Breakthrough Recognition: Competing for Attention

Sen Chai *
ESSEC Business School
Cergy-Pontoise, France.
chai@essec.edu
*corresponding author

Anoop Menon
The Wharton School
University of Pennsylvania,
Philadelphia, PA.
armenon@wharton.upenn.edu

Acknowledgments: We would like to thank the Mack Institute for Innovation Management at the Wharton School, University of Pennsylvania for supporting our work. We are also grateful to Michael Bikard, Richard Freeman, Matt Marx, Elisa Operti, Julianne Smith, Haris Tabakovic, and Sifan Zhou and for their insightful comments and feedback. All errors remain our own.

Abstract

We introduce to the literature on the recognition and spread of ideas the perspective that articles compete for the attention of researchers who might build upon them, in addition to the "bias against novelty" view documented in prior research. We investigate these effects by analyzing more than 5.3 million research publications from 1970 to 1999 in the life sciences. In keeping with the "competition for attention" view, we show that articles covering rarely addressed topics tend to receive more citations and have a higher chance of being a breakthrough paper. We also explore some conditions under which these effects might vary, by using decade subsamples, home versus foreign field forward citations, as well as short-, medium- and long-term time windows. Finally, we also find evidence consistent with the previously documented channel of "bias against novelty", as well as that both mechanisms can work simultaneously.

Introduction

The idea at the core of a successful scientific breakthrough, one that lays the foundations for further advancements, must not only be significantly novel but must also be recognized by others who then build on it (Simonton, 1999). Studies exploring the former criterion have indicated that depth of knowledge is a source of breakthrough creation, and is especially key in identifying anomalies that may open new paths and provide building blocks upon which new ideas are constructed (Gardner, 1993). However, novel recombinations of ideas from distant and diverse sources (Hargadon & Sutton, 1997) are crucial for path-breaking advances in order to break away from prevailing theories (Simonton, 1999) and avoid myopia. Thus, novel atypical combinations are key to the creation of breakthrough ideas although they also encompass many conventional combinations (Uzzi, Mukherjee, Stringer, & Jones, 2013).

However, as per its definition, breakthroughs depend not only on the quality of the idea itself, but also on whether they are recognized and built upon by the community. Hence, even if the idea is of high quality, there is no guarantee that it will be recognized. The consumption and recognition of knowledge is usually facilitated when ideas produced fit within the confines of the accepted paradigm (Kuhn, 1962; Margolis, 1993). Given that breakthrough ideas are inherently novel, their acceptance and recognition may be difficult and not immediate. Indeed, existing work mainly documents evidence for a "bias against novelty" mechanism, where idea recognition is usually facilitated when similar or "closer" ideas are recombined together (Fleming, Mingo, & Chen, 2007; Mueller, Melwani, & Goncalo, 2012; Rietzschel, Nijstad, & Stroebe, 2010; Wang, Veugelers, & Stephan, 2017). However, to the best of our knowledge, this is the only mechanism pertaining to the recognition of ideas that has received deep treatment. In this paper, we contribute to the innovation literature on breakthroughs by proposing, exploring, and finding empirical evidence for a new mechanism for the recognition and spread of

breakthrough ideas. This is the "competition for attention" view – the notion that ideas can be thought of as competing for the attention of researchers in various domains. The concept of competition for attention is borrowed from work in the organizational and marketing literatures that have pointed out the importance of people, products, and ideas having to compete for the attention of the audience they are trying to address due to the informational overload faced by decision makers (Hannan & Carroll, 1992; Hansen & Haas, 2001; Iyer & Katona, 2016; Ocasio, 2011; Shen, Hu, & Rees Ulmer, 2015).

To see this better in our context of ideas competing for attention, consider the observation that multiple instances of the same discovery that were made quasi-simultaneously (Bikard, 2018) have had different recognition patterns. For example, the trigger mechanism to RNA interference of gene silencing was discovered in the same year by two separate groups – one using c.elegans worms (Fire, Xu, Montgomery, Kostas, Driver, & Mello, 1998) and one using potatoes (Waterhouse, Graham, & Wang, 1998) as the model organism. While the latter discovery was made in a very populated area as plant researchers had been working on understanding the phenomenon since the early 1990s, it was a relatively unpopulated area for animal researchers. In turn, the paper on c.elegans worms caught on much more than the paper using potatoes, which got lost among the crowd.

More generally, ideas that emerge in topical areas that are "crowded" will have to compete harder with other ideas for the attention of the scientific community, thereby lowering their chances of being recognized, while those that emerge in relatively "lonely" topical areas will not have to face such stiff competition for attention. Researchers, like everyone else, have limited time, attention, and cognitive resources to expend in the search over the body of scientific knowledge contained in publications (Bordalo, Gennaioli, & Shleifer, 2016; Hansen & Haas, 2001; Iyer & Katona, 2016; Shen et al., 2015). They cannot be aware of every paper on a

given topic. Instead, in an abstract sense, all the papers that address a certain topic can be thought of as competing for the attention of the researcher when she is exploring that domain and looking to build on that topic. Thus, if a researcher is exploring a certain topic, an article addressing that topic is more likely to be "picked up" when there are fewer other articles addressing the topic at that point in time. It accounts for the pattern that papers covering rarely addressed topics tend to receive more citations and tend to have a higher chance of being a breakthrough paper, i.e., belonging to a certain top percentage of cited papers in its field in a given year.

We investigate this hypothesis by analyzing more than 5.3 million research publications spanning three decades (1970 to 1999) in the life sciences from the Web of Science and MedLine (Torvik & Smalheiser, 2009). In keeping with the "competition for attention" view, our data indicate that papers that cover topics that are frequently covered by other papers published in that year tend to receive many fewer citations compared to papers that cover much less frequently addressed topics. They also have a much lower chance of being in the top 1% of forward citations for papers published in that field and year. In other words, it is the papers that address rarely covered topics that have a much higher chance of being recognized and built upon by the scientific community. Our analysis also explored some conditions that moderate the effect of the mechanism of "competition for attention", using decade subsamples, home versus foreign field forward citations as well as short-, mid- and long-term time windows to run additional analysis on top of the main effects.

Interestingly, our data also corroborates the previous findings on the "bias against novelty", and both mechanisms could be thought of as operating in tandem. Hence, we also explore implications of the interaction of these two mechanisms. From a policy standpoint, our

findings provide a new dimension for decision-makers when looking to optimize public and private investment in science and technology in search of the next breakthrough idea.

Breakthrough Creation and Recognition

Creation of Breakthrough Ideas

Since Schumpeter's notion of creative destruction (1942), scientific and technological breakthroughs have held a central and recurrent prominence in the innovation literatures. The topic remains extremely important, as breakthroughs have the potential to disrupt extant industries and regions, and at the same time, provide renewed impetus for new industries, whole economies, and society. For breakthroughs to occur, they depend not only on the quality of the idea itself, but also on whether they are recognized and built upon by the community (Simonton, 1999). Though both components are crucial for such breakthroughs, much of the research has concentrated on the former, that of identifying the sources of breakthrough idea generation.

The process underlying the generation of novel, path-breaking ideas that has received significant attention is the recombination of ideas (Fleming, 2001; Henderson & Clark, 1990; Weitzman, 1998). The combination of diverse knowledge stemming from different individuals, different experiences, or different network positions driving breakthrough discovery pre-supposes combinatorial novelty in the creative search process. This notion of combinatorial novelty is prevalent in the innovation literature, and it has been directly employed and studied in settings pertaining to both scientific and technological innovation, using peer reviewed publications (Azoulay, Güler, Koçak, Murciano-Goroff, & Anttila-Hughes, 2012; Boudreau, Guinan, Lakhani, & Riedl, 2016; Uzzi et al., 2013) and patents (Fleming, 2001; Verhoeven, Bakker, & Veugelers, 2016).

For instance, collaboration is likely to improve search diversity and idea selection efficiency as the circulation of ideas for critique by collaborators decreases the likelihood of poor

outcomes, while multiple collaborators permit the recombination of more diverse ideas (Singh & Fleming, 2010; Wuchty, Jones, & Uzzi, 2007). Generalists can bring together disparate and distant components. Similarly, mobility across multiple affiliations increases exposure to a greater set of ideas, which increases the creation of potential breakthroughs (McEvily & Zaheer, 1999). From a network perspective, social brokers—group members who bring together disparate views—have shown to be more creative as brokers occupy a nexus position in which they have first access to diverse information that can be recombined (Burt, 2004). Hence, most of the work that explore sources of breakthrough creation rests on this notion of recombination of diverse ideas sources.

Recognition of Breakthrough Ideas

Bias Against Novelty

This aforementioned diversity is key in avoiding myopia and breaking away from prevailing theories and assumptions on the path to creating breakthrough advances (Chai, 2017; Hargadon & Sutton, 1997; Kuhn, 1962; Lifshitz-Assaf, 2017; Simonton, 1999). Although, the recombination of more distant ideas may increase the likelihood of producing path-breaking novelty, these ideas still need to be recognized for them to be considered breakthroughs. One cannot simply assume that the more ground breaking an idea is, the more impact it will have, and/or the more likely it is to be recognized and built upon by the scientific community.

In fact, the literature has often documented that the consumption and recognition of knowledge is usually facilitated when similar or "closer" ideas are recombined together (Fleming et al., 2007; Wang et al., 2017) as these ideas are more likely to fit within the confines of the accepted paradigm (Kuhn, 1962; Margolis, 1993). When new breakthrough ideas emerge, paradigmatic forces may lead to the resistance of these new competing paradigms. The old paradigm is incommensurable with the new one (Kuhn 1962), and proponents of each will talk

past one another, avoiding meaningful conversations between the conflicting schools of thought. It is precisely because of these more distant and less familiar recombinations that the recognition of these ground breaking novelties may be hindered or delayed (Fleming et al., 2007; Wang et al., 2017) as these ideas also lack legitimacy due to their newness (Hannan & Carroll, 1992). Hence, there is a bias against novelty in recognizing breakthrough ideas.

The difficulty in accepting novelty has also been documented from a cognitive perspective. Novelty or creativity is often associated with the aversive state of uncertainty (Mueller et al., 2012). Hence, although individuals openly assert that creativity is the fundamental driving force for scientific and technology advancement (Hennessey & Amabile, 2010), they still tend to routinely reject creative ideas and instead select more feasible and desirable ideas at the cost of originality (Rietzschel et al., 2010). This tension has also been documented in the context of cognitive distance between alliance partners and their subsequent innovation performance. Novelty increases, but the consumption or absorption of this novelty decreases with cognitive distance (Nooteboom, Van Haverbeke, Duysters, Gilsing, & Van den Oord, 2007).

Uzzi et al. (2013) offer a partial resolution to this tension by proposing that novel atypical combinations within an article are key to the creation of breakthrough ideas, although the breakthrough papers also tend to include many conventional combinations, thus "straddling" the two effects. Similarly, Foster, Rzhetsky, and Evans (2015) also speak to this essential tension between productive tradition and risky innovation, and find that for prizewinners in biomedicine and chemistry, occasional gambles for extraordinary impact account for the observed levels of risky breakthrough innovation across a researcher's portfolio of publications.

