Published in final edited form as:

Scientometrics. 2019 March ; 188(3): 787–808. doi:10.1007/s11192-018-2991-4.

Edge Factors: Scientific Frontier Positions of Nations

Mikko Packalen

University of Waterloo

Abstract

A key decision in scientific work is whether to build on novel or well-established ideas. Because exploiting new ideas is often harder than more conventional science, novel work can be especially dependent on interactions with colleagues, the training environment, and ready access to potential collaborators. Location may thus influence the tendency to pursue work that is close to the edge of the scientific frontier in the sense that it builds on recent ideas. We calculate for each nation its position relative to the edge of the scientific frontier by measuring its propensity to build on relatively new ideas in biomedical research. Text analysis of 20+ million publications shows that the United States and South Korea have the highest tendencies for novel science. China has become a leader in favoring newer ideas when working with basic science ideas and research tools, but is still slow to adopt new clinical ideas. Many locations remain far behind the leaders in terms of their tendency to work with novel ideas, indicating that the world is far from flat in this regard.

Keywords

science; novelty; impact factor; new ideas; idea adoption

1. Introduction

Knowledge production is an increasingly global endeavor. In spite of robust increases in scientific production by the traditional leaders – including the United States, the United Kingdom, and Japan – their relative share has decreased in recent decades because the pace of growth in science by other nations – including China, South Korea, India, and Brazil – has been even more rapid (Freeman, 2013; National Science Board, 2016). The share of international collaborations has also increased, as has the share of citations to papers with foreign authors (Freeman, 2013, National Science Board, 2016). This spread of knowledge production has not been unexpected. It was anticipated long ago by Marshall (1920) that improved communication technologies would make it easier to learn about new discoveries regardless of location and that this would lead to the pursuit of creative work in more diverse places.

While this perspective suggests a diminishing influence for location in scientific work, location may in fact continue to have considerable import in science. This is because learning about which new ideas exist may not have been an important benefit of location for quite some time and because location likely still impacts the fertility of creative work in other important ways. Lucas (2004), for example, has emphasized the continuing

dependence of knowledge production activities on location through the benefits that accrue from daily interactions with colleagues, the training environment, and ready access collaborators.

One potential remaining influence of location stems from the fact that when new ideas are first discovered, they are often raw and poorly understood. The ideas only gradually mature into useful advances after a community of scientists tries them out and develops them. But such work is hard, harder than work that builds on well-established ideas. One indication that work that tries out new ideas is indeed harder than more conventional science is that such work is linked with larger team size (Packalen and Bhattacharya, 2016). Thus, when a scientist seeks to build on a recent advance, it is beneficial to be surrounded by a community of scholars with whom to debate about which new ideas to try out and how (Marshall, 1920; Kuhn, 1962, 1977; Usher, 1929). Daily interactions with colleagues, the training environment, and ready access to potential collaborators thus become especially important in work that is close to the edge of the scientific frontier in the sense that the work builds on recent advances.

Because such local factors influence the fertility of the debates that seek to unlock the mysteries of new ideas, the tendency to work with new ideas can be expected to vary by location. This mechanism – and thus the import of location – may even be increasingly influential, for Jones (2010) shows that reaching the scientific frontier now involves even more work than before as evidenced by increases in training times, specialization, and teamwork.

Therefore, even as the pursuit of science spreads to more diverse places, location may well continue to have an important influence on *what kind of science* is pursued – through the impact that location may have on the ability to work with novel ideas. Identifying where barriers to knowledge adoption still exist is thus crucial for understanding the role of location in knowledge production and for designing policies that can help eliminate the remaining barriers.

We calculate each nation's propensity to publish biomedical work that is close to the edge of the scientific frontier in the sense that it builds on relatively recent ideas. The results reveal each nation's position on the scientific frontier: what share of its contributions to biomedical science build on relatively new ideas vs. well-established ideas. Our empirical analysis is focused on biomedicine because it is an important area of science and because of the availability of the Pubmed/MEDLINE database on over 24 million biomedical research papers.

We refer to our constructed measure of novelty as *the edge factor*. Whereas the familiar impact factor measures scientific influence (Garfield, 1955, 1972), the edge factor measures an aspect of novelty of scientific work – the tendency to build on ideas close to the edge of the scientific frontier. A common characteristic for these two measures is that for each entity they both quantify the average of a characteristic (rather than the total number of novel contributions or the total number of received citations).

We selected countries as the unit of analysis because borders continue to influence scientist interactions and because many important science policy decisions are set at the national level. However, similar to the impact factor, the edge factor too can be constructed also for many other units of analyses. For example, the approach can be utilized to evaluate the novelty of research published by a journal or institution. It can also be applied to analyze the novelty of individual scientists' publications and examine which scientist-level characteristics promote the trying out of new ideas. Furthermore, the edge factor can be utilized to compare to what extent funding agencies succeed in their often-stated aim of supporting novel work. The focus on countries in the present paper is thus just one possible application. But this focus is useful because it provides way to utilize this concept – the edge factor – to shed light on a long-standing question about to what extent the spread of modern communication technologies has eradicated location-based barriers to the adoption of new ideas. At the same time, this application illustrates the potential of the edge factor in other contexts.

