstracts of graphene papers and found the results acceptable. However, we believe that the results would be better if we use hybrid clustering methods and will try this in future research.

3.2.1. Text clustering

There have been many previous attempts to identify topics in a corpus of text. In the machine learning literature, several efficient methods for detecting topics have been proposed, including Latent Semantic Indexing (Deerwester, Dumais et al., 1990), Probabilistic Latent Semantic Analysis (Hofmann, 1999, 2001), and the widely popular Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003). We experimented with LDA but found that we needed first to specify the number of topics k . This means that we need another procedure to select the optimal k , and none of the existing ones seems natural. We also tried community detection on the word co-occurrence network using the Louvain algorithm (Blondel, Guillaume et al., 2008). This algorithm returns the maximum modularity Q , and the number of clusters k associated with it. However, Q is very low (0.09 for four communities). Eventually, we settled for CoClus, a novel block diagonal coclustering algorithm proposed by Ailem, Role, and Nadif (2015). This and other coclustering algorithms (Madeira & Oliveira, 2004; Van Mechelen, Bock, & De Boeck, 2004) have found applications in bio-informatics (Cheng & Church, 2000; Xie, Ma et al., 2020), web mining (Feng, Zhao, & Zhou, 2020; George & Merugu, 2005), and text mining (Celardo & Everett, 2020; Dhillon, 2001).

Our goal is to partition journal papers and the words they use into meaningful communities in the document-term bipartite network. The CoClus algorithm accomplishes this by directly maximizing the modularity, which measures the concentration of edges within each

community in comparison with the random ordering (Newman & Girvan, 2004). Given an object set $O = \{o_1, \dots, o_i, \dots, o_n\}$ and an attribute set $P = \{p_1, \dots, p_j, \dots, p_d\}$, we first decide how many clusters g we would like to partition I and J into. The goal of CoClus is to maximize the modified modularity

$$Q(A, C) = \frac{1}{\sum_{i,j} a_{ij}} \sum_{i=1}^n \sum_{j=1}^d \left(a_{ij} - \frac{\sum_{i=1, \dots, n} a_{ij} \sum_{j=1, \dots, d} a_{ij}}{\sum_{i,j} a_{ij}} \right) c_{ij}$$

over all possible partitions of $O \times P$ into g clusters $\{(O^k, P^k)\}_{k=1, \dots, g}$. Here, $A = \{a_{ij}\}_{i=1, \dots, n; j=1, \dots, d}$, a_{ij} is a weight that tells us how strongly attribute p_j is associated with object o_i , C is a characteristic matrix with elements $c_{ij} = 1$ if object o_i and attribute p_j are in the same cluster (O^k, P^k) . Readers are encouraged to study the example given by Ailem et al. (2015) to understand the procedure behind this coclustering algorithm better.

To apply CoClus to our G-S collection, we assumed that the topic of an article could be inferred from the linguistic content in the title and abstract. We first used the regular expression package (Friedl, 2006) to remove nonalphabetical contents. We then filtered out stop words using two packages: scikit-learn and Natural Language Toolkit (378 stop words) and stemmed the words by using Porter stemmer (Porter, 1980) from the Natural Language Toolkit package, before counting the number of times $a_{i,j}$ the stemmed word j ($j = 1, \dots, n$) appears in the document i ($i = 1, \dots, d$). Finally, we filtered out words that appeared in less than 0.01% of the total data set, keeping $\sim 12,000$ words for the coclustering. Most likely, we will need only a small number of words to describe the topic of each cluster, but we start with a large number of words to avoid throwing some of these out too early. After all these preprocessing steps were completed, we organized the frequencies $a_{i,j}$ into a document-term matrix:

$$A = \begin{pmatrix} a_{1,1} & \dots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{d,1} & \dots & a_{d,n} \end{pmatrix}.$$

Finally, we applied CoClus to A for a range of communities $k = 2, \dots, 9$ to determine the value g^* corresponding to the maximum modularity.

3.2.2. Communities validation

After the communities were tested to be robust, we next validated them to ensure they were also meaningful. We do so in two ways: (a) by extracting a list of the most important keywords, sorted according to a z-score that we will describe next, and (b) by inspection of the titles and abstracts of highly cited papers. In the latter, we checked that papers highly cite these papers within their respective communities 3 years after their publications. In this sense, these highly cited papers can be thought of as representatives of their communities.

3.2.2.1. z-score method for keyword identification In the keyword identification literature, methods like TF-IDF (Spärck Jones, 1972) identify keywords that frequently appear in a document but infrequently appear in other documents within the corpus. For our problem, in addition to these document-specific keywords, we also encountered keywords that frequently appear not only in a single document but with high probability in many documents belonging to the same community. We are less interested in the former but more interested in the latter. In other words, we are interested in keywords that would describe the community but not individual papers in the community. In general, we expect only a small number of communities, each containing on the order of 10,000 documents. In such situations, methods such as TF-IDF will tend to exclude community keywords but pick up keywords of individual papers instead.

The standard way to deal with this would be to simultaneously model these communities' topics and use the topics as the basis for identifying them. Instead of going to topic modeling methods, we developed a simple method to identify these community keywords.

To begin, after we have partitioned our collection of d G-S articles, with n keywords into g nonoverlapping document-term communities $\{(O^k, P^k)\}_{k=1, \dots, g}$, we compared how likely it is that individual keywords can be explained by a null model where each word is equally likely to appear in any document. In other words, for word j assigned to P^k , let m_j be the number of articles belonging to O^k that word j appears in, and M_j be the number of articles in the entire corpus that word j appears in. In this null model, we do not care how many times the word j appears in a given document, so long as it appears. Therefore, the probability of word j appearing in an article would be $p_j = \frac{M_j}{d}$. Hence, we expect word j to appear $\mu_j = |O^k| \cdot p_j$ times on average in the articles of O^k , with a standard deviation of $\sigma_j = [|O^k| \cdot p_j \cdot (1 - p_j)]^{\frac{1}{2}}$. Empirically, word j appearing m_j times in O^k is highly significant if $m_j > \mu_j$ in relation to σ_j . To quantify how effective word j is, we define its z-score to be

$$z_j = \frac{m_j - \mu_j}{\sigma_j}.$$

Expecting words with large z-score values to be highly representative of the clusters these words belong to, we are now ready to pick small sets of keywords that would describe the topics of the clusters. We do this by keeping words that have high z-scores and probabilities $\frac{m_j}{|O^k|}$ within the 98th percentile in their respective clusters.

3.2.2.2. Titles and abstracts of top papers Although the most important keywords can be discovered using the z-score method, these keywords have been taken out of their contexts and may be difficult to interpret. Therefore, we also looked at the titles and abstracts of G-S articles that are highly cited by articles within their communities 3 years after their publications. As these articles represent their respective communities, we believe that we can infer the topics from them.

4. RESULTS AND DISCUSSION

4.1. Robust Research Topics Within Graphene Science

After coclustering the collection of G-S papers, we plot the modularity as a function of the number of communities in Figure 2. Ideally, we should choose the number of communities that maximizes modularity. However, in Figure 2, we see that the modularity peaks for $n = 4$ and $n = 6$ communities. Therefore, we applied the principle of Occam's Razor to accept the smaller number of communities. The sparsity plot after reorganizing the G-S papers into $n = 4$ document-term clusters is shown in Figure 3. In Section C of the Supplementary material, we showed that these four document-term clusters are robust.

4.2. Keywords and Validation for Graphene Research Topics

After checking that the clusters were robust, we proceeded to assign topics to them. Figure 4 shows the lists of the top 25 keywords of four G-S topics, sorted according to their z-scores. In Group (0), we found keywords like "modulu" (i.e., modulus), "tensil" (tensile), "wt" (short form for "weight," commonly used when talking about weight percentage), "thermal," and "load." Some of these keywords are associated with material synthesis, while others are related to material characterization, which is commonly done after new materials are synthesized. In

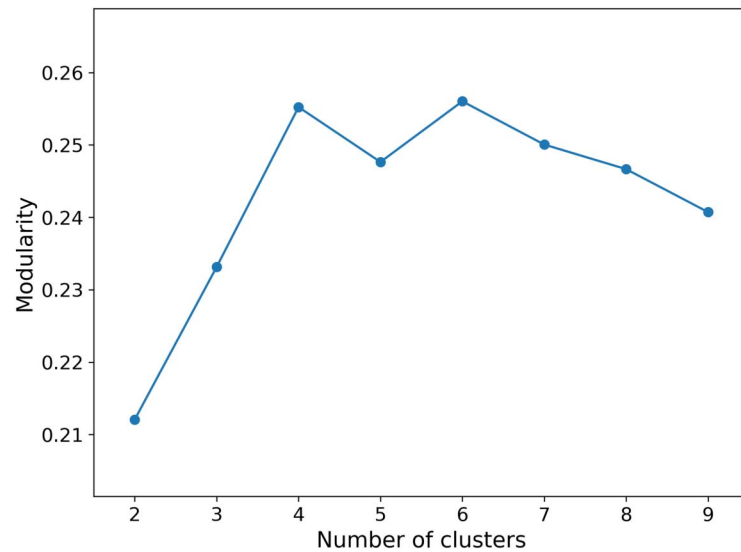


Figure 2. Plot of the modularity discovered by the CoClus algorithm for organizing G-S papers into the different number of communities.