Competition for Attention

While the bias against novelty/creativity view has been very insightful, we propose that there is more than one mechanism driving the acceptance and recognition of groundbreaking ideas. A recent stream of work in the organizational and marketing literatures has pointed out the importance of people, products, and ideas having to compete for the attention of the audience they are trying to address (Hansen & Haas, 2001; Iyer & Katona, 2016; Ocasio, 2011). Given the informational overload that is faced by customers and/or decision makers, agents and firms can develop strategic behaviors to "compete for the attention" of their target audiences, trying to make sure that their own products or ideas are attended to (Iyer & Katona, 2016; Shen et al., 2015).

At the organizational level, this focus on attention and attentional mechanisms has a long intellectual history. In their seminal work, Simon and Barnard (1947) argued for the how organizations could be thought of as structures that focus and channel the attention of individuals within those organizations. Several others have followed in this tradition (Cyert & March, 1963; Daft & Weick, 1984; March & Shapira, 1987; March & Simon, 1958) and explored how attention and its differential distribution in an organization can significantly affect organizational decision making. More recently, an "Attention Based View" of the firm has been proposed (Ocasio, 1997) that includes not just how rules and structures within an organizational affect the distribution of the scarce resource of attention within the organization and affect decision making, but also how these attentional structures play a large role in the evolution of the organization over time. Other work has noted how critical events can focus the attention of a community or field, and how that can affect institutional change (Hoffman & Ocasio, 2001; Nigam & Ocasio, 2010). (Please see (Ocasio, 2011) for a good review of the various traditions within organizational studies that have explored the importance of attention and attention allocation.)

Another tradition within this domain that has explored the role of attention in organizational processes, and one whose perspective comes closest to the views we express here, is the "ecological perspective". In population ecology, the "density dependence" argument explores the interplay between competition and legitimacy (Hannan & Carroll, 1992). When new domains emerge, organizations operating in it have limited legitimacy. However, when the domain is less crowded, there is less competition between organizations. Translating this onto the idea space entails higher bias against novelty due to lack of legitimacy, but because the domain is not yet crowded there is limited competition for attention. Conversely, competition between and legitimacy of organizations both increase in more crowded domains. In our context, this implies more competition for attention but lower bias against novelty.

The competition for attention aspect of this argument has been demonstrated in a few studies. In marketing, faced with market competition, firms have to strategically enhance certain product features in order to draw consumer attention (Bordalo et al., 2016). Recent digitization trends associated with the expansion and explosion of available electronic data has made attention rather than information the scarcer resource. Therefore, information senders have to compete for the attention of targeted receivers. Hence, when faced with an abundance of information, research has found that the less information offered translates into more usage (Hansen & Haas, 2001).

We borrow from these literatures and posit that a similar "competition for attention" mechanism is at play in the domain of scientific publications as well. Ideas embedded in publications are constantly competing for the limited attention of the other researchers in the scientific community. Those papers that contain ideas that are commonly studied will be competing for the attention of the community against all the other papers studying those ideas, while those papers that contain rarely studied ideas will not have to engage in such a tough fight

for attention. Thus, we posit that ideas that are less commonly used by the broader scientific community will garner more recognition as they need to compete less for attention.

Hypothesis 1: Ceteris paribus, publication recognition decreases when the topics contained in it are more commonly studied, due to increased competition for attention that it will have to face.

Similarly, along the same line of reasoning, the more topics a publication spans, the greater its odds are of being widely recognized. When a researcher explores and expands on existing knowledge, she usually searches using keywords. But given that the attention of the researcher is limited, if a publication covers a broader set of topics, it is more likely to be found and picked up, and thus gain more attention.

Hypothesis 2: Ceteris paribus, publication recognition increases when the publication spans more topics.

Competition for Attention & Bias against Novelty

Having laid out the competition for attention mechanism, we now consider what one might expect in terms of observable patterns in idea recognition once both these mechanisms (bias against novelty and competition for attention) are taken into account simultaneously. To recap, the bias against novelty mechanism is hypothesized to affect the adoption of ideas by skewing it away from those that are highly novel, usually captured by rare or "distant" combinations of topics. One the other hand, the competition for attention mechanism is hypothesized to affect the adoption of ideas by skewing it away from commonly studied topics, captured by the frequency with which other papers are covering the same topic(s), and also helped along by covering more topics (Bordalo et al., 2016; Hansen & Haas, 2001; Iyer & Katona, 2016; Shen et al., 2015). Thus, while the latter mechanism is concerned with whether and how much the scientific community becomes aware of a publication, the former is about adoption based on the degree of fit of the idea with existing paradigms.

Given these separate channels of action, one can reasonably expect both these mechanisms to be at work simultaneously. When might they reinforce each other, and when might they cancel out? Before we get to this interaction, we first hypothesize the expected pattern that would result from the bias against novelty mechanism (Fleming et al., 2007; Mueller et al., 2012; Nooteboom et al., 2007; Wang et al., 2017).

Hypothesis 3: Ceteris paribus, publication recognition decreases when the publication recombines more distant topics (i.e., topics that are rarely combined).

Moving on to the joint action of both mechanisms, the innovation literature is rather mute given that one of the mechanisms, that of competition for attention introduced above, was borrowed from other literatures. However, logically deducing from the arguments for each mechanism, one should expect that ideas with commonly covered topics will have to compete hard for attention (competition for attention), while common combinations of ideas will have an easier time being adopted as the recombined idea has already gained traction and legitimacy (bias against novelty (in reverse)). In fact, that would be what one would predict if one were to translate the notions of the density dependence argument (Hannan and Carroll (1992), presented earlier) from the organizational context into our context of innovation and the spread of ideas.

Note that one of the channels is about the topics in a publication considered separately (how commonly studied they are, how many there are), while the other is about pairs or combinations of ideas (how common is a given combination). With that in mind, one should expect the greatest rate of adoption for those ideas that cover topics that are less commonly addressed by themselves as they benefit from reduced competition for attention, but at the same time are more commonly seen in combination, thus without being negatively biased by novelty[1].

[1] For example, continuing with the case of RNA interference, on top of writing in a less crowded space the paper on c.elegans worms also combined two keywords, dsRNA and gene expression regulation, that had been commonly combined in both the communities studying gene expression and interferon responses. The paper on potatoes, on the other hand, combined sense and antisense RNA with gene expression regulation, which was much less common.

On the other hand, those ideas that cover topics that are commonly addressed by other publications (thus suffering from increased competition for attention), but are rarely paired/combined together due to the novelty of the idea (thus also suffering from increased bias against novelty) will have the hardest time being adopted.

Hypothesis 4: Ceteris paribus, publication recognition decreases when the topics covered in a publication are commonly studied, but are rarely studied together, and vice versa.

A similar prediction should also be expected when using the span of topics in a publication as a measure of the competition for attention (as described earlier). Thus:

Hypothesis 5: Ceteris paribus, publication recognition decreases when a publication covers fewer topics, and when those topics are rarely studied together, and vice versa.

Data and Methods

We analyzed 5.3 million scientific publications in the life sciences published between 1970 and 1999. We focused on the life sciences in order to take advantage of the independent indexing of each publication using Medical Subject Heading (MeSH) keywords performed by indexers of the United States National Library of Medicine (NLM) at the National Institutes of Health (NIH). Since these controlled keywords are not assigned by the authors but by an independent indexer, it is believed to be a relatively objective classification scheme. It is therefore more difficult for authors to game the assignment of keywords by choosing keywords that are more or less popular at the time of publication or by introducing more or less keywords to manipulate the apparent breadth and depth of the article. Moreover, new MeSH keywords are only introduced into the lexicon once the phenomenon or concept has gained enough traction. For instance, the mechanism of RNA interference and its trigger (a breakthrough discovery in molecular biology that was awarded the 2006 Nobel Prize in Physiology or Medicine) and was discovered in 1998 but RNA interference only entered the MeSH lexicon in 2002. Therefore,

although new MeSH keywords represent novel ideas or concepts, they are only added onto the MeSH lexicon after a new and steady stream of work has already been recognized. Hence, their frequency of use immediately after they are introduced is already high, and we do not run the risk of overestimating the novelty of those of the follower papers that are assigned the new MeSH keywords.

We chose 1999 as the ending year of our sample to ensure that a total of 15 years has elapsed for every publication, for the accurate collection of the longest specification of our dependent variables based on forward citations. Our data draws mainly from the disambiguated MedLine database[2] (Torvik & Smalheiser, 2009), as it provides MeSH keywords as well as individual disambiguated author information. We supplemented this database using the Thomson Reuters Web of Science (WOS) database by matching on the unique publication identification number (PMID) assigned on the MedLine database in order to obtain broad and narrow subfield classification for each article according to subject area of the journal it is published in. Our sample contains 178 unique narrow subfields and 16 broad subfields in the life sciences.

The main data is at the publication level. Table 1 shows the descriptive statistics, including sample size, mean, standard deviation, and minimum and maximum of each variable used in our analyses. Table A1 in the appendix shows the correlation matrix.

[Insert Table 1 about here]

Dependent variables

The consumption and recognition of scientific publications was proxied for by the widely-used measure of forward citations. We measured both the total number of citations that a

[2] MedLine is a database of references and abstracts on life sciences and biomedical topics maintained by the NLM.

paper garnered, as well as whether or not the paper belonged to the top 1% of papers in terms of forward citations for all the papers published in that year and in that particular subfield.

Citations. We tabulated the number of forward citations that each publication garnered within 15 years, 10 years, and 3 years after its publication *(# of cites 15yr, 10yr, and 3yr)*. We chose these three time windows to capture the temporal patterns of forward citations, as prior research has shown that the recognition of ideas takes time and is associated with a non-linear accumulation pattern.

Aside from the three time window variants, we also separated citations into those from the same narrow subfield, i.e. home field citations with 3-year, 10-year, and 15-year forward citations windows (*# home cites 3yr, 10yr and 15yr*), and those from outside the narrow subfield, i.e. foreign field citations with 3-year, 10-year, and 15-year forward citations windows (*# foreign cites 3yr, 10yr*, and *15yr*).

We also included the standard deviation of the number of forward citations in the 15-year window in order to plot dispersion graphs of the variable.

Top 1% of the citation. To capture the right tail of the citation distribution, i.e. breakthrough papers, we use a binary variable indicating whether the number of total forward citations received by a publication is within the top 1% of the distribution of publications in terms of their total forward citations (with 3-year, 10-year, and 15-year time windows), clustered by year of publication and narrow subfield.

Independent variables

The "competition for attention" view was first proxied for by a measure of frequency using the number of times each MeSH keyword was assigned in the entire population of life sciences articles on an annual basis. In the three decades that our data covers, the number of distinct MeSHs per year ranged from 7,820 to 16,786. For each publication, the average

frequency with which MeSHs of the focal publication were used to index other articles provides us an indication of how much attention the topics covered were receiving during the year it was published. As mentioned earlier, we also used the number of MeSH keywords assigned to each publication as another measure for competition for attention. Finally, to replicate the bias against novelty view, we also created a measure of combinatorial novelty.