We emphasize that the impact factor and the edge factor capture distinct aspects of science – impact and novelty – regardless of the specific setting in which they are employed. Moreover, the edge factor and the impact factor are complementary tools in policy evaluation and design. For optimal science policy requires that both influence and novelty are rewarded. One reason why rewarding influence alone is not enough is that rewarding novelty directly helps solve a coordination problem that is inherent in the formation of a vibrant scientific community to a new area of investigation (Besancenot and Vranceanu, 2015; Packalen and Bhattacharya, 2017). A singular focus on citation counts can lead to stagnant science because impact factors under-reward scientists who try out new ideas, thereby stifling work that helps ideas mature and makes more meaningful advances possible. Another related reason to reward novel work is that useful work that tries out a new idea need not be influential in the traditional sense; such work can have scientific value – in terms of helping unlock the mysteries of the new idea – even when it merely demonstrates which research paths do not work.

In recent decades, many – including the editor of *Science* (Alberts, 2013) – have raised alarm about the science community's obsession with impact factors. The obsession with impact may have already led to less healthy science, as the rise of citation metrics has coincided with a decline in the novelty of biomedicine (Rzhetsky et al., 2015). By using measures like the edge factor in conjunction with impact-based metrics, university administrators and funding agencies can strike a better balance between rewarding innovative but risky work that develops ideas early on and rewarding work that takes advantage of the ideas in their more mature stages. Making reward structures even slightly more favorable to scientific novelty would encourage scientists to pursue more innovative research paths and lead to healthier, less stagnant, science. To be sure, the edge factor is not meant to displace the impact factor. Instead, it is ideally used as a complementary metric that captures a different aspect of science.

As in the closest prior work (Packalen and Bhattacharya, 2017), the focus here is on the novelty of idea inputs, as opposed to the novelty of the combination of idea inputs. Novelty of combinations is a focus in several recent analyses (Wang et al., 2016; Lee et al., 2015;

Rzhetsky et al., 2015; Boudreau et al., 2016; Foster et al., 2015). Both foci come with their advantages, as discussed in Packalen and Bhattacharya (2017). The focus on the use of new ideas makes it possible to include on a larger number of ideas in the analysis than is computationally feasible in an analysis of combinatorial novelty. Analysis of the use of new ideas is also important because the trying out of new ideas is so central to scientific progress. For without new ideas science eventually stagnates – combinatorial novelty alone cannot overcome it.

Similar to the closest prior work (Packalen and Bhattacharya, 2017), we use text analysis to determine the ideas that each paper built upon and also the vintage of those ideas. There are two main differences between the present study and this prior work. First, there is a shift in substantive focus – from ranking journals to ranking nations. Second, in the present approach the novelty of each contribution is allowed to depend not just on the vintage of ideas employed in it but also on what types of ideas they are. This is important because a paper that employs a 10-year old research tool may represent novel work but the same need not be true for a paper that examines a gene of the same vintage. This methodological innovation yields a more finely grained measurement of each entity's distance to the edge of the scientific frontier. Importantly, this methodological advance can be utilized in various applications beyond the present analysis.

2. Methods

In this section, we first describe the data sources (MEDLINE biomedical publications database, NLM Journal Categories, the UMLS Metathesaurus, and the MeSH vocabulary). In subsection 2.2 we explain why the analysis is focused on years 1988–2016, and in subsection 2.3 we explain why the location of each contribution is determined based on the first author's affiliation. In subsection 2.4 we explain how the research area of each contribution is determined.

In subsection 2.5 we explain how we use the UMLS metathesaurus to determine the ideas that each paper built upon and also how the vintage of each idea is determined. In subsection 2.6 we explain how we define a contribution (by determining links from each paper to each research area and each idea category), how we measure the idea category of each idea based on the UMLS metathesaurus, why the resulting contribution-level analysis is preferable to a paper-level analysis, and also how the novelty of each contribution is determined.

In subsection 2.7 we explain how the overall edge factor for each nation is calculated: we first determine the nation's edge factor separately for each (idea category, research area) pair based on all the nation's contributions linked to that pair; afterwards we calculate the nation's overall edge factor as a weighted sum of its edge factors across all (idea category, research area) pairs. Finally, in subsection 2.8 we explain how the approach developed here differs from the approach used in closest prior work.

2.1 Data Sources

2.1.1 MEDLINE Database on Biomedical Research Publications—Our source for information on scientific publications is the MEDLINE database (<https://>

www.nlm.nih.gov/bsd/pmresources.html). MEDLINE is a comprehensive database on life sciences with a focus on biomedicine in particular. The database contains information on over 24 million journal articles.

For each journal article in MEDLINE, we make use the following variables: publication year, affiliation of the first author, text of the title and abstract, journal where the article was published, and MeSH keywords. The acronym “MeSH” stands for *Medical Subject Headings*; the MeSH vocabulary is a controlled vocabulary of over 87,000 terms (<https://www.nlm.nih.gov/mesh/>). Publications in MEDLINE indexed with MeSH keywords; we use the MeSH keywords to determine article type (we focus the analysis on original research articles) and whether an article represents applied or basic science (section 2.4).

2.1.2 Broad Subject Terms for MEDLINE Journals—Our source for the research area of each article is the broad subject terms that are assigned by the National Library of Medicine for journals in the MEDLINE database (<https://wwwcf.nlm.nih.gov/serials/journals/index.cfm>). We show further below how articles in the MEDLINE database are distributed across the journal categories in this database (section 2.4).

2.1.3 Unified Medical Language System (UMLS) Metathesaurus—As our source for information on which words and word sequences represent meaningful concepts in biomedicine and which concepts are synonyms, we use the 2017 version of the Unified Medical Language System (UMLS) metathesaurus (<https://www.nlm.nih.gov/research/umls/>). The UMLS metathesaurus links over 5 million terms that appear in one or more of over 150 medical vocabularies.