Group (1), we found keywords like “batteri” (i.e., battery), “lithium,” “supercapacitor,” “stomag” (i.e., storage), “capac” (capacity), and “electrod” (electrode). All these keywords are associated with using graphene to make supercapacitors, which can be thought of as a physical analog of batteries. In Group (2), we found keywords like “detect,” “sensor,” “sensit” (i.e., sensitive), and “sampl” (samples). These are related to the application of graphene in sensors. Finally, in Group (3), we found keywords like “dirac,” “theori” (i.e., theory), “gap,” “principl” (principle), “spin,” “calcul” (calculations), “band,” “simul” (simulations), and “theoret” (theoretical). These



Figure 3. Sparsity plot of the G-S papers and potential keywords that appear in them after they have been reorganized into $n = 4$ document-term communities by the CoClus algorithm. In this figure, the pixel along the i th row and in the j th column is colored blue if term j appears in document i , or white otherwise. As we can see, the potential keywords are not uniformly distributed across the documents. Instead, a cluster of possible keywords is preferentially found in one cluster of documents and less so in the other document clusters. Hence, the four diagonal blocks marked by red dashed lines are darker than the off-diagonal parts of the matrix.

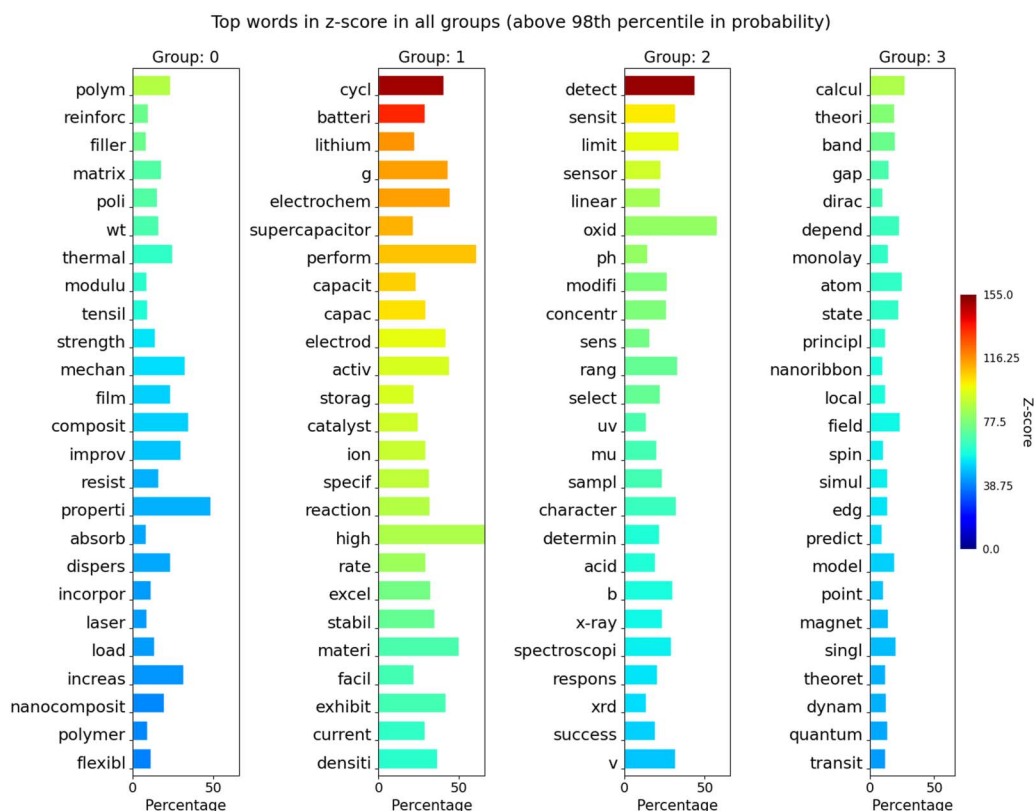


Figure 4. The lists of the top 25 keywords for the four G-S topics, sorted according to their z-scores. For each keyword, we also show a bar (color-coded according to the z-score of the keyword) whose length represents the percentage of documents in the cluster containing the keyword.

are keywords that appear commonly in theoretical and simulation papers on graphene. Therefore, based on these top keywords as well as the word clouds shown in Figure S6 to Figure S9 in the Supplementary material, we named the four topics as (0) *synthesis*, (1) *supercapacitors*, (2) *sensors*, and (3) *theory and simulation*.

To validate our assignment of topics, we looked at the five papers from within each topic with the highest number of citations 3 years after their publications. As shown in Table S2 in the Supplementary material, the titles of three (10.1038/nnano.2010.132, 10.1038/nnano.2007.451, 10.1038/nmat3944) of the top five papers in Group (0) are clearly about the synthesis of graphene. The titles of the other two are on graphene-based polymer composites, one of the popular approaches to functionalize graphene. In the Supplementary material, we list up to the top 20 papers in each group, and we find more Group (0) papers whose titles are on the functionalization of graphene. Therefore, we changed the topic of Group (0) to *synthesis and functionalization*. For Group (1), the titles of two (10.1126/science.1200770, 10.1126/science.1216744) of the papers confirm that they are on the supercapacitor application of graphene. However, the titles of the remaining three are on the electrocatalyst application of graphene. Therefore, we renamed the topic of Group (1) to *supercapacitors and electrocatalysts*. For Group (2), four papers contain “sensor” or “sensing” in their titles. The one paper (10.1021/nn901221k) whose title does not contain “sensor” or “sensing” is also a sensor paper because any application of graphene as a sensor for visible light requires some chemical groups (P25-Graphene Composite) to be sensitive to light. For Group (3), the most

highly cited paper is, in fact a review of sorts, even though it was not classified as such. Ignoring this paper, two (10.1038/nature04233, 10.1038/nature12385) of these papers are theoretical papers, and one of them (10.1038/nature04235) is an experimental paper that tested a specific theoretical prediction on graphene. From the top 20 papers listed in the Supplementary material for this group, we find many more experimental papers focused on testing various theoretical predictions. Therefore, we rename this group *theory and experimental tests*.

For the rest of this paper, we worked with (3) *theory and experimental tests* and (1) *supercapacitors and electrocatalysts* as case studies. We made these choices for the following reasons: (a) we have one pure case study, *theory and experimental tests*, and one applied case study, *supercapacitors and electrocatalysts*; (b) we chose *theory and experimental tests* over *synthesis and functionalization* because topic (3) remains largely the same even if we chose to work with $n = 6$ topics; (c) we chose *supercapacitors and electrocatalysts* over *sensors* because the former breaks up mainly into two of $n = 6$ topics, but the latter breaks up into at least three of $n = 6$ topics (see Figures S10, S11 in the Supplementary material).

4.3. The Sequence of Emergences of Graphene Research Topics

To answer our first scientific question on the sequence of emergences of the graphene research topics, and perhaps also to understand the logical reasons behind this sequence, we first plotted the number of G-S papers for each topic and its proportion among all G-S papers over the years. Figure 5(B) suggests that Group (3) is the first G-S topic to emerge. The number of G-S papers in the other three topics started increasing around the same time, making it difficult to tell the order of their emergences. When we plotted the proportions of G-S papers over the years for the four topics, the *theory and experimental tests* curve peaked first. The proportion of

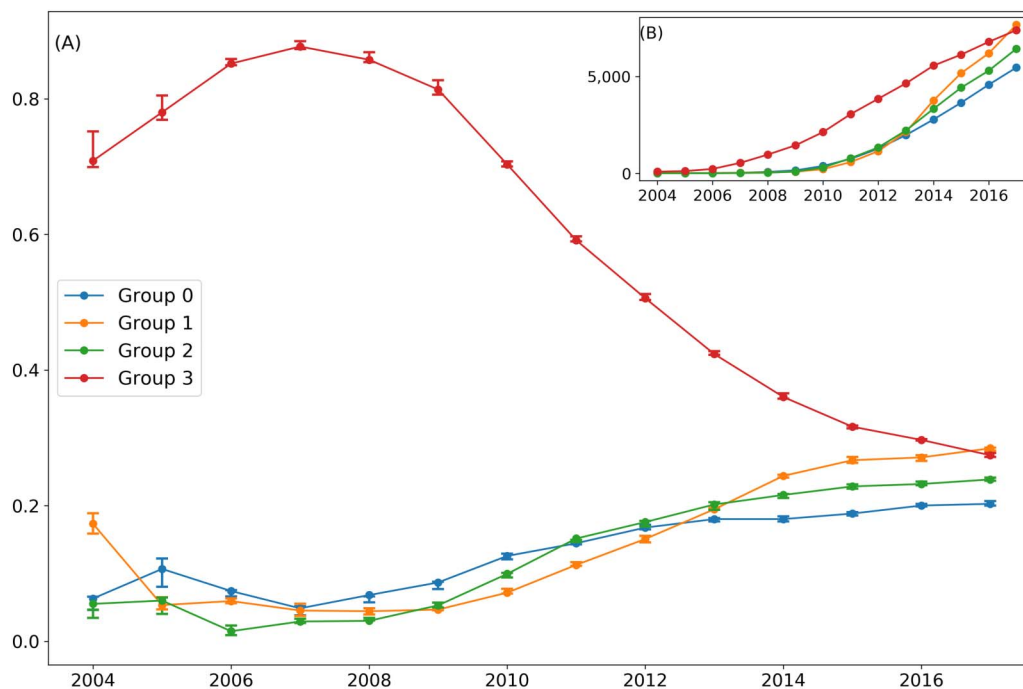


Figure 5. (A) Publication proportions of four topics from 2004 to 2017, where the error bars represent 25th–75th percentile boundaries from the respective proportions of 10 document subsets, and (B) the actual number of publications in the four topics after coclustering the full data set.