Competition for Attention – average MeSH keyword frequency. For each MeSH keyword in the focal publication, we first tabulated the frequency with which it was used by all articles published in the same year (total count). Then for each publication, we took the frequency of each of the MeSH keywords in that publication and calculated their average. The higher the average frequency, the more the topics covered in the focal publication were receiving intense study by the scientific community. Since the average frequency values tend to be quite large as shown in Table 1, the regressions were run with the normalized value of average frequency divided by 10,000.

Competition for Attention – number of MeSH keywords. We counted the number of MeSH keywords indexed for each focal publication. The higher the number of MeSH keywords used to index a publication, the more topics are covered in the focal publication, and the greater its odds are of being widely recognized

Combinatorial Novelty – average/median/maximum MeSH keyword dyad distance. For each dyadic combination of MeSH keywords in the focal publication, we first tabulated the frequency with which the same dyad appeared in all articles published in the same year (total count). The distance between each dyad is the reciprocal of the dyad frequency. Then for each publication, we took the distance of each of the MeSH dyads in that publication and calculated their average. The higher this measure the more distant are the ideas being recombined, as less

frequent MeSH co-occurrences indicate rarer recombination from farther idea spaces. We also calculated the median and the maximum dyad distance for each publication.

Control variables

Given prior research we also included the following control variables.

Average cumulative publications. In order to address the possible concern that perhaps highly skilled, more published authors were more likely to work in low frequency topics (though unlikely), or that they would tend to use more MeSH keywords, we also ran our analyses controlling for the average number of prior publications of the author team[3].For each author on the focal publication, we tabulated the number of articles published by that author cumulated from first publication to the year of the focal publication. We then take the average of the cumulative publications for all authors of the focal publication, to control for the publishing experience of authors.

Number of authors. We controlled for the number of co-authors as studies have shown an increasing trend of teamwork (Wuchty et al., 2007) and a positive correlation with citations (Singh & Fleming, 2010). We counted the number of authors for each focal publication.

Impact factor. We also controlled for the impact factor of the journal in which the article was published as articles published in higher impact journals are more likely to be viewed and, hence, picked up. We calculated the impact factor using the Thomson Reuters method defined as citations from research articles to the journal in the current year to items published in the previous two years, divided by the total number of scholarly items (these comprise articles, reviews, and proceedings papers) published in the journal in the previous two years.

[3] We did find, however, that more prior publications are indeed strongly correlated with the focal paper receiving more citations.

Additionally, in the full regression specifications, we also included subfield fixed effects and *publication year* fixed effects.

Model Estimation

The findings that we observed are first represented graphically. We also performed further analyses using regression models to control for potentially confounding variables and to explore the variation of the effects across different contingencies. We used two regression models for our tests, along with a few robustness checks mentioned below.

The first model we employed was a negative binomial regression with *# of cites* as the dependent variable since it is an over-dispersed count variable, and the different independent variables and controls mentioned above. We chose negative binomial models instead of simple Poisson models in order to circumvent the assumption of equal mean and variance distribution to minimize estimation bias, as follows:

$$Citations_i = \beta_0 + \beta_1.avg\ frequency_i + \beta_2.\#\ of\ MeSHes_i + \beta_3.avg\ distance_i + \gamma.Controls_i + \delta_i + \tau_i + \varepsilon_i$$

where for publication i, β_1 and β_2 are the coefficients of interest for the competition for attention independent variables, β_3 is the coefficient of interest for the bias against novelty independent variable, δ_i is a dummy variable for the narrow subfield that the publication belongs to – thus incorporating subfield fixed effects, τ_i is its year of publication – thus incorporating publication-year fixed effects, and ε_i is the noise term.

The second model was a logistic regression with *Top 1% of the citation*, an indicator, as the dependent variable, and the different independent variables and controls mentioned above introduced sequentially, and then finally with all of them in the full model:

$$Top\ 1\%\ of\ the\ citation_i = \beta_0 + \beta_1.avg\ frequency_i + \beta_2.\#\ of\ MeSHes_i + \beta_3.avg\ distance_i + \gamma.Controls_i + \tau_i + \varepsilon_i$$

where for publication i, β_1 and β_2 are the coefficients of interest for the competition for attention independent variables, β_3 is the coefficient of interest for the bias against novelty independent variable, τ_i is its year of publication – thus incorporating publication-year fixed effects, and ε_i is the noise term. Note that we do not include subfield fixed effects here since the dependent variable is constructed at the subfield level.

Findings

Competition for Attention

Main effects. We first present our main findings (using the longest forward citation accumulation window of 15 years after publication) graphically. Figure 1 illustrates the relationship between competition for attention as operationalized using average frequency and idea recognition. We find that the number of forward cites and the probability of being in the top 1% (as well as the 95% confidence interval for both) fall in average frequency, as shown respectively in Figures 1A and 1B. Hence, the more a publication covers familiar topics the more it has to compete for attention, the less cites it will receive and the less likely it is to become a breakthrough paper. This supports Hypothesis 1.

[Insert Figure 1 about here]

Our graphical evidence shows raw correlations between the dependent and independent variables and does not account for the control variables. However, our regression results in Tables 2 and 3, which ultimately includes all controls and fixed effects, are reassuringly consistent with the plots. We find in Model 1 of Table 2 that a ten thousand unit increase in average frequency translates to a 9.3%[4] significant decrease on forward cites. If we include all independent, control variables and subfield as well as time fixed effects into the regression as

[4] $e^{coefficient} - 1 = e^{-0.0971} - 1 = -0.0925$

shown in Model 5, we find a 11.8% significant decrease in forward citations for a ten thousand unit increase in average frequency. The effect sizes are similar for a publication to be in the top 1% of forward citations using logistic analysis. Model 1 of Table 3 shows a 10.2%[5] significant decrease in the odds of being in the top 1% of forward citations with a ten thousand unit increase in average frequency, while Model 5 shows a 11.8% significant decrease in the same odds of being in the top 1% of forward citations with a ten thousand unit increase in average frequency when accounting for all independent, control variables and fixed effects. These regression results fully support our graphical evidence in that articles obtain fewer citations on average and are less likely to become a highly cited breakthrough paper when they cover popular topics of current conversation, and thus, Hypothesis 1 is supported.

[Insert Table 2 and 3 about here]

Figure 2 plots the relationship between competition for attention as operationalized using number of MeSHs and idea recognition. Our findings indicate a strong pattern of increasing cites and top 1% probability (and the 95% confidence interval) in the number of MeSHs as shown respectively in Figures 2A and 2B. Thus, the more topics a publication covers, the more likely it can be found when searched, the more cites it will receive, and the more likely it is to become a breakthrough paper. Thus, Hypothesis 2 is supported.

[Insert Figure 2 about here]

Again, to ensure robustness of our findings, we also ran regressions and found full support for the graphical evidence. We find in Model 2 of Table 2 that one additional MeSH keyword indexed on the publication translates to a 7.2% significant increase in forward citations.

[5] $e^{coefficient} - 1 = e^{-0.108} - 1 = -0.1024$

There might be a concern that perhaps the increasing number of citations associated with a larger number of MeSH keywords might be due to an omitted variable – time – wherein over the years, papers have tended to include more MeSH keywords, and the average number of citations has also increased. While this would not address the frequency results, we did rerun our analysis with year fixed effects and the results were unchanged. If we include all independent variables, control variables and fixed effects into the regression as shown in Model 5, we find a 4.4% significant increase in forward citations for one unit increase in MeSH keyword. The effect sizes are even more pronounced for a publication to be in the top 1% of forward citations using logistic analysis. Model 2 of Table 3 shows a 10.6% significant increase in the odds of being in the top 1% of forward citations for one additional MeSH keyword indexed, while Model 5 shows an 9.2% significant increase in the same odds when accounting for all independent, control variables and fixed effects. These regression results fully support our graphical evidence in that articles obtain fewer citations on average and are less likely to become a highly cited breakthrough paper when they cover popular topics of current conversation.

Aside from focusing on mean regressions, we also investigate the dispersion of forward citations. We know from our findings above that the less competition for attention present, the more likely a paper is picked up and cited. However, when articles focus on rarely discussed topics we would expect high potential for gains but also high failure rate since if the topic is too far from mainstream it will be difficult for it to be recognized. This pattern was found in the data as illustrated in Figure 3, which plots the standard deviation (and its 95% confidence interval) of the number of forward citations on quartiles of both measures of competition for attention.[6]

[Insert Figure 3 about here]

[6] We thank one of the referees for pointing us in this direction.

Variation of Effects. Given these results, we wanted to explore trends of how competition for attention may vary based on various factors. First, we explored how competition for attention trended throughout the three-decade span that our data covers. We divided our full sample into three subsamples according to the decade of publication – 1970s, 1980s, and 1990s. Second, we separated the forward citation window into short, medium, and long term windows. Our main results use the long-term window of 15 years, but we also investigate how competition of attention is linked to a short-term forward citation window of 3 years and a mid-term window of 10 years. Finally, we divided total forward citations into home field citations and foreign field citations. Home field citations are forward citations from publications in the same narrow subfield as the focal paper, while foreign field citations are forward citations from publications outside the same narrow subfield. Tables 4 to 6 respectively show all these variations for citation windows of 15, 10, and 3 years.

[Insert Table 4, 5, and 6 about here]

Model 1 in Tables 4, 5 and 6 show the full regression model on the full data sample using total number of cites as the dependent variable. Comparing the effect sizes for average frequency of MeSH keywords in Table 4 for the 15-year window, Table 5 for the 10-year window and finally Table 6 for the 3-year window, we observe stronger effects for the shorter forward citation windows. Specifically, a ten thousand unit increase in average frequency translates to an 11.8% significant decrease on forward cites in the 15-year window, while the effects are stronger in the 10-year and 3-year windows with respectively a 12.7% and 15.5% significant decrease. These findings suggest that shortly after an article has been published, competition for attention as measured by average frequency is stronger since the contributions are still uncertain. As time passes the article's contribution has settled and the effect of competition for attention is weaker.

When measuring competition for attention using the number of MeSH keywords, we observe different but less pronounced trends. The effect sizes are very similar between the 15-year and 10-year window with an additional keyword translating to a 4.4% significant increase in citations, while for the 3-year window the increase in citations is 3.8% and weaker.

In order to observe whether competition for attention has strengthened or weakened over time, Models 2 to 4 in Tables 4 depict the division of our data into respectively the 1990s, 1980s, and 1970s decades. We ran these regressions without time fixed effects as it is precisely these time trends that we are exploring. The findings show that for both measures of competition for attention the effects decrease over time as they are weaker in the 1990s and stronger in the 1970s. For instance, a ten thousand unit increase in average frequency translates to an 21.1% significant decrease on forward cites in the 1970s, whereas in 1990s the decreasing effect shrank to 11.1%. We posit that this weakening trend may be because fields have become more and more specialized (our data contains 118 narrow subfields in the 1970s, 154 in the 1980s and 174 in the 1990s) even as science has expanded. Additionally, it may also be because of improvements in search tools through the decades which enable more focused search or other reasons that would be interesting to further investigate in future studies. These trends are robust to the 10-year mid-term and 3-year short-term forward citation time windows as evidenced in Tables 5 and 6.