In addition to determining the synonyms for each term, the UMLS database assigns each term to one or more of 127 semantic types (<https://semanticnetwork.nlm.nih.gov>). We use the semantic type of each term to represent the idea category of the term (section 2.6.1). Further below we list examples of ideas and idea categories captured by this approach (section 2.6.2).

2.2 Sample of Papers

When we determine the vintage of each idea (section 2.5.2), we use the sample of all papers in the MEDLINE database. By contrast, when we calculate for each location its propensity to publish novel work, we limit the sample of papers in several ways. First, we limit the analysis to original research papers, thereby excluding editorials, reviews, etc. However, in a robustness analysis, we include all papers in the sample. Second, we limit the analysis to papers published during 1988–2016. This is because the coverage for affiliation data in the MEDLINE database begins in 1988. Third, we limit the analysis to papers for which the available text on the title and the abstract of the article in the database includes at least 200 characters and no more than 5000 characters. However, in a robustness analysis, we conduct the analysis without this character limit. The number of articles that are included in our main specification is shown by publication year in Figure S1 (Web Appendix).

2.3 Country of Each Scientific Publication

We assign each paper to a country based on the affiliation string for the first author of the paper. We limit the analysis to first authors because for most papers published before 2014 the affiliation information in MEDLINE is limited to the first author of each paper. Figure S2 (Web Appendix) shows by publication year the share of papers that we were able to match to a country. For ease of exposition we limit the number of locations by combining some countries that publish a smaller number of biomedical publications to regions. Figure S3 (Web Appendix) shows the share of papers by location (country or region) and time period.

2.4 Journal Categories and Journal Category Groups

We use the journal categories (Broad Subject Terms) to represent the research area of each paper. On average, each original research article published during 2015–2016 is linked to 1.49 journal categories. Table S1 (Web Appendix) shows the distribution of links from papers to journal categories during this time period. As is discussed in detail further below (section 2.6.1), papers from journals that are linked to multiple journal categories are considered to have contributed to multiple research areas.

In our main specification, all journal categories are included in the analysis. In secondary analyses, we conduct three separate analyses – each limits the analysis to one of the following three groups of journal categories: “Applied”, “Basic Science”, and “Other (Both Applied and Basic Science)”.

To conduct these secondary analyses, we assign each journal category to one of the three journal category groups. Here we make use of the MeSH keywords affixed to each MEDLINE article and the “A-C-H” model (Weber, 2014) that classifies papers along the translational axis based on the MeSH keywords. Specifically, using the MeSH codes we first determine each paper’s position on the translational axis as specified by the A-C-H model: “H status” (human) is assigned to papers with either the MeSH code B01.050.150.900.649.801. 400.112.400.400 (Human) or the Mesh code M01 (Person), “C status” (cells and molecules) is assigned to papers with any of the following MeSH codes (or codes that appear in the MeSH subtrees of these MeSH codes): A11 (Cells), B02 (Archaea), B03 (Bacteria), B04 (Viruses), G02.111.570 (Molecular Structures), and G02.149 (Chemical Processes), and “A status” (animal) is assigned to papers with the MeSH code B01 (Eukaryota) and papers with any of the codes in the subtree of this MeSH code B01 except the aforementioned MeSH code for “Human”.

We thus construct three separate indicator variables (“H status”, “A status”, “C status”). In the A-C-H model, papers with “H status” have an applied aspect to them, and papers with either “A status” or “C status” have a basic science aspect to them. More than one of these indicator variables will be positive for papers that have both an applied and a basic science aspect to them.

For each journal category, we next calculate the average of each of these three dummy variables (“H status”, “A status”, “C status”) among all papers linked to that journal category. Denoting these variables as “Average H status”, “Average A status”, and “Average

C status”, we use them to classify journal categories to three journal category groups as follows. Journal categories that satisfy conditions “Average H status > Average C status” and “Average H status > 0.2” are assigned to journal category group “Applied”. Journal categories that satisfy “Average H status < Average C status” and “Average A status < 0.8” and “Average C status > 0.5” are assigned to journal category group “Basic Science”. (We thus exclude journal categories that focus heavily on veterinary medicine from this category even though such journal categories are located early along the translational axis in the A-C-H model; this happens in the A-C-H model because the model does not distinguish between veterinary medicine and animal studies as precursor to human medicine). The remaining journal categories are assigned to journal group category “Other (Both Applied and Basic Science)”. The result of this approach for determining the journal category group of each journal category is shown in the last column of Table S1.

2.5 Identifying Ideas and the Vintage of Ideas

2.5.1 Using the UMLS metathesaurus to identify ideas from text—We employ text analysis to discern which ideas each research paper built upon. We treat each of the 5+ million terms in the comprehensive United Medical Language System (UMLS) metathesaurus as representing ideas. To identify which of these ideas each research paper in the MEDLINE database built upon, we search the title and abstract of each publication for all the terms in the UMLS metathesaurus.

Thus, the first step in the text analysis is to determine for each article in the MEDLINE database which UMLS terms appear in it. Further below we also show a list of examples of ideas identified by this approach (section 2.6.2).

2.5.2 Calculating the vintage of each idea—The vintage of the idea represented by a UMLS term is determined based on how long ago the UMLS term was first mentioned in a biomedical research paper. We interpret the mention of a relatively new term as indicative of work that builds on ideas close to the edge of the scientific frontier. We refer to the year of first appearance of a term as the cohort year of the term. In a robustness analysis, we set the cohort year of each term as the earliest year the UMLS term or any of its synonyms appears in the MEDLINE data (synonyms are determined based on the synonym information in the UMLS metathesaurus). The results from this robustness analysis show that our results are not driven by relabeling of old ideas.