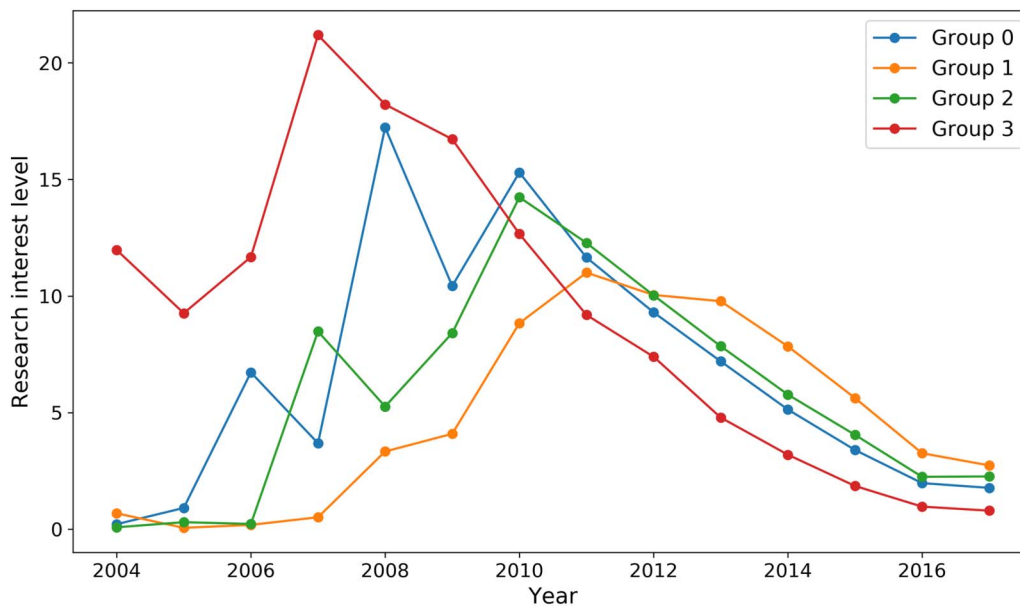


Figure 6. The levels of scientific interest on the four G-S topics.

G-S papers belonging to Group (0) is the second highest from 2005 to 2010, while Group (2) is second highest between 2011 and 2013. Finally, after a single year as the second highest in 2004, the proportion of G-S papers for Group (1) became consistently second highest from 2014 to 2016 and eventually became the highest in 2017. These suggested that the order of emergences was (3)-(0)-(2)-(1).

In our previous paper (Nguyen et al., 2020), we figured out how to deal with the problem of plotting bibliometric quantities that increased with time by measuring the average rate at which papers published in particular years are attracting citations. We do the same here to find the scientific interest in the four topics, first to rise and then fall. We explained how the interest curve could be computed in Section G of the Supplementary material. From Figure 6, we see Group (3) *theory and experimental tests* peaking clearly in 2007. For (0) *synthesis and functionalization*, the global peak was in 2008, and the global peak for (2) *sensors* was 2 years later, in 2010. Finally, for (1) *supercapacitors and electrocatalysts*, the global peak was in 2011. This suggests that pure theoretical research preceded pure experimental research, an order commonly seen in science. We also see a collective shift in the interest from pure topics to applied topics, again in agreement with the standard model of innovation.

This answers our first scientific question on which topic emerges first, and which other topics follow.

4.4. Incubation and Emergence Analysis

Our second scientific question is on the parents of the G-S topics: where their seeds were first planted and where they incubated in their embryonic stages before they emerged as independent topics. To be a parent of a G-S topic, the parent must have become a separate topic earlier. With this in mind, we reasoned that NT-S could potentially be the parent of one or more of the G-S topics. At the same time, B-S can potentially be the parent of (1) *supercapacitors and electrocatalysts*. It is conceivable that a G-S topic can have more than one parent, and therefore there is the possibility of parent fields beyond NT-S and B-S. At this early stage of

our research, we chose not to worry whether we have a complete or overcomplete list of candidate parents to test but simply focus on testing which indicators are better at confirming a potential parent. If other fields are suggested to be parents of the G-S topics, we would like the indicators to tell us convincingly that this is the case.

4.4.1. Explorers Versus specialists

Intuitively speaking, if topic B is incubated in topic A before emerging as an independent stream, we expect that the first scientists to dabble in topic B would be from topic A. We called such scientists *explorers*. In some sense, the existence of this group of scientists has been anticipated by Salatino, Osborne, and Motta (2017), when they proved the existence of embryonic stages of emergent topics in Computer Science by measuring the density of cross-references between mature topics. At the same time, most scientists in topic A would have no interest in topic B. Similarly, topic B would start attracting *specialists* who do not publish in topic A after its emergence. Therefore, we first identified the set of all authors working on topics A or B and split them into three disjoint subsets: (a) those working on topic A only, (b) those working on topic B only, and (c) those working on both topics A and B. From a data processing perspective, we do not wish to include opportunistic authors who publish one paper on a topic once every few years because they make the data noisy when the number of committed researchers is small. Therefore, an author will be counted if he or she publishes at least two papers a year for at least two years. For example, if an author publishes three papers in 2005 and one paper in 2007, this author is excluded from the count. On the other hand, if another author publishes two papers in 2006, one paper in 2007, and three papers in 2009, this author is included in the count for 2006 and 2009, but not for 2007. We expect that the number of authors from a subset (b) would be lower than those from a subset (c) but eventually become higher. Therefore, the signature we should look out for comprises a crossing between the number of authors working on topic B and the number of authors working on both topics.

We do this first for the four topics compared against NT-S. A cross-over from explorer-dominated to specialist-dominated is clearest when we plot the numbers of authors on the semilog scale (Figure 7). When we do this for groups (0), (1), and (2), the numbers of authors

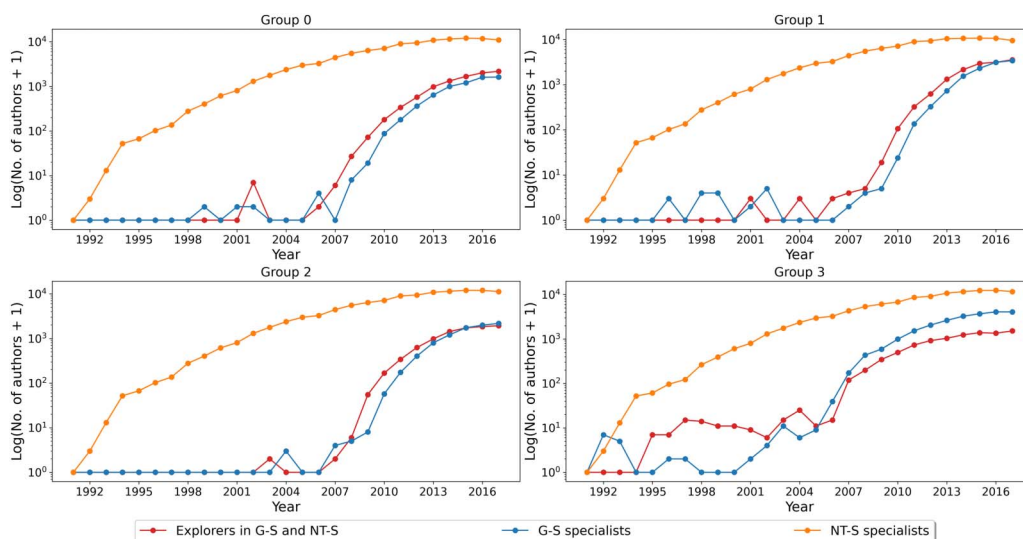


Figure 7. The number of authors who have publications in G-S only (G-S specialists), NT-S only (NT-S specialists), and in both collections (G-S and NT-S explorers) for each year in four G-S topics.

in subset (b) publishing in G-S only and subset (c) publishing in both G-S and NT-S started increasing rapidly between 2005 and 2007. The only exception is group (3), *theory and experimental tests*, where we find authors from subset (c) publishing as early as 1995. The number of such authors rose rapidly after 2006 but was overtaken by authors from subset (ii) around this time. This suggests that G-S *theory and experimental tests*' parent stream is likely to be a similar stream within NT-S.

Next, we compared the four topics against B-S (Figure 8). For groups (0), (2), and (3), the numbers of specialist authors rose before the number of explorer authors. This tells us that B-S could not be the parent stream of these topics. For group (1), the persistent rise of specialist authors also preceded the corresponding continuous rise of explorer authors, but earlier episodes of interdisciplinary exploration occurred between 1996 and 2002. This signature is weak. Thus, we can say that B-S is likely to be a secondary parent stream for G-S group (1) *supercapacitors and electrocatalysts*, with a yet unidentified primary parent stream.

In general, the name of an author can appear differently in different records. For example, some journals may publish only the first and last names of an author, whereas other journals may include the middle name(s) of the author. Some journals publish the full name of an author, whereas other journals may publish only the last name in full, and use initials for the first and middle names. For some regions, there might also be different authors with the same English name, because their distinct native names might be Romanized the same way. There might even be authors with the same native names. The problem of figuring out the distinct individuals' different names correspond to is called *disambiguation*. We arrived at the above results by using the author full names as they were extracted from the Web of Science records (i.e., the results were obtained without disambiguation). In Section H of the Supplementary material, we described the disambiguation algorithm proposed by Sinatra, Wang et al. (2016), and redid the analysis. The results we obtained after disambiguation are slightly different from the ones before disambiguation, but the differences are not enough to change our conclusions. This robustness of our author-based analyses towards disambiguation is also true for the results we present in Section 4.4.2 and Section 4.5.