Switching the dependent variable to the indicator variable of being in the top 1% of the citation distribution, we find similar trends in Table 7 as with the number of citations. First, when comparing the effect sizes contingent on the forward citation time window, we find stronger effects for the shorter window of 3 years (Models 1 to 3) than the longer 15-year window (Models 4 to 6). Second, when exploring time trends by decade, we also find that

competition for attention weakens from the 1970s to 1990s as shown respectively in Models 1 to 3 and Models 4 to 6.

[Insert Table 7 about here]

Finally, we also contrasted forward citations from the same narrow subfield and those from outside the subfield in Models 5 and 6 of Table 4. We observe similar effect sizes for average frequency for both home and foreign citations, and slightly stronger effects on home citations for number of MeSHs – 5.0% significant increase for home citations vs. 3.7% significant increase for foreign citations. Since the home subfield is more focused, articles need to compete slightly more for attention when the citations are from the same home subfield. It is also interesting to note that citations from foreign subfields depend more on signaling to due stronger information asymmetry in the foreign field, as the effect size for the average cumulative publication of authors, the number of authors, and the impact factor of the journal which are all signals of quality are all positive and stronger for foreign citations than home citations. These trends are also robust to the 10-year mid-term and 3-year short-term forward citation time windows as evidenced in Tables 5 and 6.

Bias Against Novelty

Main Effects. Using the same dataset and forward citations garnered 15 years after publication, we also replicated previously documented results pertaining to the "bias against novelty". There is an interesting tension between the finding that novel, "distant" recombinations of concepts lead to truly novel ideas (Simonton, 1999; Weitzman, 1998), while these ideas then have a hard time being recognized by the scientific community precisely due to their novelty (Kuhn, 1962; Margolis, 1993; Wang et al., 2017). Uzzi et al (2013) indicate a partial resolution to this seeming paradox through the finding that breakthrough papers tend to have mostly

conventional combinations of ideas, but also have a few highly atypical combinations as well, thus in some ways "straddling" the two effects. On an annual basis, we tabulated for the entire population of published articles that year in the life sciences the number of co-occurrences of all pairwise combinations of MeSH keywords.

Figure 4 illustrates the bias against novelty mechanism as operationalized using distance of recombination. As expected, we document a bias against novelty where citations and top 1% of cites 15 years after publication compared to articles in the same field and published in the same year (and their 95% confidence interval) decrease in average distance (see Figures 3A and 3B). The more a publication recombines distant topics the more novel are the recombined ideas, the less cites it will receive and the less likely it is to become a breakthrough paper. Thus, Hypothesis 3 is supported. Moreover, we also replicated the finding that breakthrough papers tend to contain mostly conventional ideas sprinkled with atypical combinations. We operationalized conventionality as the median distance of a paper and atypicality as the maximum distance of that paper, and found as expected that cites are decreasing in median distance, while they are increasing in maximum distance of dyadic recombination used in the publication (see Figures 3C and 3D)[7]. Conventionality of small median recombination distance is recognized, but so is atypicality with large maximum recombination distance.

[Insert Figure 4 about here]

Model 3 in Tables 2 and 3 document the bias against novelty view where the dyadic recombination distance of MeSH keywords is negatively correlated with forward citations with high significance and with being top cited with high significance respectively. Specifically, one

[7] I.e., for all the MeSH pairs in a paper, find the MeSH pairs that have the highest distance, and those that are at the median distance for that paper.

additional unit in average dyadic distance between MeSH keywords on the publication translates to a 62.8% significant decrease in forward citations (model 3 of Table 2) and an 87.5% significant decrease in the odds of the publication being in the top 1% of forward cites (model 3 of Table 3). Thus, publications recombining more distant (or novel) ideas tend to be associated with lower forward citation counts on average or with a decrease in the likelihood of being top cited, and is also consistent with our findings in Figures 3A and 3B.

All results above remain robust to combining the three independent variables, adding all control variables and time fixed effects, as shown in Models 4 and 5 in Tables 2 and 3. The findings are also robust to including heteroscedasticity robust standard errors. Model 5 of the two tables respectively show a 77.6% significant decrease in forward citations and a 95.3% significant decrease in the odds of the publication being in the top 1% of forward cites associated with one additional unit in average dyadic distance between MeSH keywords on the publication.

Variation of Effects. Similar to competition for attention, we explored trends and variations in bias against novelty contingent on the decade, the forward citation time window, as well as home versus foreign citations as shown in Tables 4 to 7. In line with prior research, we find stronger effects of novelty in foreign citations since these foreign and distant subfields are more likely to accept more diverse recombination of ideas, and stronger effects of bias against novelty for short-term recognition as the article's contributions are still uncertain (Wang et al., 2017). Moreover, we find that bias against novelty strengthens over the decades from the 1970s to the 1990s. With the expansion of science and increased specialization into more subfields, the distance of recombination increases over time and so might the bias against novelty effect.

Competition for Attention & Bias Against Novelty

Given the evidence that both mechanisms can work simultaneously, we also explored some ways their joint action could influence idea recognition. As mentioned in Hypotheses 4 and

5, we should expect low distance and low frequency to be best, and high distance and high frequency to be worst, for idea recognition. We divided our sample into high and low average frequency and average distance from each measure's median, and plotted the probability of being a top 1% forward cited breakthrough paper 15 years after publication, conditional on distance of recombination and competition for attention in Figure 5. We observed the expected pattern where articles recombining ideas from less common topics that are closer to one another are the most likely to be in the top 1%. Conversely, articles recombining ideas from more distant and familiar topics are the least likely highly recognized, as they not only suffer from the bias against novelty but also have to compete more for attention (see Figure 5A). Thus, Hypothesis 4 is supported. Similarly, using median splits of the number of MeSHs per publication, we also observed highest top 1% probability of recognition for low distance and high number of MeSHs (see Figure 5B). Again, both the competition for attention channels and bias against novelty are present, and Hypothesis 5 is supported.

[Insert Figure 5 about here]

Additional robustness checks

In addition to the robustness checks mentioned above, we ran a few additional ones in our regression analysis to ensure the consistency of our findings. In the interest of conciseness, regression results for these robustness tests are not shown herein, but available upon request from the corresponding author.

Narrow vs. Broad subfields. To ensure that our results are not sensitive to the Web of Science's subfield classifications, we also performed the same set of analyses using the broad subfield classification (*n = 16*) instead of the narrow classification (*n = 178*). This includes generating a new dependent variable indicating the top 1% of the citation distribution clustered by year of publication and broad subfield. All our results remain robust to this specification.

Different regression specifications. Given that the dependent variable tabulating the number of forward citations is non-negative and could be an over-dispersed count, we also used ordinary least squares models with robust standard errors. All our results using negative binomial specifications remain robust to the OLS specification (please refer to Tables A2 in the Appendix). Similarly, for the indicator dependent variable, we also ran probit regression models that returned robust and similar results to the logistic regression.

Discussion and Conclusion

By analyzing three decades of publication data in the life sciences, we add a more nuanced view to the recognition and spread of ideas by not only focusing on how it relates to combinatorial novelty but also introducing the perspective that articles compete for the attention of researchers who might build upon them. The findings indicate a highly significant negative relationship between average frequency and forward citations as well as with being a top cited paper. These results respectively imply that publishing in areas of increasing popularity is associated with fewer forward citations, and that publishing in areas of increasing popularity correlates with a decrease in the likelihood of being in the 1% top cited paper within the same narrow subfield and publication year. We also observed a highly significant positive relationship between the number of MeSH keywords of a publication with forward citations and with a publication with being top cited. This could be because the more indexed keywords facilitate discovery of the article for subsequent works to build onto and tend to increase attention brought on the article by subsequent works. Hence, this new perspective, what we refer to as the "competition for attention" view, makes sharp predictions about the recognition of ideas that seems to be supported by empirical evidence.

Furthermore, when exploring a few conditions under which these effects might vary, we find that competition for attention is stronger for the shorter citation time window, since shortly

after an article has been published its contributions are still uncertain. As time passes the article's contribution has settled and the effect of competition for attention becomes weaker. We also observe that the effects of competition for attention decrease over time as they are weaker in the 1990s and stronger in the 1970s. Finally, we find similar effect sizes for both home and foreign citations.

It is interesting to note that our empirical operationalization of these two mechanisms suggests that the attention view applies to single MeSH keywords, while novelty acts on pairs. Even though the bias against novelty might have been expected to apply to single MeSH keywords as well, it appears that the effect of the competition for attention is strong enough to override and reverse that effect, as was seen by outcome measures moving in the opposite direction than would be predicted by a bias against novelty measure applied at the single MeSH level. It might be interesting to speculate on why this might be the case. One plausible reason could be that when researchers search for and cite prior articles, they usually characterize the work by single keywords and cite them for a single idea. For instance, Einstein is cited for his theory of relativity, while Newton for gravity. They are probably much less likely to be searching for keyword pairs or higher order combinations. Thus, the effect of the competition for attention mechanism could be expected to be strongest at the single keyword level. In contrast, combinatorial novelty is manifested by the recombination of ideas (Fleming, 2001; Henderson & Clark, 1990; Weitzman, 1998) or keywords, and arguably, its effect gets stronger (or at least not decrease) as higher order combinations are introduced. Future work can hopefully explore further this intriguing symmetry of effects.

In sum, this work provides evidence that there is not only a bias against novelty in the recognition of ideas, but that the spread of knowledge is also governed by competition for attention. Hence, we introduce into the innovation literature on breakthrough recognition insights

from works in the organizational and marketing literatures on how people, products, and ideas having to compete for the attention of the audience they are trying to address due to the informational overload faced by decision makers (Hannan & Carroll, 1992; Hansen & Haas, 2001; Iyer & Katona, 2016; Ocasio, 2011; Shen et al., 2015). Taken together, the mechanisms of bias against novelty and competition for attention are reminiscent of the density dependence argument in population ecology (Hannan & Carroll, 1992). When new domains emerge – whether they are industries or ideas – they are less crowded but tend to have limited legitimacy. Hence, there is less competition for attention but also more bias against this novel area. Conversely, once a domain becomes more crowded, it has gained legitimacy hence there is there is less bias against it, but competition for attention increases.

These results call for an emphasis on the principles of competition for attention when exploring the spread of ideas. This competitive view departs from most prior work that have focused on cooperation and collaboration as drivers of idea recognition. Differentiating one's idea from other more common ones is in line with the priority-reward system in science (Merton, 1957). Coining a term or being early in studying a phenomenon or proposing a theory enables the researcher to direct the scientific conversation and claim a domain, which translates into further usage and recognition. Of course, areas evolve as greenfield spaces gain popularity with time and become more crowded. Having taken a first cross-sectional measurement of frequency and distance at the time of publication, future research can explore in more depth how the dynamics of these features affect idea recognition.