Because of the sparsity of publications in MEDLINE with a publication year before 1946, the cohort year of ideas (i.e. the year of first appearance) does not reflect the ideas’ true vintage well for ideas that are new to biomedicine before 1950. Thus, we exclude from the analysis all terms with cohort before 1950.

Further below we show examples of cohort years assigned to terms using this approach (section 2.6.2).

2.6 Contribution-Level Analysis

2.6.1 Defining a contribution as a link from a paper to an (idea category, research area) pair—In determining the novelty of biomedical work, we seek to control

for the idea category of each idea. Thus, we aim to compare the use of novel ideas against the use of more established ideas from the same idea category. The rationale for seeking to control for the idea category is the following: how recent ideas should be considered novel depends on what type of an idea it is. For example, a paper that employs a 10-year old research tool may represent novel work but the same need not be true for a paper that examines a gene of the same vintage.

To control for the idea category in the present analysis, we take advantage of the fact that the UMLS metathesaurus classifies terms to 127 semantic types (these are listed further below in section 2.6.2). We treat each of these UMLS semantic types as representing one idea category.

We make use of these idea categories as follows. After determining which UMLS terms are mentioned in each paper, we determine which UMLS categories are represented by these terms. We then treat a paper that mentions terms from K different idea categories as K separate contributions. The underlying assumption in this approach is that work that mentions at least one idea from an idea category advances our understanding of how ideas from that idea category work. Thus, work that mentions ideas from multiple categories advances our understanding on multiple dimensions; hence we treat it as multiple contributions.

Table S2 (Web Appendix) shows the number of links to each idea category from papers published during 2015–2016. As was mentioned above in section 2.5.2, we only include in the analysis those terms that have cohort year 1950 or later.

In our main analysis, we calculate the overall edge factor based on links to any of the 127 idea categories (the edge factor is an average across all idea categories). In a secondary analysis, we calculate the edge factor separately for each of the following four groupings of idea categories: “Clinical and Anatomy”, “Drugs and Chemicals”, “Basic Science and Research Tools”, and “Miscellaneous”. To conduct this analysis, we link each idea category to one of these four idea category groups. The last column of Table S2 shows which idea category belongs to which idea category group. This secondary analysis reveals for each country whether any barriers to new idea adoption are limited by to certain types of ideas. The decision to link each of the 127 idea categories to one of four groups was made for expositional purposes – it implies that for each country we must report four separate numbers.

What kind of work should be considered novel is likely to depend also on the research area. For example, use of a 10-year old research tool may be novel work in public health research but not in biotechnology research. To address this issue, we also determine the links from papers to research areas. We use the National Library of Medicine (NLM) journal categories as proxies for research areas (these journal categories were listed in section 2.4, Table S1).

Thus, after determining the ideas mentioned in each paper, we determine which idea categories are linked to these ideas as well as which research areas are linked to the journal where the paper is published. We define a *contribution* as an (idea category, research area) pair linked to a paper.

In our approach, a paper is considered to contribute to our understanding of all the (idea category, research area) pairs linked to it. A paper can make multiple contributions, depending on how many (idea category, research area) pairs are linked to it. A paper that mentions ideas from K idea categories, and is published in a journal that is linked to J journal categories, is treated as $K \times J$ separate contributions.

Note that a paper that mentions multiple ideas from an idea category results in the same number contributions as a paper that mentions only one idea from the idea category.

The number of links listed in the second column of Table S1 is the number of links to (idea category, research area) pairs associated with each research area. Similarly, the number of links listed in the second column of Table S2 is the number of links to (idea category, research area) pairs associated with each idea category. On average, each paper published during 2015–2016 is linked to 6.26 (idea category, research area) pairs. Therefore, in our approach each paper is, on average, counted as 6.26 contributions.

A limitation of the approach pursued here is the inherent assumption that, synonyms aside, all ideas within an idea category are treated as though they are equally close to one another although in reality some are closer than others. Similarly, all ideas within a category are treated as though they are equally distance is the same to all ideas outside the idea category. A potentially important direction for future work is to explore a more fine-grained approach that either calculates a pairwise between each idea category or calculates a pairwise distance between each idea.

When determining whether a contribution represents novel work, we only consider the age of the newest term linked to the (idea category, research area) pair from the paper in question. Researcher's choice is between using any new ideas or only well-established ideas from this idea category. This is discussed in more detail next.

2.6.2 The novelty of a contribution—Above we defined a contribution as a link from a paper to an (idea category, research area) pair; these links are inferred from the UMLS terms that appear in the title and abstract of the paper. The novelty of each contribution is determined in three steps.

Step 1. Age of each UMLS term that links a paper to the (idea category, research area) pair. First, for each contribution associated with a paper, we determine the age of each term that links the paper to the (idea category, research area) pair in question. Age of each term is calculated by subtracting the cohort year of the term from the publication year of the paper.

Step 2. Age of the newest UMLS term that links a paper to the (idea category, research area) pair. Second, for each contribution we determine the age of the newest term that links the paper to the (idea category, research area) pair. We refer to the cohort year of the newest term that links a paper to the (idea category, research area) pair as the cohort year of the contribution.

Step 3. Novelty of the contribution relative to other contributions to the (idea category, research area) pair among papers published in the same year. The relative novelty of a

contribution is then determined by comparing the vintage of the contribution to the vintages of all the other contributions linked to the same (idea category, research area) pair, among papers published in the same year. The interpretation is that a paper that links to an idea category reflects a choice faced by a scientist: one can choose work with at least one relatively new idea from this idea category, or one can choose to work with only well-established ideas from this idea category. The comparison is also limited by research area because whether the use of an idea represents novel work is expected to depend on the context where it is used. The reason for limiting the comparison to papers published in the same year is obvious: because the rate of scientific progress need not be the same over time, the use of a 10-year old research tool may represent novel work in one year but not in some other year.