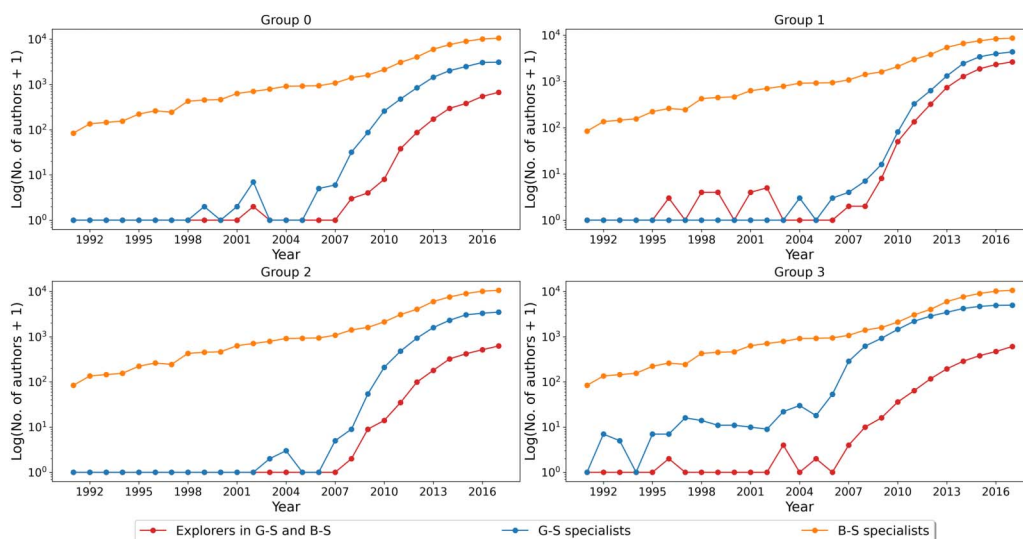


Figure 8. The number of authors who have publications in G-S only (G-S specialists), B-S only (B-S specialists), and in both collections (G-S and B-S explorers) for each year in four G-S topics.

4.4.2. Debutant authors

As we can see from Figures 7 and 8, all three subsets of authors are increasing with time. This means that subtle changes in the rate of increase would be hard to detect. A good indicator should rise and then fall with time, like the interest curves shown in Figure 6. However, we cannot use the interest curve here because it was not designed to identify the incubation period and its emergence. We therefore propose a second indicator that we feel provides more information and has the property of rising and then falling. To motivate this second indicator, let us observe that the numbers of explorer and specialist authors in a given year consist of those who are publishing their first papers in the field, together with those who are publishing their second, third, ... papers in the field. This adds unnecessary noise to our first indicator. In general, a productive scientist would have worked on a succession of topics over his or her (ongoing) career. For each topic, this scientist must have published a *debut* or *maiden* paper, whatever career stage the scientist might be at. For example, the famous physicist Richard Feynman worked on quantum electrodynamics (the quantum theory describing the interactions between electromagnetic waves and charged particles) as well as superfluidity (a quantum phenomenon in which liquid helium becomes nonviscous when cooled below a certain critical temperature), among other topics over his long and prolific career. Feynman first published on quantum electrodynamics in 1949 (Feynman, 1949), and his last paper on this topic was in 1954 (Feynman & Speisman, 1954). For his work on this topic, he shared the 1965 Nobel Prize in Physics with Julian Schwinger and Sin-Itiro Tomonaga. Similarly, Feynman's first paper on superfluidity was in 1953 (Feynman, 1953), and he last published on this topic in 1958 (Feynman, 1958). We can then say that Feynman's debut paper in quantum electrodynamics was in 1949, while his debut paper in superfluidity was in 1953. For G-S, we identify the debut year for all scientists who have at least two publications a year for at least two years, and we plot the number of debuts in different years for these debutant authors.

For a field where not much is happening, we expect a constant rate of debuts. On the other hand, for an emerging field that is attracting a lot of attention, we expect an increase in the rate of debuts in the few years immediately following the emergence, but as the rate of discovery slows in this field, the rate of debuts also drops to a low level. We will track this rate of debuts for the same three subsets of authors. For a mature parent stream, we expect the rate of debuts to be constant, whereas, for the child stream, we expect the pace of debuts to rise and peak and then fall back to a low level after the emergence. For scientists that work on both topics, they will debut in one or the other topic. Some of these scientists would do so in the incubation phase, so therefore we expect the debut curve of these explorer scientists to rise and perhaps peak earlier than the debut curve of the child stream.

From Figure 9, we see all four G-S topics showing clear incubation signatures (i.e., the debut curve of explorer authors rises and peaks earlier than the debut curve of the specialist G-S authors). Suppose we define the incubation period to be the time the explorer debut curve is higher than the G-S specialist debut curve. In that case, we find that this varies across the four topics, from 3 years in G-S *theory and experimental tests* to 6 years in G-S *supercapacitors and electrocatalysts* and G-S *synthesis and functionalization*. This suggests that the undifferentiated NT-S is the parent stream of all four G-S topics. Surprisingly, the debut curve for NT-S is not constant but shows a significant jump between 2005 and 2006. We will have more to say on this phenomenon in Section 4.5.

When we used the same indicator to test the G-S topics against B-S, we find from Figure 10 no incubation signatures from topics (0), (2), and (3). Even for (1) G-S *supercapacitors and electrocatalysts*, which we expected from logical considerations to have incubated in B-S, the

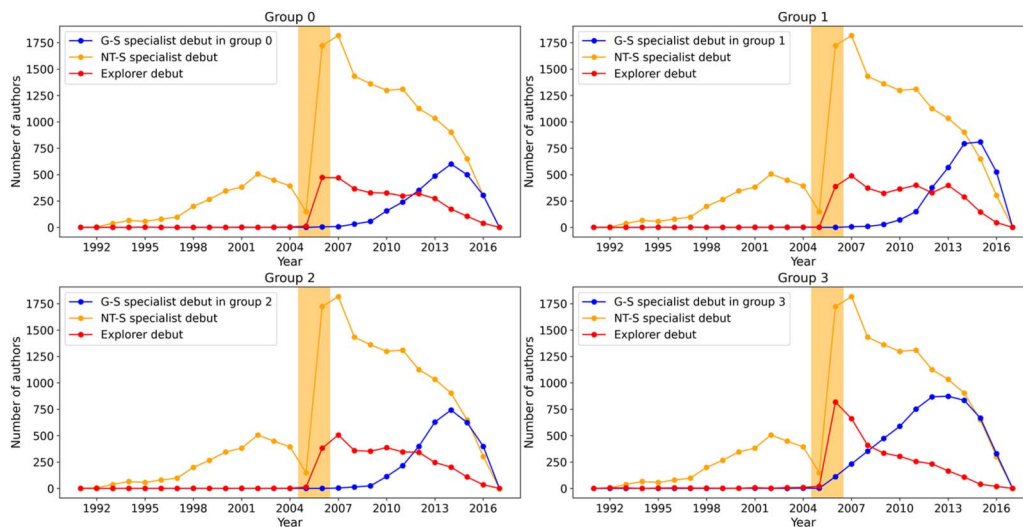


Figure 9. The number of specialist and explorer authors in their debut years from four G-S topics and NT-S. An explorer can debut in either G-S or NT-S, whichever is earlier.

incubation signature is much weaker than the corresponding one in the test against NT-S. Even more unexpectedly, there was a jump in the debut rate of B-S between 2005 and 2006. Before this jump, the debut rate was increasing slowly, as a result of the increases of the yearly numbers of G-S, NT-S, and B-S papers. This is shown in Figure S44 of the Supplementary material. From Figure 9, we see that the number of debutant NT-S authors was increasing from 1992 to 2004 as expected, based on new authors joining NT-S every year. However, the number of debutant authors dropped sharply in 2005 and increased sharply to a maximum in 2007 before starting to decrease. This was in spite of the number of NT-S papers published increasing monotonically over this period.

From Figure 10, we see that the number of debutant B-S authors increased from 1991 to 2001 and decreased from 2001 to 2005. This then rose sharply in 2006 and fell for 2 years,

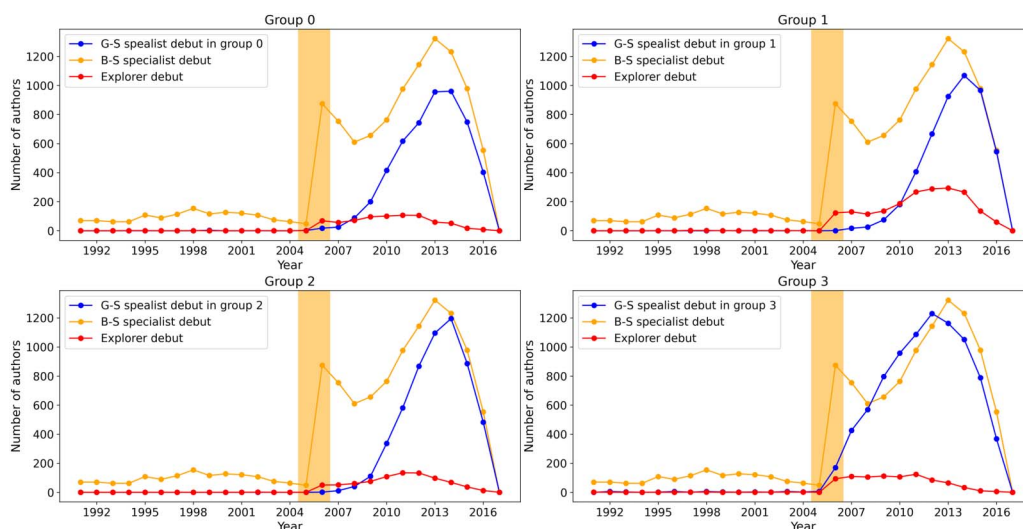


Figure 10. The number of specialist and explorer authors in their debut years from four G-S topics and B-S. An explorer can debut in either G-S or B-S, whichever is earlier.

before increasing strongly to a maximum in 2013 before decreasing thereafter. If we look at Figure S44, we see again that there is no slowing down in the yearly numbers of B-S papers over this period. Therefore, when the number of papers in a given field is increasing, the number of debut authors in the same field need not share the same increasing trend.

In fact, the rate of increase of debut NT-S papers in Figure 9 shortly after 2006 appears to be faster than the rate of increase of debut NT-S papers before 2004. While the numbers for 2016 and 2017 may not be accurate because of our stringent criterion for admitting debutant authors, the sharp rise follow by sharp fall in the number of debutant authors seen in Figures 9 and 10 for certain groups appears to be a genuine phenomenon. Instead of the number of debutant authors increasing in proportion with the yearly number of publications, the sudden increase in debutant authors over and above that due to the growth of the yearly number of publications is most likely due to the sudden increase in interest in the particular topic. Therefore, as expected, they are always debut authors every year, but the rate of debuts is not constant over time, but reacts very sensitively to sudden surges in interest (such as the emergence of a new field).