Finally, from a policy standpoint that looks to justify and optimize public and private investment in science in search of the next breakthrough, our findings provide a new dimension that policy makers and corporate lab managers can use in their decision-making process.

References

Azoulay, P., Güler, I., Koçak, Ö., Murciano-Goroff, R., & Anttila-Hughes, J. 2012. Are recombinant scientific articles more impactful? ***Science***.

Bikard, M. 2018. Made in Academia: The Effect of Institutional Origin on Inventors' Attention to Science. ***Organization Science***, Conditionally accepted.

Bordalo, P., Gennaioli, N., & Shleifer, A. 2016. Competition for Attention. ***The Review of Economic Studies***, 83(2): 481-513.

Boudreau, K. J., Guinan, E. C., Lakhani, K. R., & Riedl, C. 2016. Looking across and looking beyond the knowledge frontier: Intellectual distance, novelty, and resource allocation in science. ***Management Science***, 62(10): 2765-2783.

Burt, R. S. 2004. Structural Holes and Good Ideas. ***The American Journal of Sociology***, 110(2): 349-399.

Chai, S. 2017. Near Misses in the Breakthrough Discovery Process. ***Organization Science***, 28(3): 411-428.

Cyert, R. M. & March, J. G. 1963. A behavioral theory of the firm. ***Englewood Cliffs, NJ***, 2: 169-187.

Daft, R. L. & Weick, K. E. 1984. Toward a model of organizations as interpretation systems. ***Academy of management review***, 9(2): 284-295.

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., & Mello, C. C. 1998. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. ***Nature***, 391(6669): 806.

Fleming, L. 2001. Recombinant Uncertainty in Technological Search. ***Management Science***, 47(1): 117-132.

Fleming, L., Mingo, S., & Chen, D. 2007. Collaborative Brokerage, Generative Creativity, and Creative Success. ***Administrative Science Quarterly***, 52(3): 443-475.

Foster, J. G., Rzhetsky, A., & Evans, J. A. 2015. Tradition and Innovation in Scientists' Research Strategies. ***American Sociological Review***, 80(5): 875-908.

Gardner, H. 1993. ***Creating minds: An anatomy of creativity seen through the lives of Freud, Einstein, Picasso, Stravinsky, Eliot, Graham, and Gandhi***. New York: Basic Books.

Hannan, M. T. & Carroll, G. R. 1992. ***Dynamics of organizational populations: Density, legitimation, and competition***: Oxford University Press.

Hansen, M. T. & Haas, M. R. 2001. Competing for attention in knowledge markets: Electronic document dissemination in a management consulting company. ***Administrative Science Quarterly***, 46(1): 1-28.

Hargadon, A. & Sutton, R. I. 1997. Technology Brokering and Innovation in a Product Development Firm. ***Administrative Science Quarterly***, 42(4): 716-749.

Henderson, R. M. & Clark, K. B. 1990. Architectural Innovation: The Reconfiguration of Existing Product Technologies and the Failure of Established Firms. ***Administrative Science Quarterly***, 35(1): 9-30.

Hennessey, B. A. & Amabile, T. M. 2010. Creativity. ***Annual Review of Psychology***, 61(1): 569-598.

Hoffman, A. J. & Ocasio, W. 2001. Not all events are attended equally: Toward a middle-range theory of industry attention to external events. ***Organization science***, 12(4): 414-434.

Iyer, G. & Katona, Z. 2016. Competing for Attention in Social Communication Markets. ***Management Science***, 62(8): 2304-2320.

Kuhn, T. S. 1962. ***The structure of scientific revolutions***. Chicago: University of Chicago Press.

Lifshitz-Assaf, H. 2017. Dismantling Knowledge Boundaries at NASA: From Problem Solvers to Solution Seekers. ***Administrative Science Quarterly***.
March, J. G. & Simon, H. A. 1958. Organizations.
March, J. G. & Shapira, Z. 1987. Managerial perspectives on risk and risk taking. ***Management science***, 33(11): 1404-1418.
Margolis, H. 1993. ***Paradigms & barriers: How habits of mind govern scientific belief***. Chicago: University of Chicago Press.
McEvily, B. & Zaheer, A. 1999. Bridging Ties: A Source of Firm Heterogeneity in Competitive Capabilities. ***Strategic Management Journal***, 20(12): 1133-1156.
Merton, R. K. 1957. Priorities in Scientific Discovery: A Chapter in the Sociology of Science. ***American Sociological Review***, 22(6): 635-659.
Mueller, J. S., Melwani, S., & Goncalo, J. A. 2012. The bias against creativity: Why people desire but reject creative ideas. ***Psychological science***, 23(1): 13-17.
Nigam, A. & Ocasio, W. 2010. Event attention, environmental sensemaking, and change in institutional logics: An inductive analysis of the effects of public attention to Clinton's health care reform initiative. ***Organization Science***, 21(4): 823-841.
Nooteboom, B., Van Haverbeke, W., Duysters, G., Gilsing, V., & Van den Oord, A. 2007. Optimal cognitive distance and absorptive capacity. ***Research policy***, 36(7): 1016-1034.
Ocasio, W. 1997. Towards an attention-based view of the firm. ***Strategic management journal***: 187-206.
Ocasio, W. 2011. Attention to Attention. ***Organization Science***, 22(5): 1286-1296.
Rietzschel, E. F., Nijstad, B. A., & Stroebe, W. 2010. The selection of creative ideas after individual idea generation: Choosing between creativity and impact. ***British journal of psychology***, 101(1): 47-68.
Schumpeter, J. A. 1942. ***Capitalism, socialism, and democracy***. New York: Harper Perennial.
Shen, W., Hu, Y. J., & Rees Ulmer, J. 2015. Competing for attention: an empirical study of online reviewers' strategic behavior. ***Management Information Systems Quarterly***, 39(3): 683-696.
Simon, H. A. & Barnard, C. I. 1947. ***Administrative behavior: A study of decision-making processes in administrative organization***: Macmillan.
Simonton, D. K. 1999. ***Origins of genius: Darwinian perspectives on creativity***. Oxford: Oxford University Press.
Singh, J. & Fleming, L. 2010. Lone Inventors as Sources of Breakthroughs: Myth or Reality? ***Management Science***, 56(1): 41-56.
Torvik, V. I. & Smalheiser, N. R. 2009. Author Name Disambiguation in MEDLINE. ***ACM transactions on knowledge discovery from data***, 3(3): 1-29.
Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. 2013. Atypical combinations and scientific impact. ***Science***, 342(6157): 468-472.
Verhoeven, D., Bakker, J., & Veugelers, R. 2016. Measuring technological novelty with patent-based indicators. ***Research Policy***, 45(3): 707-723.
Wang, J., Veugelers, R., & Stephan, P. 2017. Bias against novelty in science: A cautionary tale for users of bibliometric indicators. ***Research Policy***.
Waterhouse, P. M., Graham, M. W., & Wang, M.-B. 1998. Virus Resistance and Gene Silencing in Plants can be Induced by Simultaneous Expression of Sense and Antisense RNA. ***Proceedings of the National Academy of Sciences of the United States of America***, 95(23): 13959-13964.

Weitzman, M. L. 1998. Recombinant Growth. ***The Quarterly Journal of Economics***, 113(2): 331-360.

Wuchty, S., Jones, B. F., & Uzzi, B. 2007. The Increasing Dominance of Teams in Production of Knowledge. ***Science***, 316(5827): 1036-1039.

Variable	Observation	Mean	Std. Dev.	Min	Max
publication year	5,300,029	1988.033	8.034179	1970	1999
# of cites (15yr)	5,300,029	26.98885	78.01378	0	49299
# of cites (10yr)	5,300,029	22.57974	57.42935	0	30330
# of cites (3yr)	5,300,029	8.156929	17.84104	0	3969
top1% indicator (15yr)	5,300,029	0.0099428	0.0992165	0	1
top1% indicator (10yr)	5,300,029	0.0099447	0.0992258	0	1
top1% indicator (3yr)	5,300,029	0.0099543	0.0992734	0	1
home cites (15yr)	5,300,029	14.46142	39.43114	0	31166
home cites (10yr)	5,300,029	12.18305	30.57669	0	19288
home cites (3yr)	5,300,029	4.520233	9.699633	0	2199
foreign cites (15yr)	5,300,029	12.52743	52.71717	0	39139
foreign cites (10yr)	5,300,029	10.39669	37.24478	0	15828
foreign cites (3yr)	5,300,029	3.636696	11.72168	0	3550
avg frequency(/10k)	5,232,299	2.900894	1.953521	0.0001	16.06245
# of MeSHs	5,232,299	11.5193	4.393472	1	53
average distance	5,230,303	0.1759706	0.1228729	0.0000263	1
median distance	5,230,303	0.0637036	0.112808	0.0000149	1
maximum distance	5,230,303	0.8704594	0.276203	0.0000263	1
minimum distance	5,230,303	0.0020553	0.0246364	0.00000776	1
avg cumulative pubs	5,160,163	35.01917	39.43043	1	1361
# of authors	5,300,029	3.609467	2.397941	1	394
impact factor	5,149,805	0.5902668	1.877485	0	70.5

Table 1.

This table shows the descriptive statistics for all variables in our sample of ~5.3M articles in the life sciences.

Note: The variation in the number of observations across the variables was due to missing data on those variables.

# of cites (15yr)	Model 1	Model 2	Model 3	Model 4	Model 5
avg frequency(/10k)	-0.0971***			-0.0980***	-0.126***
	(0.000702)			(0.000920)	(0.000967)
# of MeSHs		0.0698***		0.0591***	0.0430***
		(0.000301)		(0.000312)	(0.000312)
avg distance			-0.990***	-1.573***	-1.495***
			(0.00984)	(0.0127)	(0.0111)
avg cumulative pubs					0.00388***
					(0.0000348)
# of authors					0.0602***
					(0.000554)
impact factor					0.143***
					(0.000764)
constant	3.560***	2.444***	3.464***	3.106***	2.672***
	(0.00244)	(0.00428)	(0.00218)	(0.00707)	(0.00964)
lnalpha					
constant	0.359***	0.325***	0.370***	0.299***	0.224***
	(0.00119)	(0.00146)	(0.00122)	(0.00136)	(0.00150)
subfield fe	yes	yes	yes	yes	yes
time fe					yes
N	5232299	5232299	5230303	5230303	5050334
Log lik.	-22264426.0	-22153207.3	-22294057.4	-22063798.2	-21035192.2

Standard errors in parentheses

* p<0.1, ** p<0.05, *** p<0.01

Table 2.