Having determined all contributions linked to an (idea category, research area) pair among papers published in the same year, we order the contributions based on their vintage (age of the newest term linked to that (idea category, research area) pair from each paper). We then construct an indicator variable that captures the relative novelty of each contribution: in our baseline specification, contributions that are in the top 5% based on their vintage are considered novel work (the indicator variable is 1 for such contributions and 0 otherwise). In robustness analyses, we construct the indicator variable using alternative choices for the cutoff percentile (top 1%, top 5%, or top 20%). Figure S4 (Web Appendix) shows the distribution of the cohort of *all* contributions based on papers published during 2015–2016. Figure S5 (Web Appendix) in turn shows the distribution of cohort of *novel* contributions when novelty of a contribution is determined based on the top 5% status and also the when novelty of a contribution is determined based on one of the alternative cutoffs (top 20%, top 10%, or top 1%). When the top 5% cutoff is used, then for all post-2004 cohorts the majority of contributions with that cohort are deemed novel by our approach, and for all pre-2004 cohorts at most a minority of contributions are deemed novel by our approach.

Table S3 (Web Appendix) shows examples of ideas, as represented by UMLS terms, captured by our approach. The table also shows the idea category of each term. Some terms appear multiple times because these terms are linked to multiple UMLS categories by the UMLS metathesaurus. As in related prior work (15), the list of terms shows that the approach used here captures ideas that are widely recognized to have been important inputs in biomedical work in recent decades (for expositional reasons the list is focused on popular ideas – there are of course also many unpopular, less important ideas that are captured by our approach) and that for most terms the cohort year assigned to the term reflects the era when the idea represented by the term entered biomedicine.

2.7 Calculation of the Edge Factor

2.7.1 Novelty of a nation's contributions linked to a specific each (idea category, research area) pair—Normalization of contribution-level novelty indicators. Having determined the contributions of each paper (i.e. which (idea category, research area) pairs are linked to from each paper) and which contributions are novel (i.e. which contributions have the top 5% status based on their vintage), we next normalize the novelty variable within contributions to each (idea category, research area) pair so that the average of

the normalized novelty variable is 100 within each (idea category, research area) pair. In implementing the normalization, we combine data from multiple years. For example, in our main specification we combine data from 2015–2016.

Location-level novelty scores for each (idea category, research area) pair. Using the normalized contribution-level novelty variable, we then calculate for each location its propensity for novel work within each (idea category, research area) pair. That is, for each location we calculate the mean of the normalized novelty variable based on all of the location's contributions linked to a specific (idea category, research area) pair. We refer to each such average of the novelty variables as the edge factor of the location for the specific (idea category, research area) pair. An edge factor above (below) 100 indicates an above (below) average tendency for work that builds on relatively novel ideas. In our main specification, these edge factors are calculated based on papers published during 2015–2016.

2.7.2 Calculating the overall edge factor—Having determined the relative novelty of each location's contributions separately for each (idea category, research area) pair – the location's edge factor for that (idea category, research area) pair – we construct the overall edge factor for each nation as a weighted sum of these (idea category, research area) pair specific edge scores.

Weights. In our main specification, we use as weights the frequency at which each (idea category, research area) pair is encountered in biomedicine. In other words, the weight of an edge factor for an (idea category, research area) pair is the total number of papers linked to it from any location during the time period.

A justification for selecting these weights is that those (idea category, research area) pairs that are encountered more often in biomedicine are, by revealed preference, considered more important by scientists. The ability to pursue cutting-edge work in an often-encountered (idea category, research area) pair is thus arguably more valuable than is the ability to pursue cutting-edge work in a rarely encountered (idea category, research area) pair. The implicit assumption in this approach is that, even though it is not yet known which (idea category, research area) pairs will be the most important sources of future progress in biomedicine, the past is the best predictor of the future.

Because the overall scientific frontier position for a nation (the edge factor) is calculated as a weighted sum over its position across all (idea category, research area) pairs, the resulting measure for the nation reflects its overall capability across all of biomedicine, as opposed to only the nation's capabilities in areas where it has concentrated most of its own activities. Accordingly, the edge factor is high only if the country has significant capabilities across different areas of biomedicine; expertise in a narrow subset of biomedicine is not enough.

However, in a secondary analysis we show that the results are robust to the case when the edge factor is calculated using as weights each country's own number of papers that link to a given (idea category, research area) pair. Hence, the results from this alternative specification reflect the novelty of the work actually pursued by the nation – emphasizing more the

novelty of the nation's work in areas where it publishes a lot – rather than the nation's capabilities across all of biomedicine.

Cells with missing observations. Not all locations have publications linked to every (idea category, research area) pair. The results are mainly insensitive to how such missing cells are handled. One reason for this is that while there are 13,394 (idea category, research area) pairs with positive weights, the 3,000 largest (idea category, research area) pairs as measured by their weight account for the vast majority (82%) of the total weight with most of them large enough for also the small and mid-sized nations to be active in them. In our main specification, we handle the cells with missing observations by replacing the nation's edge factor for that cell with the nation's weighted average across the other cells (those (idea category, research area) pairs for which the location does have publications linked to it). The weights used in this calculation are the same weights as discussed above. In an alternative specification, we replace cells with missing edge scores with 0 (the worst possible edge score). In another alternative specification, we replace cells with missing edge scores with 100 (by definition the average novelty score for every (idea category, research area) pair). In both cases the results are similar to the results for the main specification (see section 3.2).