Finally, we did a sanity check in Section I of the Supplementary material, to confirm that the four G-S topics are not parents of each other. We also showed that results from the two indicators are robust. For the first indicator, we omit 10% of publications in the three subsets before plotting Figures S33 and S34 shown in Section J of the Supplementary material. We arrived at the same conclusions as those from Figures 7 and 8. For the second indicator, we first identified the debutant publications and removed 10% of them. We then went through the remaining data set to determine the debutant publications anew, to plot Figures S35 and S36 shown in the Supplementary material. Again, we found no changes to the conclusions we have arrived at from Figures 9 and 10.

Thus, on the scientific question of who the parents of the G-S topics are, we found that NT-S is the primary parent of all four G-S topics. In contrast, B-S is a secondary parent of G-S group (1) *supercapacitors and electrocatalysts*.

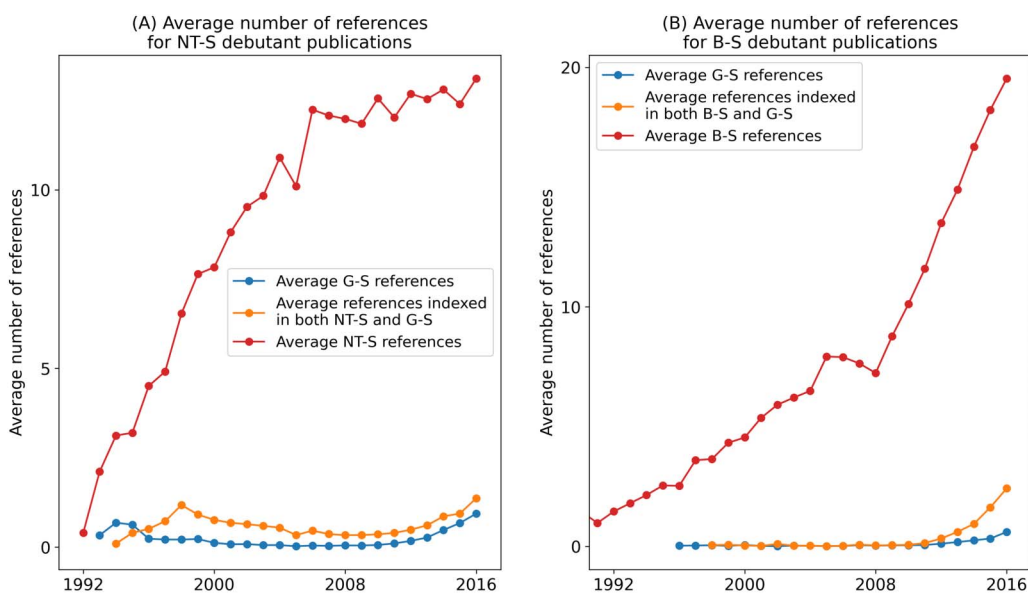


Figure 11. The average citations for debutant publications from (A) NT-S and from (B) B-S.

4.5. Interaction Analysis

For an emergent field, we expect its incubation and eventual emergence to be driven by exogenous events, such as breakthroughs in distant fields noticed by the parent stream. We call these *pre-emergence interactions*. These breakthroughs seeded the incubation of the emerging stream, which then grew to become an independent stream or faded after a series of failures. Salatino et al. (2017) measured these interactions at the aggregate level. When the new stream emerged, it would also be seen as a breakthrough by other streams and might seed emergent streams elsewhere. We call these *postemergence interactions*. This is the picture of interactions between streams that we believe can enrich the theory of innovation.

As the G-S topics are emergent streams, we expected to find evidence of pre-emergence (between stream X and the parent streams of the G-S topics) and postemergence (between the G-S topics and stream Y) interactions. We also expected the signatures of these interactions to be weak. Therefore, we were surprised by sharp increases in specialist authors publishing for the first time in NT-S/B-S. This signature is much stronger than we expected. In the next part of this paper, we focus on demonstrating this signature's connection to the 2004 G-S breakthrough (Novoselov et al., 2004).

First, for each specialist author, we identified the first paper they published in NT-S/B-S. Then, the numbers of such debut specialist papers were more or less constant over the years until 2004, when we saw a sharp increase. If this sharp increase is due to the graphene breakthrough in 2004, then we would expect to find in these debut specialist papers no references to G-S papers before the 2004 breakthrough, and a very sudden increase in the number of references to G-S papers shortly afterwards (1–3 years). Therefore, plotting the average number of G-S references in the NT-S/B-S debut specialist papers, we expected a pulse right after 2004 that decayed gradually to very low levels. However, in Figure 11, we do not see any sharp changes to the average number of references to G-S papers around 2004. On the contrary, we see NT-S debut specialist papers citing on average more NT-S papers in 2006 and B-S debut specialist papers citing on average more B-S papers in 2005.

Upon closer inspection, we find many references to a small collection of highly cited G-S papers published around 1992. It turned out that these were considered a breakthrough by both NT-S and G-S. These consisted of theoretical papers explaining why specific nanotubes are metallic while others are semiconducting (Saito, Fujita et al., 1992a, 1992b). Over the years, these papers indexed by Web of Science as belonging to both NT-S and G-S were highly cited by papers from both streams, so much so that the signature of the 2004 G-S breakthrough was masked. Unlike the 1992 breakthrough, the 2004 breakthrough involved demonstrations of the feasibility of preparing single-layer graphene samples. To see the signature of the latter more clearly, we grouped the G-S references into eight time windows: 1992–1994, 1995–1997, 1998–2000, 2001–2003, 2004–2006, 2007–2010, 2011–2013, and 2014–2016, and counted the number of times these eight groups of G-S papers were cited on average in NT-S debut specialist papers published in different years. We expected more references to the 1992–1994 and 2004–2006 groups of G-S papers than G-S papers in the other time windows because of their temporal proximities to the 1992 and 2004 breakthroughs.

As expected, we can see from Figure 12 that the 1992–1994 group of G-S references containing the first series of breakthrough papers were more highly cited than the 1995–1997, 1998–2000, and 2001–2003 groups. However, the average number of citations of these papers decreases with time, starting with more than one in every 10 NT-S debut specialist papers, suggesting that the NT-S community was losing interest in these papers as time went on. We suspected that the 1995–1997, 1998–2000, and 2001–2003 groups of G-S references were likely

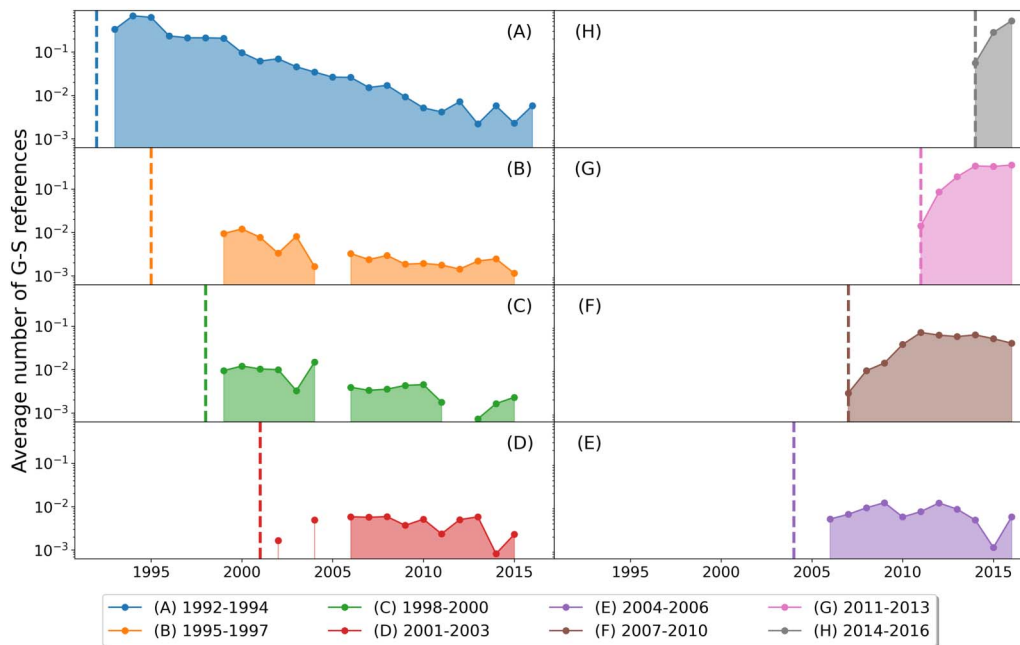


Figure 12. For each of the eight time windows, we plot the number of G-S references within the time window, averaged over the number of NT-S debut specialist publications citing them. The horizontal axis gives the publication years of the NT-S debut specialist papers, while the vertical axis gives the average number of G-S references in these years. In these figures, the dashed vertical lines indicate the starting years of the groups of G-S publications. For the latter groups, there were citations from as early as the starting years of the groups, whereas for the earlier groups, we see delays of 1–3 years.

to be following up on the 1992 breakthrough but contained no significant innovations of their own. In contrast, the average number of citations started at a level of 10^{-2} for the 2004–2006 group of G-S references, increased to a level of 10^{-1} for the 2011–2013 group of G-S references, and eventually to a level $>10^{-1}$ for the 2014–2016 group of G-S references. This suggests a growing interest from the NT-S community in the 2004 breakthrough and studies that follow up on it.