This table shows negative binomial regression models with narrow subfield fixed effects. The dependent variable for all models are forward citations each article garners up to 15 years after its publication. The independent variables are: (a) in Models 1 and 2, respectively the average frequency of all MeSH keywords and the total number of MeSH keywords in an article as measures of competition for attention, (b) in Model 3, the average distance between all dyads of MeSH keywords in an article as measure of novelty, (c) in Model 4, all measures of competition for attention and novelty, and (d) in Model 5, all measures controlling for the number of authors in the article, the average number of cumulative publications for authors of the article, the journal impact factor and time fixed effects.

top 1% of cites (15yr)	Model 1	Model 2	Model 3	Model 4	Model 5
avg frequency(/10k)	-0.108***			-0.133***	-0.126***
	(0.00244)			(0.00301)	(0.00361)
# of MeSHs		0.101***		0.0936***	0.0884***
		(0.000891)		(0.000938)	(0.00116)
avg distance			-2.077***	-3.237***	-3.048***
			(0.0414)	(0.0508)	(0.0539)
avg cumulative pubs					0.00576***
					(0.0000639)
# of authors					0.0779***
					(0.00286)
impact factor					0.143***
					(0.00122)
constant	-4.303***	-5.861***	-4.260***	-4.892***	-5.478***
	(0.00758)	(0.0130)	(0.00760)	(0.0191)	(0.0474)
times fe					yes
N	5232299	5232299	5230303	5230303	5050334
Log lik.	-291491.0	-286490.1	-291050.8	-283608.6	-260344.8

Standard errors in parentheses
* p<0.1, ** p<0.05, *** p<0.01

Table 3.

This table shows logistic regression models. The dependent variable for all models are a dummy variable of whether the 15-year cumulated forward citations of the article is in the top 1% of the citation distribution clustered by publication year and narrow subfield. Given that the dependent variable is constructed at the subfield level, we do not include subfield fixed effects. The independent variables are: (a) in Models 1 and 2, respectively the average frequency of all MeSH keywords and the total number of MeSH keywords in an article as measures of competition for attention, (b) in Model 3, the average distance between all dyads of MeSH keywords in an article as measure of novelty, (c) in Model 4, all measures of competition for attention and novelty, and (d) in Model 5, all measures controlling for the number of authors in the article, the average number of cumulative publications for authors of the article, the journal impact factor and time fixed effects.

# of cites (15yr)	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
	total cites	total cites (90s)	total cites (80s)	total cites (70s)	home cites	foreign cites
avg frequency(/10k)	-0.126***	-0.118***	-0.176***	-0.246***	-0.127***	-0.126***
	(0.000967)	(0.00113)	(0.00175)	(0.00482)	(0.00088)	(0.00141)
# of MeSHs	0.0430***	0.0304***	0.0584***	0.0514***	0.0485***	0.0367***
	(0.000312)	(0.00040)	(0.00071)	(0.00088)	(0.00031)	(0.00041)
avg distance	-1.495***	-1.819***	-1.521***	-1.381***	-1.819***	-1.131***
	(0.0111)	(0.01590)	(0.01870)	(0.02920)	(0.01150)	(0.01460)
avg cumulative pubs	0.00388***	0.00337***	0.00474***	0.00436***	0.00332***	0.00451***
	(0.0000348)	(0.00004)	(0.00008)	(0.00011)	(0.00003)	(0.00005)
# of authors	0.0602***	0.0524***	0.0849***	0.119***	0.0500***	0.0719***
	(0.000554)	(0.00060)	(0.00123)	(0.00251)	(0.00057)	(0.00073)
impact factor	0.143***	0.0771***	27.44***	53.14***	0.125***	0.159***
	(0.000764)	(0.00045)	(1.23800)	(3.59400)	(0.00084)	(0.00094)
constant	2.672***	3.147***	2.803***	2.667***	1.912***	2.036***
	(0.00964)	(0.00957)	(0.01130)	(0.02880)	(0.00962)	(0.01280)
lnalpha						
constant	0.224***	0.188***	0.284***	0.267***	0.425***	0.606***
	(0.00150)	(0.00165)	(0.00290)	(0.00484)	(0.00134)	(0.00174)
subfield fe	yes	yes	yes	yes	yes	yes
time fe	yes				yes	yes
N	5050334	2541121	1590997	918216	5050334	5050334
Log lik.	-21035192.2	-10803361.2	-6555584.4	-3698690.1	-17872506.8	-16765801.7

Standard errors in parentheses
* p<0.1, ** p<0.05, *** p<0.01

Table 4.

This table shows negative binomial regression models with narrow subfield fixed effects. The dependent variables for all models are forward citations with the *15-year window* after publication with (a) total number of cites in Model 1 (b) in total number of cites for 90s, 80s and 70s in Models 2 to 4, (c) number of home cites in Model 5, and (d) number of foreign cites in Model 6. The independent variables are all measures of competition for attention and bias against novelty while controlling for the number of authors in the article, the average number of cumulative publications for authors of the article, the journal impact factor and time fixed effects.

# of cites (10yr)	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
	total cites	total cites (90s)	total cites (80s)	total cites (70s)	home cites	foreign cites
avg frequency(/10k)	-0.136***	-0.119***	-0.194***	-0.253***	-0.138***	-0.134***
	(0.000873)	(0.00102)	(0.00174)	(0.00294)	(0.000830)	(0.00125)
# of MeSHs	0.0427***	0.0308***	0.0576***	0.0507***	0.0478***	0.0367***
	(0.000270)	(0.000361)	(0.000619)	(0.000569)	(0.000279)	(0.000349)
avg distance	-1.607***	-1.875***	-1.666***	-1.444***	-1.924***	-1.248***
	(0.00964)	(0.0141)	(0.0172)	(0.0205)	(0.0105)	(0.0121)
avg cumulative pubs	0.00415***	0.00357***	0.00518***	0.00472***	0.00355***	0.00481***
	(0.0000334)	(0.0000414)	(0.0000730)	(0.0000899)	(0.0000326)	(0.0000453)
# of authors	0.0627***	0.0539***	0.0895***	0.126***	0.0521***	0.0751***
	(0.000513)	(0.000577)	(0.00112)	(0.00173)	(0.000528)	(0.000684)
impact factor	0.142***	0.0967***	27.20***	54.60***	0.124***	0.159***
	(0.000762)	(0.000458)	(1.274)	(3.469)	(0.000842)	(0.000933)
constant	2.447***	2.960***	2.616***	2.442***	1.693***	1.805***
	(0.00859)	(0.00833)	(0.0103)	(0.0165)	(0.00870)	(0.0113)
lnalpha						
constant	0.212***	0.167***	0.274***	0.225***	0.410***	0.605***
	(0.00130)	(0.00159)	(0.00265)	(0.00331)	(0.00124)	(0.00148)
subfield fe	yes	yes	yes	yes	yes	yes
time fe	yes				yes	yes
N	5050334	2541121	1590997	918216	5050334	5050334
Log lik.	-20112742.1	-10418975.0	-6187641.7	-3499277.0	-17015481.4	-15864243.2

Standard errors in parentheses
* p<0.1, ** p<0.05, *** p<0.01

Table 5.

This table shows negative binomial regression models with narrow subfield fixed effects. The dependent variables for all models are forward citations with the *10-year window* after publication with (a) total number of cites in Model 1 (b) in total number of cites for 90s, 80s and 70s in Models 2 to 4, (c) number of home cites in Model 5, and (d) number of foreign cites in Model 6. The independent variables are all measures of competition for attention and bias against novelty while controlling for the number of authors in the article, the average number of cumulative publications for authors of the article, the journal impact factor and time fixed effects.

# of cites (3yr)	Model 1 total cites	Model 2 total cites (90s)	Model 3 total cites (80s)	Model 4 total cites (70s)	Model 5 home cites	Model 6 foreign cites
avg frequency(/10k)	-0.168***	-0.150***	-0.235***	-0.301***	-0.173***	-0.167***
	(0.000727)	(0.000846)	(0.00152)	(0.00176)	(0.000718)	(0.00101)
# of MeSHs	0.0376***	0.0281***	0.0509***	0.0438***	0.0416***	0.0331***
	(0.000208)	(0.000298)	(0.000447)	(0.000367)	(0.000217)	(0.000284)
avg distance	-1.902***	-2.266***	-1.924***	-1.707***	-2.185***	-1.582***
	(0.0078)	(0.0123)	(0.0137)	(0.0145)	(0.00850)	(0.0102)
avg cumulative pubs	0.00473***	0.00413***	0.00565***	0.00527***	0.00399***	0.00550***
	(0.0000293)	(0.0000375)	(0.0000624)	(0.0000712)	(0.0000288)	(0.0000397)
# of authors	0.0685***	0.0582***	0.0961***	0.132***	0.0564***	0.0836***
	(0.000435)	(0.000534)	(0.000859)	(0.00115)	(0.000429)	(0.000613)
impact factor	0.157***	0.106***	24.10***	51.28***	0.140***	0.173***
	(0.000721)	(0.000417)	(1.243)	(3.267)	(0.000824)	(0.000825)
constant	1.609***	2.056***	1.777***	1.604***	0.893***	0.930***
	(0.00709)	(0.00622)	(0.00808)	(0.00772)	(0.00741)	(0.00924)
lnalpha						
constant	0.230***	0.229***	0.268***	0.162***	0.404***	0.685***
	(0.00116)	(0.00166)	(0.00212)	(0.00227)	(0.00119)	(0.00148)
subfield fe	yes	yes	yes	yes	yes	yes
time fe	yes				yes	yes
N	5050334	2541121	1590997	918216	5050334	5050334
Log lik.	-15112118.9	-7834514.9	-4643349.3	-2629650.8	-12396985.8	-11059971.6

Standard errors in parentheses
* p<0.1, ** p<0.05, *** p<0.01

Table 6.

This table shows negative binomial regression models with narrow subfield fixed effects. The dependent variables for all models are forward citations with the *3-year window* after publication with (a) total number of cites in Model 1 (b) in total number of cites for 90s, 80s and 70s in Models 2 to 4, (c) number of home cites in Model 5, and (d) number of foreign cites in Model 6. The independent variables are all measures of competition for attention and bias against novelty while controlling for the number of authors in the article, the average number of cumulative publications for authors of the article, the journal impact factor and time fixed effects.

top 1% of cites	Model 1 15 yr - 90s	Model 2 15 yr -80s	Model 3 15 yr - 70s	Model 4 3 yr - 90s	Model 5 3 yr - 80s	Model 6 3 yr - 70s
avg frequency(/10k)	-0.0877***	-0.283***	-0.319***	-0.133***	-0.366***	-0.442***
	(0.00407)	(0.00704)	(0.0126)	(0.00412)	(0.00716)	(0.0130)
# of MeSHs	0.0562***	0.120***	0.0966***	0.0632***	0.127***	0.0943***
	(0.00156)	(0.00223)	(0.00201)	(0.00153)	(0.00220)	(0.00199)
avg distance	-3.837***	-3.013***	-2.328***	-5.234***	-4.416***	-3.807***
	(0.0839)	(0.0883)	(0.111)	(0.0874)	(0.0945)	(0.123)
avg cumulative pubs	0.00493***	0.00630***	0.00651***	0.00515***	0.00715***	0.00761***
	(0.0000776)	(0.000132)	(0.000195)	(0.0000789)	(0.000135)	(0.000213)
# of authors	0.0594***	0.145***	0.177***	0.0681***	0.177***	0.227***
	(0.00300)	(0.00350)	(0.00576)	(0.00309)	(0.00342)	(0.00571)
impact factor	0.118***	2.038***	9.835***	0.125***	1.599***	11.72***
	(0.00108)	(0.271)	(1.138)	(0.00108)	(0.227)	(1.069)
constant	-5.225***	-5.585***	-5.690***	-5.064***	-5.448***	-5.461***
	(0.0329)	(0.0399)	(0.0479)	(0.0316)	(0.0378)	(0.0460)
times fe	yes	yes	yes	yes	yes	yes
N	2541121	1590997	918216	2541121	1590997	918216
Log lik.	-130050.0	-82953.2	-47235.1	-128098.1	-81297.6	-46259.2

Standard errors in parentheses
* p<0.1, ** p<0.05, *** p<0.01

Table 7.