2.8 Comparison of the Approach with the Approach Used in Closest Prior Work

The approach employed here has two main differences with the closest prior work (Packalen and Bhattacharya, 2017). In addition to the shift in substantive focus – from ranking journals to ranking nations – the present analysis is conducted at the contribution-level, with contribution defined as a link from a paper to an (idea category, research area) pair. By contrast, in this prior work, the analysis was conducted at the paper-level. That is, here the novelty score for an entity is calculated first at the contribution-level separately for each (idea category, research area) pair and the overall novelty score for the entity is then calculated as a weighted sum across these each (idea category, research area) pairs. By contrast, in the prior work (Packalen and Bhattacharya, 2017) the novelty score for an entity was calculated at the paper-level either without controlling for either the idea category or the research area, or by only controlling for the research in a manner that essentially uses as weights the entity's own involvement in the research area (in this prior work the research area was determined based on the appearance of 6-digit MeSH terms; the entity of interest in this prior work was a journal, here it is a nation).

The advantage of the approach pursued in the present analysis is thus not only that the present approach controls for the idea category but also that the present approach uses as weights the (idea category, research area) pair's overall importance in biomedicine (as measured by the total number of contributions linked to it). This yields a better reflection of an entity's capabilities in biomedicine compared to the case when the weights represent the distribution of the entity's own involvement across different areas of biomedicine.

3. Results

3.1 Main Results

Figure 1 shows the edge factor for each nation based on papers published during 2015–2016. The edge factor is normalized so that the average edge factor across all contributions is 100. An edge factor of 110 for a nation indicates that on average the nation's contributions build on relatively new ideas 10% more often compared to the average contribution in the same research area. Markers drawn in red (blue) indicate edge factors that are well above (well below) average. Markers drawn in gray indicate edge factors that are approximately average.

The results show that the United States and South Korea have the most advanced positions on the scientific frontier: scientists working in these nations build on cutting-edge ideas more often than do scientists in other locations. The propensity for novel science is well above average also in Singapore and Taiwan. Countries that come after these four countries have approximately average propensity for novel science. Such countries include China, Canada, most western European countries (including the United Kingdom and Germany), Australia, and South Africa. Other countries (including Turkey, India, Brazil, and Iran) come further behind – scientists in these countries have clearly below average propensities for novel work.

Confidence intervals and results for alternative specifications (shown in Table S4 (Web Appendix) and discussed in detail below in subsection 3.2) indicate that in most cases these results are robust. The one exception is Saudi Arabia, for which results from alternative specifications suggest a below average tendency for novel work.

Countries examined here thus have quite different propensities for work with newer ideas in biomedicine. This indicates that location continues to exert considerable influence on what kind of science is pursued. Furthermore, even developed nations are not on an equal footing in the pursuit of novel scientific work: in some developed nations scientists take advantage of opportunities created by the arrival of new ideas much more often than do scientists in other nations.

Figure 2 shows the change in the edge factor for each nation from the 1990s to present. South Korea, Taiwan, and China have leapfrogged most developed nations. Whereas the United States is still among the leaders, the relative positions of Switzerland and the United Kingdom are less advanced now than they were in the 1990s. Overall some convergence appears to have taken place as the lagging nations are no longer as far behind the leaders. This suggests that the world of ideas may have become somewhat flatter. Analysis of the edge factor by 5-year time periods (shown in Table S5 (Web Appendix)) indicates that most changes that occur are persistent. The changes thus reflect systematic changes in capabilities rather than merely year-to-year random variations.

In our approach, we compare each contribution only to other contributions that use ideas from the same idea category and are linked to the same research area (the 127 idea categories include “Amino Acid, Peptide, or Protein” and “Pharmacologic Substance”; the 125 research areas include “Biochemistry” and “Neoplasms”; see Table S1 and Table S2 for

the full lists). Table 1 shows the edge factor separately for four groupings of idea categories: “Clinical and Anatomy”, “Drugs and Chemicals”, “Basic Science and Research Tools”, and “Miscellaneous”, and for three groupings of research areas: “Applied”, “Basic Science”, and “Other (Both Applied and Basic Science)”.

For most nations, the edge factor is similar across these groupings, suggesting that the pursuit of novel work is generally dependent on capabilities that some countries possess but others lack. One important exception is China. China’s contributions linked to the idea category grouping “Basic Science and Research Tools” now have the second highest propensity for novel work (after Singapore), but its contributions linked to idea category groupings “Clinical and Anatomy” and “Drugs and Chemicals” are well below average in terms of their novelty. This result serves to highlight an important feature of our approach: it can be used to reveal not just whether a nation is facing barriers in new idea adoption but where in the idea space those barriers lie.

While in Table 1 we divided papers to just three groups based on their research area (i.e. basic, applied, and other), it is important to note that the restriction to just three bins on this dimension was made for expositional purposes. The edge factors can also be reported separately for each of the 125 research areas (listed in Table S2). For example, funding agencies could use such field-level analyses of novelty across to determine in which scientific fields their own country is closest to the frontier and use this information (in conjunction with other relevant metrics) when deciding where to allocate limited research dollars.

3.2 Confidence Intervals and Results from Alternative Specifications

Table S4 (Web Appendix) shows the confidence intervals and results from a variety of alternative specifications. For ease of comparison, the results from the main specification are reported again in column (1d).