We repeated this analysis for B-S debut specialist publications in Figure 13. Unlike for NT-S, the average number of G-S references for the 1992–1994 group is sporadic (i.e., present in some years but absent in the rest). This average number of G-S references is then low and roughly constant in time for the 1995–1997, 1998–2000, and 2001–2003 groups. This suggests that G-S’s highly theoretical 1992 breakthrough was not attractive to the B-S community, which is understandable. The average number of citations ($<10^{-2}$) to the 2004–2006 group of G-S references (containing the 2004 breakthrough papers) is also almost constant in time and is comparable to that in the 2001–2003 group. From the 2007–2010 group onwards, the average number of citations to G-S references started increasing, indicating an increase in the interest from the B-S community. This increase was most notable in 2010 when Andre Geim and Kostya Novoselov won the Nobel Prize in Physics. Comparing this slow increase after 2010 with the sharp rise in 2006 that we saw in Figure 10, we realized a slight discrepancy that we needed to explain. To do this, we checked all the post-2004 groups carefully but saw no strong signature in 2006. Therefore, we cannot conclude that the sharp rise in the number of B-S debut specialist papers in 2006 was directly caused by the 2004 breakthrough in G-S, but we cannot rule out that this is an indirect phenomenon.

The story we discovered here is that NT-S was influenced by and followed up on the 1992 breakthrough (many of the papers are labeled as both NT-S and G-S by the Web of Science)

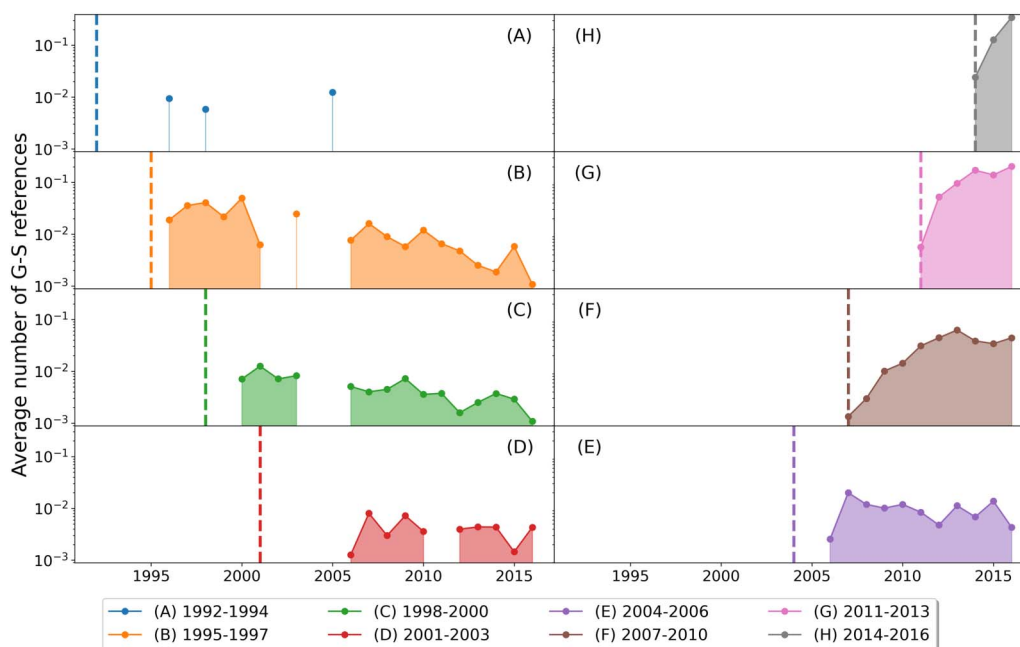


Figure 13. For each of the eight time windows, we plot the number of G-S references within the time window, averaged over the number of B-S debut specialist publications citing them. The horizontal axis gives the publication years of the B-S debut specialist papers, while the vertical axis gives the average number of G-S references in these years. In these figures, the dashed vertical lines indicate the starting years of the groups of G-S publications. For the latter groups after 2007, there were citations from as early as the starting years of the groups, whereas we see delays of one to four years for the earlier groups.

before the 2004 G-S breakthrough came along. The NT-S community then reacted vigorously to this breakthrough in the form of a sharp jump in the NT-S debut specialist curve in 2006. The evidence for this reaction being driven by the 2004 G-S breakthrough can be seen from the change in the time-dependence of the average number of citations to G-S references by NT-S debut specialist papers.

In contrast, before the 2004 G-S breakthrough, the B-S community showed little interest in what was happening in G-S. From the debut specialist curve of B-S shown in Figure 10, we see signs of increased interest starting in 2006. In contrast, Figure 13 showed a more gradual change in attitude beginning in 2010 when the Nobel Prize in Physics was awarded for the discovery of graphene. More importantly, the B-S community was citing G-S references published several years after the 2004 breakthrough. Let's look specifically at the G-S references of B-S debut specialist publications in 2011. The most dominant proportion is from (1) *supercapacitors and electrocatalysts* (79.3%), while 10.9% are from (3) *theory and experimental tests*, 8.4% and 1.5% are from (0) *synthesis and functionalization* and (2) *sensors*, respectively. This suggests that the B-S community reacted to the rise of the G-S (1) *supercapacitors and electrocatalysts* topic in 2007 (see Figure 6) instead of directly responding to the 2004 breakthrough.

The same story is borne out when we plotted the citation profiles of G-S and non-G-S references of NT-S and B-S publications in different years. We show these series of citation profiles in Figures S37 and S38 in the Supplementary material. In summary, we found clear evidence of postemergence interactions between the 2004 G-S breakthrough and NT-S. In contrast, for B-S, the nature of its postemergence interactions with G-S appears to be more complex. We also tried to answer our third scientific question on what pre-emergence interactions seeded our two case studies. We did so by looking at the references of the explorers'

publications that were outside of NT-S and G-S but were at the same time rarely cited by the NT-S specialists. Figure S39 of the Supplementary material showed how the explorers cited more references from outside G-S and NT-S. For example, in the subset of G-S debut explorer publications from *supercapacitors and electrocatalysts*, we found between 2.5 and 5 references to the B-S specialist collection during the incubation period of this topic. Outside of the B-S collection, we also found references to papers on the synthesis of graphitic oxide (Hummers & Offeman, 1958), which were included in Figure S41 of the Supplementary material. Explorers in the other three topics did not cite papers in the B-S collection. For the explorers of G-S *theory and experimental tests*, we looked at their most highly cited references that are not in the NT-S and G-S collections and found that the most crucial subset (one in a hundred to one in ten) consists of simulation method papers (Plimpton, 1995; Stuart, Tutein, & Harrison, 2000) (see Figure S43 in the Supplementary material). These papers were cited with increasing consistency over the incubation periods and were likely to be the seeds of the topics. As expected, these signatures of pre-emergence interactions are weak.

5. CONCLUSIONS AND OUTLOOK

This paper described a stream-based picture of the scientific innovation process, focusing on the interplay between pure and applied research. Instead of the simple picture of an applied research stream emerging from the pure research stream, we argued that for an emerging field like graphene science, both its pure and applied streams must have emerged from different parents after incubation periods seeded through interactions between mature streams. The scientific questions we aimed to answer are: (a) Is it true that pure research streams emerge before applied research streams?; (b) What are the parents of the pure and applied G-S research streams?; and (c) Can we identify the interactions between streams that led to the emergence of G-S as a field? In principle, to answer these questions, we need to collect data on all interactions (including social interactions) between scientists, for example, the conferences, workshops, symposia, and seminars they attended, the papers they read (including those they do not cite), and research visits to other universities or research institutions. However, these data are hard to come by, and thus we limited ourselves to publications.

Using a method of coclustering to analyze the linguistic information within the titles and abstracts of a collection of G-S papers downloaded from the Web of Science, we found that the G-S papers can be organized into four topics: (0) *synthesis and functionalization*, (1) *supercapacitors and electrocatalysts*, (2) *sensors*, and (3) *theory and experimental tests*. These topics were tested and found to be robust and also meaningful. Topics (0) and (3) can be considered pure G-S research, whereas (1) and (2) are applied G-S research. By plotting the proportions of G-S papers as well as the interest curves in the four topics, we found that the topics emerged in the order (3), (0), (2), and (1), in agreement with our expectation that pure research topics emerge before applied research topics. This is our answer to the first scientific question.

For each G-S topic, we then plotted the number of authors publishing exclusively in G-S, NT-S/B-S, as well as the number of authors publishing in both G-S and NT-S/B-S over the years and found for the topic (3) that the number of authors publishing in both G-S and NT-S increased first before the number of authors publishing exclusively in G-S *theory and experimental tests*. This incubation signature suggests that NT-S is the parent stream of G-S *theory and experimental tests*. There were no clear incubation signatures for the other G-S topics when we tested them against NT-S. None of the G-S topics has clear incubation signatures when tested against B-S. We also looked for incubation signatures for the G-S topics in the numbers of authors who debut in different years, for those who published exclusively in

G-S, NT-S/B-S, as well as those who published in both. Using this second indicator, we see clear incubation signatures in NT-S for all four G-S topics but no incubation signatures in B-S, except for a weak incubation signature for (1) *G-S supercapacitors and electrocatalysts*, which is closely related to batteries. This tells us that the main parent stream of all four G-S topics is NT-S, while (1) *G-S supercapacitors and electrocatalysts* may have B-S as a secondary parent stream. This is our answer to the second scientific question.

Finally, we examined the sudden surge in debutant specialists in NT-S and B-S in 2006. Based on the timing of these surges, we believe they were the results of postemergence interactions with G-S. Using the proportion of G-S references in the NT-S and B-S debutant papers as an indicator, we found no explanation for these surges. We then used a second indicator, where we separated G-S references into eight time periods and plotted the average number of citations to these by NT-S/B-S debutant papers. We saw from these plots that the earlier NT-S debutant papers preferentially cited G-S references from 1992–1994. In contrast, later NT-S debutant papers preferentially cited G-S references from 2004 onwards. Furthermore, B-S debutant papers rarely cited G-S references before 2004, but those published after 2010 cited many later G-S references. This suggests that the G-S breakthrough of 2004 “overwrote” the community’s memory of the 1992 NT-S/G-S theoretical breakthrough in the NT-S community. The B-S community did not react to the 1992 NT-S/G-S breakthrough because it is not relevant to the research field, but responded strongly to *G-S supercapacitors and electrocatalysts* papers published a few years after 2004. We also tried to answer our original third scientific question to identify pre-emergence interactions that seeded the G-S research field and confirmed that these are weak but reasonable.