This table shows logistic regression models. The dependent variable for all models are a dummy variable of whether the cumulated forward citations of the article is in the top 1% of the citation distribution clustered by publication year and narrow subfield with the (a) 15-year window in Models 1 to 3, and (b) 3-year window in Models 4 to 6. Given that the dependent variable is constructed at the subfield level, we do not include subfield fixed effects. The independent variables are all measures of competition for attention and bias against novelty while controlling for the number of authors in the article, the average number of cumulative publications for authors of the article, the journal impact factor and time fixed effects.

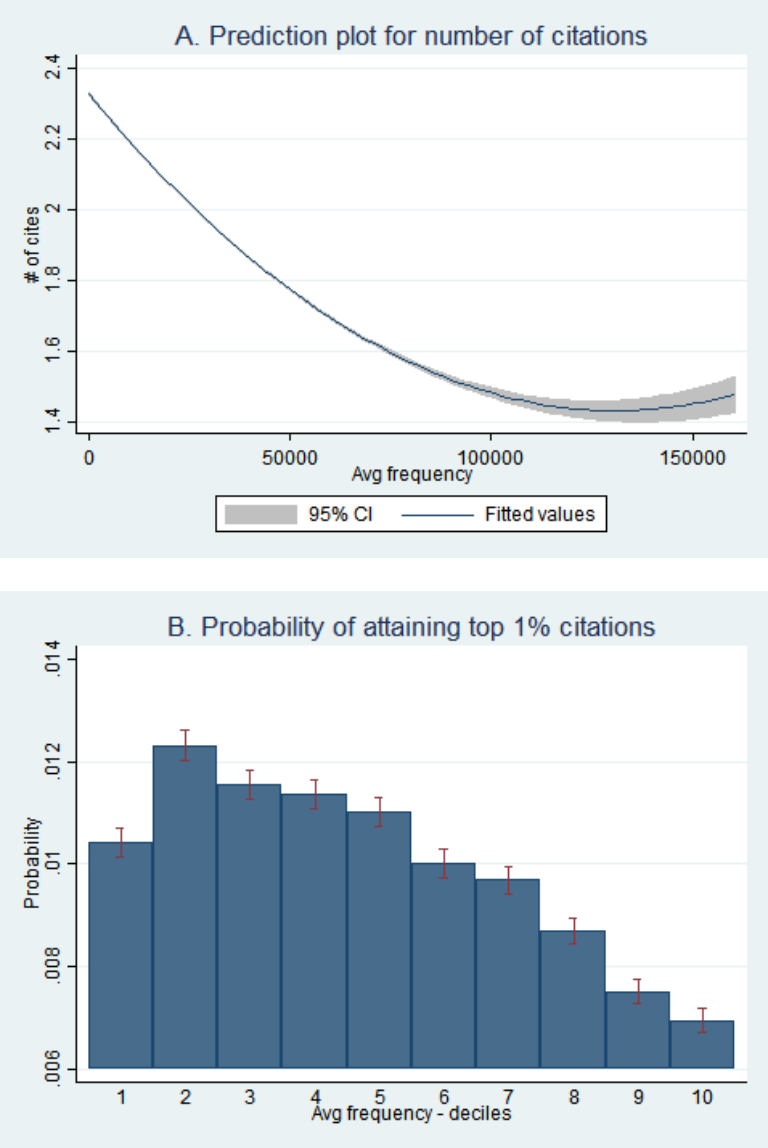


Figure 1. Competition for attention operationalized using average frequency. (A) The total number of forward citations 15 years after publication (and the 95% confidence interval) are decreasing as the average frequency of MeSH keywords used in the publication increase. (B) The probability of attaining the top 1% of cites 15 years after publication (and the 95% confidence interval) compared to articles in the same field and published in the same year are also decreasing in average frequency. The more a publication covers familiar topics the more it has to compete for attention, the less cites it will receive and the less likely it is to become a breakthrough paper.

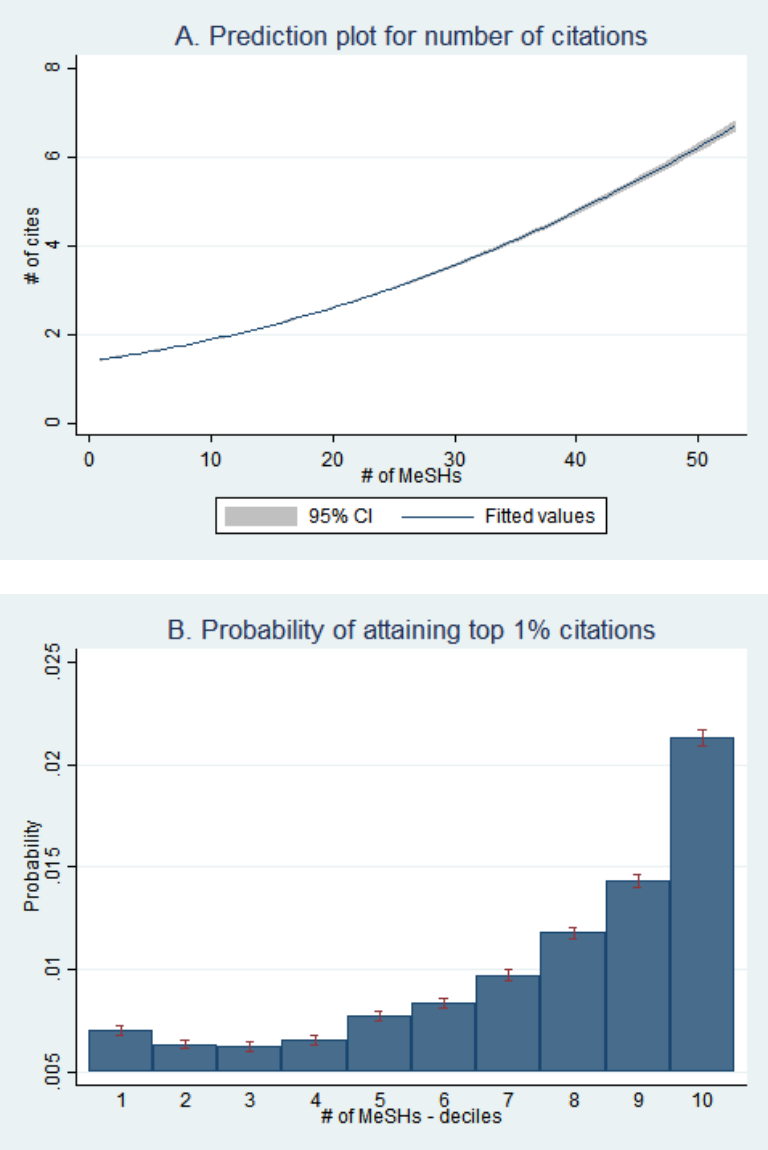
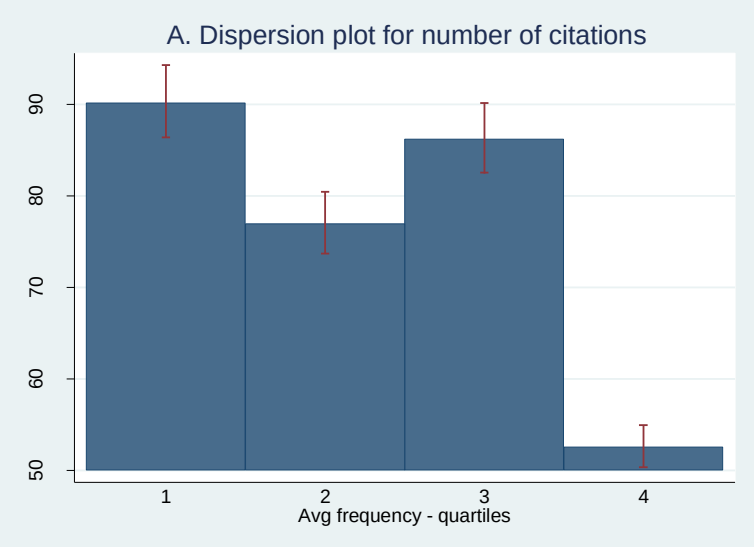


Figure 2. Competition for attention operationalized using number of MeSHs. (A) The total number of forward citations 15 years after publication (and the 95% confidence interval) are increasing in the number of MeSH keywords used in the publication. (B) The probability of attaining the top 1% of cites 15 years after publication (and the 95% confidence interval) compared to articles in the same field and published in the same year are also increasing in the number of MeSHs. The more topics a publication covers the more likely it is found when searched, the more cites it will receive and the more likely it is to become a breakthrough paper.



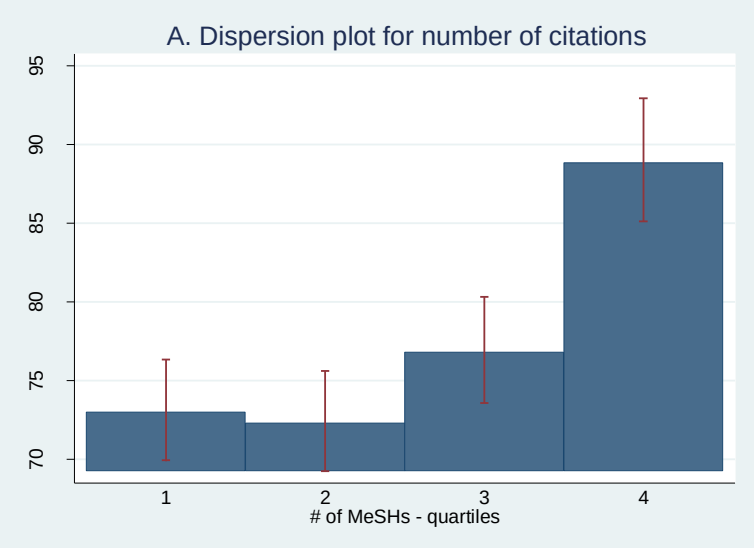


Figure 3. Dispersion (depicted using the standard deviation of the number of citations) quartile plots for competition for attention operationalized using average frequency and number of MeSHs. (A) Dispersion for the total number of forward citations 15 years after publication (and the 95% confidence interval) are decreasing as the average frequency of MeSH keywords used in the publication increase. (B) Dispersion for the total number of forward citations 15 years after publication (and the 95% confidence interval) are decreasing in the number of MeSH keywords used in the publication. The more competition for attention present (high average frequency or low # of MeSHs) is associated with more uncertainty in idea recognition as measured by the number of forward citations.