Confidence intervals reported column (1e) are constructed using a bootstrap method. We first generate each of 1000 artificial samples by re-sampling with replacement from the (idea category, research area) pairs until the total weighted number of observations (i.e. contributions) in each constructed sample is at least as large as the total weighted number of observations is in the original sample. Next, we calculate the edge factor for each nation in each constructed artificial sample. We then eliminate the largest 2.5% and the smallest 2.5% of the values in the edge factor distribution for each nation among these constructed bootstrapped samples. The extremes of remaining edge factor values form the 95% confidence interval for the edge factor of each nation.

The calculated confidence intervals indicate that scientists in the four top nations are clearly above average in their propensity to use new ideas, that scientists in most developed nations have approximately an average propensity to use new ideas, and that scientists in developing nations have a below average propensity to use new ideas.

The analysis reported in column (2) differs from the main specification in terms of how those (idea category, research area) pairs are treated for which a nation has no contributions

linked to it: now the edge factor for such (idea category, research area) pairs are replaced with 0, reflecting the most pessimistic scenario about the nation's capabilities for that (idea category, research area) pair. By contrast, in the main specification these missing observations are replaced with the average edge score for the nation for (idea category, research area) pairs for which the nation does have observations. Comparison of the main results (column 1c) against the results in column (2) shows that while the edge factor decreases somewhat for the smaller nations (as expected), the results remain qualitatively unchanged. The fact that regardless of the approach the nations near the top of the rankings included also smaller and medium sized nations (in terms of their scientific output, as indicated by Table S3), and the fact that some larger countries are far down in the rankings (most notably India), demonstrate that the results are not driven solely by the size of a country's scientific workforce.

The analysis reported in column (3) differs from the main specification in how the weights for the edge factor for each (idea category, research area pair) are calculated. Here, for each country the weight for an (idea category, research area) pair is the country's own total number of research publications linked to the same (idea category, research area) pair. Thus, the overall edge factor is the same as the average of the nation's novelty scores across all of its contributions. By contrast, in the main specification weight for each (idea category, research area) pair is the same for all nations: it is the total number of papers linked to that (idea category, research area) pair. Comparison of the main results (column 1) against the results reported in column (3) shows that the results are robust to this alternative specification.

The analyses reported in columns (4–6) differ from the main specification in that the dummy variable indicating novelty of a contribution is now constructed using top 20%, top 10% and top 1% cutoffs. By contrast, in the main specification this dummy variable is constructed using the top 5% cutoff. Comparison of the main results (column 1) against results reported in columns (4–6) indicates that while the main results are qualitatively robust – leaders do better than laggards regardless of the measure – the relative position of the United States improves as one moves to a narrower cutoff (from 5% to 1%) and China's relative position improves when one moves to a wider cutoff (from 5% to 10% and 10% to 20%). A possible explanation is that countries may differ in terms of how many of their institutions are on the very edge of the frontier ("the bleeding edge"), so that some countries fare better when novelty is calculated based on a narrower measure. For example, the U.S. may have many of the very top institutions in the world (in terms of their tendency to work with new ideas) but most of its institutions may be further down in the pack. In another country, such as China, institutions may be more homogenous in terms of the scientists' tendency to work with new ideas. The differences may also be driven by variation in where the new ideas are first born (the United States may be disproportionately the origin of new ideas – and thus receive a disproportionate share of the very first mentions of new terms – but scientists working in China may be relatively more eager to build on the new ideas).

The analysis reported in column (7) differs from the main specification in that now the cohort of each UMLS term is the year of the earliest mention of that term or any of its synonyms, with synonyms specified by the UMLS. In contrast, in the main specification the

cohort year is the year of the earliest mention of the term itself. Comparison of the main results (column 1) against the results reported in column (7) shows that the conclusions from the main specification are robust in this way as well.

The analysis reported in column (8) differs from the main specification in that the analysis now includes all publications in MEDLINE as opposed to only regular research articles. The analysis reported in column (9) in turn differs from the main specification in that the analysis now includes also publications for which the text information on the title and abstract is less than 200 characters or more than 5000 characters – in the main specification such publications were excluded from the analysis. Comparison of the main results (column 1) against the results reported in columns (8) and (9) show that the results are robust also to these alternative specifications.

4. Discussion

While our results show that differences persist even among developed nations in their propensity to work with new ideas, the results do not reveal the specific mechanisms driving these differences. One potential driver of these cross-locational differences stems from the difficulty of working with new ideas. Because novel science is harder than conventional science, novel science is more dependent on interactions with colleagues. The fertility of these scientist interactions depends on factors such as the extent of complementary tacit knowledge that is embedded in people and is transferred to others in meetings (Lucas 2004; Lucas and Moll, 2014). Cross-national variation in the extent and depth of human capital investments can thus lead to cross-national variation in the tendency to adopt new ideas.

Of course, not all fruitful interactions are limited by location, as is evidenced by the fact that a quarter of science now involves international collaborations (Freeman, 2013; National Science Board, 2016). However, the rise of long-distance collaborations can also be a source of cross-national differences in new idea adoption: a nation can gain an advantage if its scientists can form distant collaborations relatively easily. In this regard, China's special relationship with the United States in science (Freeman and Huang, 2015) has likely helped propel it to the scientific frontier. An important topic for future work is to explore to what extent Chinese scientists working at the frontier started their work in the U.S. This link would potentially have major implications for other nations that are seeking to advance their position relative to the scientific frontier. A related topic worthy of future exploration is quantifying to what extent national borders still limit collaboration opportunities and what are the implications of collaboration barriers for each country in terms of the novelty of its scientific output.