The stream-based hypotheses in this paper were stated in general terms but tested only in graphene. These hypotheses are reasonable and supported by evidence from the graphene literature but ought to be tested in other fields. Leaving their expected generality aside, what do these findings mean to researchers working on graphene, or innovators developing technologies based on graphene, or grant agencies funding graphene science and technology? For graphene scientists and innovators, these findings must be combined with equivalent results from the downstream stages of the innovation process (patents and commercial products) to piece together the complete life cycle from pure research to applied research to innovation to commercial product, so that they can decide for themselves whether to continue efforts in the same topics or to move on to other topics. For scientists and innovators in general, the methods we developed for identifying the emergence times, the parent streams, and interactions can also be applied to other research fields to accelerate the innovation processes. We expect bad or ill-conceived scientific ideas to die a natural cause during the incubation phases. Still, many good scientific ideas also fail during these phases because of insufficient resources devoted to them. While many grant agencies fund scientific research in newly emergent fields generously, they do so when these fields are already easily recognizable. In global science and technology competition, it probably makes more sense to fund good scientific ideas in the incubation phases to push them past the threshold of emergence. As mentioned above, countries or governments that could do this would command leading positions in the competition.

ACKNOWLEDGMENTS

We thank Ting Yu for suggesting what keywords to use in our search for publications on graphene.

AUTHOR CONTRIBUTIONS

Ai Linh Nguyen: Conceptualization; Data curation; Methodology; Validation; Visualization; Writing—original draft; Writing—review & editing. Wenyan Liu: Conceptualization;

Methodology; Supervision; Validation; Writing—original draft; Writing—review & editing. Khiam Aik Khor: Conceptualization; Funding acquisition; Writing—review & editing. Andrea Nanetti: Conceptualization; Funding acquisition; Writing—review & editing. Siew Ann Cheong: Conceptualization; Funding acquisition; Methodology; Project administration; Supervision; Visualization; Writing—original draft; Writing—review & editing.

COMPETING INTERESTS

The authors have no competing interests.

FUNDING INFORMATION

This research is supported by the Singapore Ministry of Education Academic Research Fund, under the grant number MOE2017-T2-2-075.

DATA AVAILABILITY

We do not own the publication data from the Web of Science. Anyone with a subscription to the database can easily download the data using “graphene,” “nanotube,” and “battery” as topics.

REFERENCES

- Adams, J., & Light, R. (2014). Mapping interdisciplinary fields: Efficiencies, gaps and redundancies in HIV/AIDS research. *PLOS ONE*, 9(12), e115092. <https://doi.org/10.1371/journal.pone.0115092>, PubMed: 25506703
- Ailem, M., Role, F., & Nadif, M. (2015). Cocustering document-term matrices by direct maximization of graph modularity. *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* (pp. 1807–1810). <https://doi.org/10.1145/2806416.2806639>
- Ajayan, P. M. (1999). Nanotubes from carbon. *Chemical Reviews*, 99(7), 1787–1800. <https://doi.org/10.1021/cr970102g>, PubMed: 11849010
- Aleta, A., Meloni, S., Perra, N., & Moreno, Y. (2019). Explore with caution: Mapping the evolution of scientific interest in physics. *EPJ Data Science*, 8(1), 1–15. <https://doi.org/10.1140/epjds/s13688-019-0205-9>
- Battiston, F., Musciotto, F., Wang, D., Barabási, A.-L., Szell, M., & Sinatra, R. (2019). Taking census of physics. *Nature Reviews Physics*, 1(1), 89–97. <https://doi.org/10.1038/s42254-018-0005-3>
- Bethune, D. S., Johnson, R. D., Salem, J. R., De Vries, M. S., & Yannoni, C. S. (1993). Atoms in carbon cages: The structure and properties of endohedral fullerenes. *Nature*, 366(6451), 123–128. <https://doi.org/10.1038/366123a0>
- Bettencourt, L. M., Kaiser, D. I., & Kaur, J. (2009). Scientific discovery and topological transitions in collaboration networks. *Journal of Informetrics*, 3(3), 210–221. <https://doi.org/10.1016/j.joi.2009.03.001>
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan), 993–1022.
- Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), P10008. <https://doi.org/10.1088/1742-5468/2008/10/P10008>
- Boyack, K. W., & Klavans, R. (2010). Cocitation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology*, 61(12), 2389–2404. <https://doi.org/10.1002/asi.21419>
- Boyack, K. W., & Klavans, R. (2019). Creation and analysis of large-scale bibliometric networks. In *Springer handbook of science and technology indicators* (pp. 187–212). Cham: Springer. https://doi.org/10.1007/978-3-030-02511-3_8
- Boyack, K. W., Klavans, R., Small, H., & Ungar, L. (2014). Characterizing the emergence of two nanotechnology topics using a contemporaneous global micro-model of science. *Journal of Engineering and Technology Management*, 32, 147–159. <https://doi.org/10.1016/j.jengtecman.2013.07.001>
- Boyack, K. W., Newman, D., Duhon, R. J., Klavans, R., Patek, M., ... Börner, K. (2011). Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches. *PLOS ONE*, 6(3), e18029. <https://doi.org/10.1371/journal.pone.0018029>, PubMed: 21437291
- Boyack, K. W., Small, H., & Klavans, R. (2013). Improving the accuracy of cocitation clustering using full text. *Journal of the American Society for Information Science and Technology*, 64(9), 1759–1767. <https://doi.org/10.1002/asi.22896>
- Braam, R. R., Moed, H. F., & Van Raan, A. F. (1991a). Mapping of science by combined cocitation and word analysis. I: Structural aspects. *Journal of the American Society for Information Science*, 42(4), 233–251. [https://doi.org/10.1002/\(SICI\)1097-4571\(199105\)42:4<233::AID-AS11>3.0.CO;2-I](https://doi.org/10.1002/(SICI)1097-4571(199105)42:4<233::AID-AS11>3.0.CO;2-I)
- Braam, R. R., Moed, H. F., & Van Raan, A. F. (1991b). Mapping of science by combined cocitation and word analysis. II: Dynamical aspects. *Journal of the American Society for Information Science*, 42(4), 252–266. [https://doi.org/10.1002/\(SICI\)1097-4571\(199105\)42:4<252::AID-AS12>3.0.CO;2-G](https://doi.org/10.1002/(SICI)1097-4571(199105)42:4<252::AID-AS12>3.0.CO;2-G)
- Callon, M., Courtial, J.-P., & Laville, F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry. *Scientometrics*, 22(1), 155–205. <https://doi.org/10.1007/BF02019280>
- Callon, M., Courtial, J.-P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word