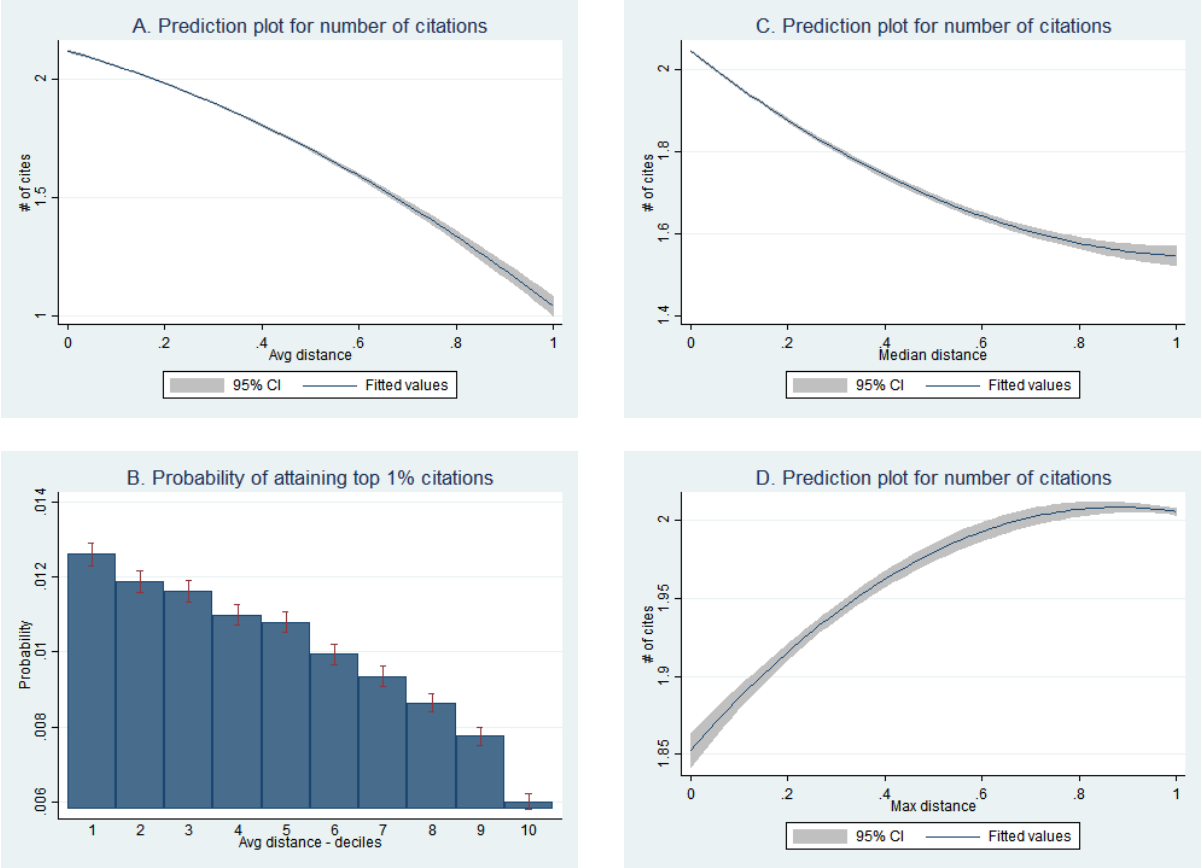
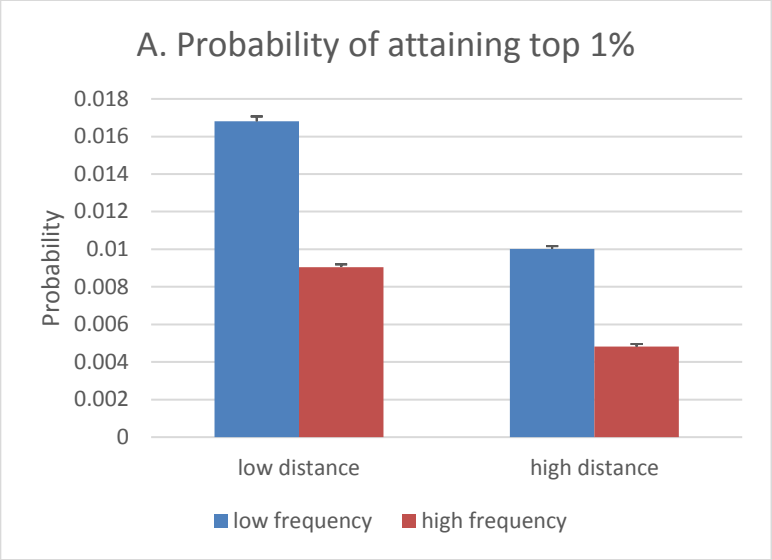


Figure 4. Bias against novelty operationalized using distance of recombination. (A) The total number of forward citations 15 years after publication (and the 95% confidence interval) are decreasing in the average distance of dyadic recombination used in the publication. (B) The probability of attaining the top 1% of cites 15 years after publication (and the 95% confidence interval) compared to articles in the same field and published in the same year are also decreasing in average distance. The more a publication recombines distant topics the more novel are the recombined ideas, the less cites it will receive and the less likely it is to become a breakthrough paper. (C & D) The total number of forward citations 15 years after publication (and the 95% confidence interval) are decreasing in the median distance and increasing in the maximum distance of dyadic recombination used in the publication. Conventionality of small median recombination distance is recognized, but so is atypicality with large maximum recombination distance.



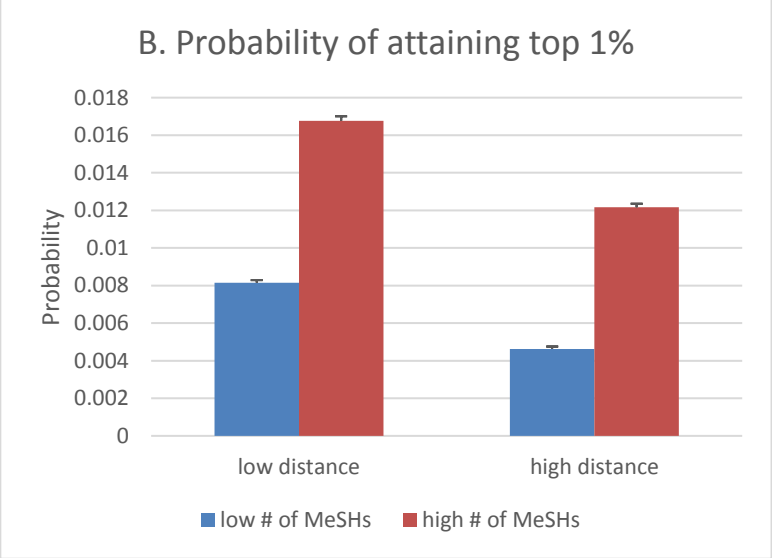


Figure 5. Probability of breakthrough paper, conditional on distance of recombination and competition for attention. Competition for attention is operationalized using (A) the average frequency of MeSH keywords, and (B) the number of MeSH keywords. Both the competition for attention channels and bias against novelty are present as articles covering most common topics (or with the least number of MeSHs) that recombine most distant ideas from one another are the least likely to be in the top 1% of cites 15 years after publication.

	publication year	# of cites (15yr)	# of cites (3yr)	top1% indicator (15yr)	top1% indicator (3yr)	avg frequency (/10k)	# of MeSHs
publication year	1						
# of cites (15yr)	0.0302	1					
# of cites (3yr)	0.0587	0.7517	1				
top1% indicator (15yr)	0.001	0.4902	0.4954	1			
top1% indicator (3yr)	0.0015	0.4267	0.5641	0.6266	1		
avg frequency (/10k)	0.3732	-0.0596	-0.0933	-0.0182	-0.0221	1	
# of MeSHs	0.0979	0.1141	0.1599	0.0505	0.0554	-0.2459	1
avg distance	-0.057	-0.0393	-0.0625	-0.0225	-0.03	-0.3957	-0.0025
median distance	-0.07	-0.0293	-0.0436	-0.0144	-0.0185	-0.3472	-0.1075
maximum distance	0.0173	0.0115	0.0088	-0.0031	-0.007	-0.2774	0.3547
minimum distance	-0.0316	-0.0124	-0.0161	-0.004	-0.005	-0.1057	-0.1276
avg cumulative pubs	0.1769	0.0513	0.0795	0.0356	0.039	0.1219	0.0381
# of authors	0.2483	0.0908	0.1352	0.0504	0.0599	0.0805	0.2306
impact factor	0.3658	0.1357	0.2389	0.0749	0.0857	0.0313	0.1378

	avg distance	median distance	max distance	min distance	avg cumulative pubs	# of authors	impact factor
avg distance	1						
median distance	0.8047	1					
maximum distance	0.4856	0.1835	1				
minimum distance	0.2493	0.3434	-0.0092	1			
avg cumulative pubs	-0.083	-0.0719	-0.0488	-0.0251	1		
# of authors	-0.0808	-0.0916	0.0705	-0.0358	0.0336	1	
impact factor	-0.0336	-0.0327	0.0149	-0.0094	0.091	0.1331	1

Table A1.

This table shows the correlation for major variables in our sample of ~5.3M articles in the life sciences.

# of cites (15yr)	Model 1	Model 2	Model 3	Model 4	Model 5
avg frequency(/10k)	-1.542***			-1.424***	-2.058***
	(0.0186)			(0.0207)	(0.0227)
# of MeSHs		1.893***		1.752***	1.316***
		(0.0078)		(0.0080)	(0.0084)
avg distance			-26.75***	-34.84***	-31.91***
			(0.2800)	(0.3020)	(0.3000)
avg cumulative pubs					0.0749***
					(0.0009)
# of authors					1.990***
					(0.0146)
impact factor					6.139***
					(0.0220)
constant	31.48***	5.199***	31.72***	17.09***	6.080***
	(0.0637)	(0.0956)	(0.0598)	(0.1500)	(0.3330)
subfield fe	yes	yes	yes	yes	yes
time fe					yes
N	5232299	5232299	5230303	5230303	5050334
R-sq	0.0206	0.0303	0.021	0.0329	0.0574

Standard errors in parentheses

* p<0.1, ** p<0.05, *** p<0.01

Table A2.

This table shows OLS regression models with narrow subfield fixed effects. The dependent variable for all models are forward citations each article garners up to 15 years after its publication. The independent variables are: (a) in Models 1 and 2, respectively the average frequency of all MeSH keywords and the total number of MeSH keywords in an article as measures of competition for attention, (b) in Model 3, the average distance between all dyads of MeSH keywords in an article as measure of novelty, (c) in Model 4, all measures of competition for attention and novelty, and (d) in Model 5, all measures controlling for the number of authors in the article, the average number of cumulative publications for authors of the article, the journal impact factor and time fixed effects.