- analysis. *Social Science Information*, 22(2), 191–235. <https://doi.org/10.1177/053901883022002003>
- Celardo, L., & Everett, M. G. (2020). Network text analysis: A two-way classification approach. *International Journal of Information Management*, 51, 102009. <https://doi.org/10.1016/j.ijinfomgt.2019.09.005>
- Cheng, Y., & Church, G. M. (2000). Biclustering of expression data. *Proceedings of the International Conference on Intelligent Systems for Molecular Biology (ISMB)* (pp. 93–103).
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391–407. [https://doi.org/10.1002/\(SICI\)1097-4571\(199009\)41:6<391::AID-ASII>3.0.CO;2-9](https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASII>3.0.CO;2-9)
- Dhillon, I. S. (2001). Coclustering documents and words using bipartite spectral graph partitioning. *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 269–274). <https://doi.org/10.1145/502512.502550>
- Dogan, M., & Pahre, R. (1990). *Creative marginality, innovation at the intersection of social sciences*. Boulder, San Francisco, and Oxford: Westview Press.
- Fanelli, D., & Glänzel, W. (2013). Bibliometric evidence for a hierarchy of the sciences. *PLOS ONE*, 8(6), e66938. <https://doi.org/10.1371/journal.pone.0066938>, PubMed: 23840557
- Feng, L., Zhao, Q., & Zhou, C. (2020). Improving performances of Top-N recommendations with coclustering method. *Expert Systems with Applications*, 143, 113078. <https://doi.org/10.1016/j.eswa.2019.113078>
- Feynman, R. P. (1949). Space-time approach to quantum electrodynamics. *Physical Review*, 76(6), 769. <https://doi.org/10.1103/PhysRev.76.769>
- Feynman, R. P. (1953). Atomic theory of liquid helium near absolute zero. *Physical Review*, 91(6), 1301. <https://doi.org/10.1103/PhysRev.91.1301>
- Feynman, R. P. (1958). Excitations in liquid helium. *Physica*, 24, S18–S26. [https://doi.org/10.1016/S0031-8914\(58\)80495-4](https://doi.org/10.1016/S0031-8914(58)80495-4)
- Feynman, R. P., & Speisman, G. (1954). Proton-neutron mass difference. *Physical Review*, 94(2), 500. <https://doi.org/10.1103/PhysRev.94.500>
- Foster, J. G., Rzhetsky, A., & Evans, J. A. (2015). Tradition and innovation in scientists' research strategies. *American Sociological Review*, 80(5), 875–908. <https://doi.org/10.1177/0003122415601618>
- Friedl, J. E. (2006). *Mastering regular expressions*. O'Reilly Media, Inc.
- Geim, A. K. (2009). Graphene: Status and prospects. *Science*, 324(5934), 1530–1534. <https://doi.org/10.1126/science.1158877>, PubMed: 19541989
- George, T., & Merugu, S. (2005). A scalable collaborative filtering framework based on coclustering. *Fifth IEEE International Conference on Data Mining (ICDM'05)* (pp. 625–628). <https://doi.org/10.1109/ICDM.2005.14>
- Gieryn, T. F. (1978). Problem retention and problem change in science. *Sociological Inquiry*, 48(3–4), 96–115. <https://doi.org/10.1111/j.1475-682X.1978.tb00820.x>
- Gipp, B., & Beel, J. (2009). Citation proximity analysis (CPA): A new approach for identifying related work based on cocitation analysis. *ISSI'09: 12th International Conference on Scientometrics and Informetrics* (pp. 571–575).
- Glänzel, W., & Czerwon, H. (1996). A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level. *Scientometrics*, 37(2), 195–221. <https://doi.org/10.1007/BF02093621>
- Glänzel, W., & Thijs, B. (2011). Using “core documents” for the representation of clusters and topics. *Scientometrics*, 88(1), 297–309. <https://doi.org/10.1007/s11192-011-0347-4>
- Herrera, M., Roberts, D. C., & Gulbahce, N. (2010). Mapping the evolution of scientific fields. *PLOS ONE*, 5(5), e10355. <https://doi.org/10.1371/journal.pone.0010355>, PubMed: 20463949
- Hofmann, T. (1999). Probabilistic latent semantic indexing. *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 50–57). <https://doi.org/10.1145/312624.312649>
- Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1–2), 177–196. <https://doi.org/10.1023/A:1007617005950>
- Hoonlor, A., Szymanski, B. K., & Zaki, M. J. (2013). Trends in computer science research. *Communications of the ACM*, 56(10), 74–83. <https://doi.org/10.1145/2500892>
- Hummers, W. S., Jr., & Offeman, R. E. (1958). Preparation of graphitic oxide. *Journal of the American Chemical Society*, 80(6), 1339–1339. <https://doi.org/10.1021/ja01539a017>
- Iijima, S. (1991). Helical microtubules of graphitic carbon. *Nature*, 354(6348), 56–58. <https://doi.org/10.1038/354056a0>
- Janssens, F., Zhang, L., De Moor, B., & Glänzel, W. (2009). Hybrid clustering for validation and improvement of subject-classification schemes. *Information Processing & Management*, 45(6), 683–702. <https://doi.org/10.1016/j.ipm.2009.06.003>
- Jia, T., Wang, D., & Szymanski, B. K. (2017). Quantifying patterns of research-interest evolution. *Nature Human Behaviour*, 1(4), 0078. <https://doi.org/10.1038/s41562-017-0078>
- Kelly, B. T. (1981). *Physics of graphite*. London: Applied Science Publishers.
- Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, 14(1), 10–25. <https://doi.org/10.1002/asi.5090140103>
- Kline, S. J., & Rosenberg, N. (1986). *An overview of innovation. The positive sum strategy: Harnessing technology for economic growth*. National Academy of Science, USA.
- Kroto, H. W., Heath, J. R., O'Brien, S. C., Curl, R. F., & Smalley, R. E. (1985). C60: Buckminsterfullerene. *Nature*, 318(6042), 162–163. <https://doi.org/10.1038/318162a0>
- Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press.
- Kuhn, T. S. (1977). *The essential tension: Selected studies in scientific tradition and change*. Chicago: University of Chicago Press. <https://doi.org/10.7208/chicago/9780226217239.001.0001>
- Liu, W., Nanetti, A., & Cheong, S. A. (2017). Knowledge evolution in physics research: An analysis of bibliographic coupling networks. *PLOS ONE*, 12(9), e0184821. <https://doi.org/10.1371/journal.pone.0184821>, PubMed: 28922427
- Liu, X., Yu, S., Janssens, F., Glänzel, W., Moreau, Y., & De Moor, B. (2010). Weighted hybrid clustering by combining text mining and bibliometrics on a large-scale journal database. *Journal of the American Society for Information Science and Technology*, 61(6), 1105–1119. <https://doi.org/10.1002/asi.21312>
- Madeira, S. C., & Oliveira, A. L. (2004). Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(1), 24–45. <https://doi.org/10.1109/TCBB.2004.2>, PubMed: 17048406
- Mansfield, E. (1991). Academic research and industrial innovation. *Research Policy*, 20(1), 1–12. [https://doi.org/10.1016/0048-7333\(91\)90080-A](https://doi.org/10.1016/0048-7333(91)90080-A)
- Mulkay, M. J. (1975). Three models of scientific development. *The Sociological Review*, 23(3), 509–526. <https://doi.org/10.1111/j.1467-954X.1975.tb02231.x>

- Narin, F., Hamilton, K. S., & Olivastro, D. (1995). Linkage between agency-supported research and patented industrial technology. *Research Evaluation*, 5(3), 183–187. <https://doi.org/10.1093/rev/5.3.183>
- Narin, F., Hamilton, K. S., & Olivastro, D. (1997). The increasing linkage between US technology and public science. *Research Policy*, 26(3), 317–330. [https://doi.org/10.1016/S0048-7333\(97\)00013-9](https://doi.org/10.1016/S0048-7333(97)00013-9)
- Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69(2), 026113. <https://doi.org/10.1103/PhysRevE.69.026113>, PubMed: 14995526
- Nguyen, A. L., Liu, W., Khor, K. A., Nanetti, A., & Cheong, S. A. (2020). The golden eras of graphene science and technology: Bibliographic evidences from journal and patent publications. *Journal of Informetrics*, 14(4), 101067. <https://doi.org/10.1016/j.joi.2020.101067>
- Novoselov, K. S., Geim, A. K., Morozov, S. V., Jiang, D., Zhang, Y., ... Firsov, A. A. (2004). Electric field effect in atomically thin carbon films. *Science*, 306(5696), 666–669. <https://doi.org/10.1126/science.1102896>, PubMed: 15499015
- Osborne, F., Scavo, G., & Motta, E. (2014). A hybrid semantic approach to building dynamic maps of research communities. *International Conference on Knowledge Engineering and Knowledge Management* (pp. 356–372). https://doi.org/10.1007/978-3-319-13704-9_28
- Plimpton, S. (1995). Fast parallel algorithms for short-range molecular dynamics. *Journal of Computational Physics*, 117(1), 1–19. <https://doi.org/10.1006/jcph.1995.1039>
- Popper, K. R. (1959). *The logic of scientific discovery*. London: Routledge.
- Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 40, 211–218. <https://doi.org/10.1108/00330330610681286>
- Price, D. J. d. S. (1986). *Little science, big science*. New York, NY: Columbia University Press.
- Reif, F. (1961). The competitive world of the pure scientist. *Science*, 134(3494), 1957–1962. <https://doi.org/10.1126/science.134.3494.1957>, PubMed: 17744407
- Rosenberg, N., & Birdzell, L. E. (1990). Science, technology and the Western miracle. *Scientific American*, 263(5), 42–55. <https://doi.org/10.1038/scientificamerican1190-42>
- Saito, R., Fujita, M., Dresselhaus, G., & Dresselhaus, M. S. (1992a). Electronic structure of chiral graphene tubules. *Applied Physics Letters*, 60(18), 2204–2206. <https://doi.org/10.1063/1.107080>
- Saito, R., Fujita, M., Dresselhaus, G., & Dresselhaus, M. S. (1992b). Electronic structure of graphene tubules based on C60. *Physical Review B*, 46(3), 1804. <https://doi.org/10.1103/PhysRevB.46.1804>, PubMed: 10003828
- Salatino, A. A., Osborne, F., & Motta, E. (2017). How are topics born? Understanding the research dynamics preceding the emergence of new areas. *PeerJ Computer Science*, 3, e119. <https://doi.org/10.7717/peerj-cs.119>
- Sinatra, R., Wang, D., Deville, P., Song, C., & Barabási, A.-L. (2016). Quantifying the evolution of individual scientific impact. *Science*, 354(6312), aaf5239. <https://doi.org/10.1126/science.aaf5239>, PubMed: 27811240
- Small, H. (1973). Cocitation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24(4), 265–269. <https://doi.org/10.1002/asi.4630240406>
- Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1), 11–21. <https://doi.org/10.1108/eb026526>
- Stuart, S. J., Tutein, A. B., & Harrison, J. A. (2000). A reactive potential for hydrocarbons with intermolecular interactions. *Journal of Chemical Physics*, 112(14), 6472–6486. <https://doi.org/10.1063/1.481208>
- Sun, X., Kaur, J., Milojević, S., Flammini, A., & Menczer, F. (2013). Social dynamics of science. *Scientific Reports*, 3, 1069. <https://doi.org/10.1038/srep01069>, PubMed: 23323212
- Turney, J. (1991). What drives the engines of innovation? *New Scientist*, 132, 35–40.
- Van Mechelen, I., Bock, H.-H., & De Boeck, P. (2004). Two-mode clustering methods: A structured overview. *Statistical Methods in Medical Research*, 13(5), 363–394. <https://doi.org/10.1191/0962280204sm373ra>, PubMed: 15516031
- Web of Science. (n.d.). <https://www.webofknowledge.com>
- Xie, J., Ma, A., Zhang, Y., Liu, B., Cao, S., ... Ma, Q. (2020). QUBIC2: A novel and robust biclustering algorithm for analyses and interpretation of large-scale RNA-Seq data. *Bioinformatics*, 36(4), 1143–1149. <https://doi.org/10.1093/bioinformatics/btz692>, PubMed: 31503285
- Yu, D., Wang, W., Zhang, S., Zhang, W., & Liu, R. (2017). Hybrid self-optimized clustering model based on citation links and textual features to detect research topics. *PLOS ONE*, 12(10), e0187164. <https://doi.org/10.1371/journal.pone.0187164>, PubMed: 29077747
- Zeng, A., Shen, Z., Zhou, J., Fan, Y., Di, Z., ... Havlin, S. (2019). Increasing trend of scientists to switch between topics. *Nature Communications*, 10(1), 1–11. <https://doi.org/10.1038/s41467-019-11401-8>, PubMed: 31366884