Willingness to try out new ideas can vary by location also due to differences in scientist demographics. For example, given that early-career scientists are the most likely to work with new ideas (Packalen and Bhattacharya, 2016), and given that the increase in the extent of science in China is so recent and thus many of its scientists are early on their careers, the novelty of science in China may be driven in part by the youth of its scientists. Cross-national differences in new idea adoption and China's remarkably ability to leapfrog in this regard may also be driven in part by differences in incentives to pursue novel work: it has

long been understood that nations without vested interests in existing technologies have an elevated incentive to explore new ideas (Brezis et al., 1993; Mokyr, 1994). Some of the variation in new idea adoption can also be driven by variation in where the ideas are first born, and by remaining delays in the spread of awareness about which new ideas exist.

Our results are consistent with findings from Hidalgo and Hausman (2009) which measured the complexity of each country's production structure based on its exports and found large differences in the capabilities of nations. Their analysis was motivated by the idea that each nation's capabilities determine the input varieties that it can fruitfully use in production. Our work, by contrast, is motivated by the idea that capabilities determine whether a nation's scientists can take advantage of the opportunities created by the arrival of new ideas. Moreover, whereas in this related work the complexity of goods production is measured indirectly based on exports, the edge factor is calculated directly based on the measured idea inputs. Common to these analyses is the belief that the capabilities of a nation affect which inputs it uses and both analyses are aimed at constructing new measures that reflect those capabilities.

Our finding that nations continue to differ in their ability to pursue novel science is in line with cross-country comparisons of scientific impact as measured by citations (Freeman, 2013; National Science Board, 2016). The ability to take advantage of scientific opportunities continues to vary across locations in spite of the "death of distance" phenomenon, because locational differences in capabilities persist as shown by Jones et al. (2008), Agrawal and Goldfarb (2008), Ding et al. (2010) and Packalen and Bhattacharya (2015).

But some aspects of our results also differ from the results obtained through traditional analyses of scientific productivity. Data on the tendency to produce highly cited papers point to the United States as a leader that remains far ahead of most western European nations and even further ahead of South Korea, Taiwan and China (Freeman, 2013; National Science Board, 2016; Freeman and Huang, 2015; Bornmann et al., 2017; Yu et al., 2014). Our analysis on the use of new ideas, by contrast, suggests that South Korea, Taiwan and China have caught up with western Europe and are now close to the United States in terms of their tendency to work with cutting-edge ideas. Moreover, we find that China is now a leader in favoring newer ideas when working with new basic science ideas and research tools.

The finding that some countries are among the leaders in terms of their edge factor but lag in terms of their impact is not surprising. Prior studies have found novelty and impact to correlate only imperfectly (Packalen and Bhattacharya, 2017; Wang et al., 2016; Lee et al., 2015). Moreover, work on an idea early – when the idea is still raw – may well have less impact than work that builds on more established ideas which properties are better understood. The early work on the idea is still crucial: it helps the idea develop and thus makes more significant advances possible. Moreover, countries investing heavily in novel science can reap significant benefits also for themselves from their focus: early work on an idea can help the country develop capabilities that enable it to take advantage of the later, more fertile, opportunities linked to the same idea.

5. Conclusion

Our analysis has shown that countries continue to differ in their ability to take advantage of cutting-edge ideas. Even across developed nations, sizable differences persist in terms of the tendency at which each nation's scientists build on recent advances. Hence, in spite of the arrival of modern communication technologies – which now facilitate almost instantaneous access to each new idea from almost anywhere in the world – the world of ideas is not yet flat. A likely explanation for this is that access to an idea does not guarantee that a scientist can take advantage of it in a productive way soon after its initial discovery. Instead, because new ideas are often raw and poorly understood, the ability to fruitfully build upon an idea depends on local factors such as daily interactions with colleagues, the training environment, and ready access to potential collaborators. Because these factors vary across locations but are beneficial or even necessary in helping scientists unlock the mysteries of new ideas, geographic differences in the tendency to exploit new ideas continue to persist.

Currently, the tendency to build on new ideas in biomedicine is highest in United States and South Korea. While China has leapfrogged most developed nations in terms of its overall tendency to work with new ideas, this progress has been uneven across idea types. In terms of applying basic science ideas, China has already caught up with the leaders, but in terms of applying new clinical knowledge, China remains below average. This result highlights an important benefit of the approach developed here: the approach can be used to show not only whether differences in new idea adoption persist, but also where in the idea space any remaining barriers lie.

An important direction for future work is to examine what are the best approaches for overcoming barriers in idea adoption. For example, it would be useful to uncover whether the scientific frontier is best reached through an approach where resources are first directed to a handful of fields or universities or through a more diversified approach, and whether smaller countries can elect to specialize in narrow areas or whether the unpredictability of where in the ideas space important future advances come from render it necessary for even smaller countries to develop and maintain broad capabilities so that they can take early advantage of new advances no matter where in the idea space those yet unforeseen new advances are born.

Extending the analysis to patent data would also be useful as it would facilitate exploration of links between scientific frontier positions and technological frontier positions. Such analysis could reveal to what extent funding scientists working near the edge of the scientific frontier is a pre-condition for a country to obtain the capability to pursue inventions near the edge of the technological frontier – inventions that build on recent scientific and technological advances.

In analyses of science, the metric introduced here – the edge factor – holds considerable potential beyond cross-national comparisons. Ever since Garfield (1955, 1972), the focus in empirical analyses of knowledge production has been on measuring influence. This focus on impact in science policy has recently been decried by many, including Alberts (2013) and Osterloh and Frey (2015). For the theory of knowledge production implies that science

policy decisions should be guided by not just the influence of scientific work but also what kind of science is being pursued – novel or conventional (Kuhn 1962; Besancenot and Vranceanu 2015). Because new ideas are raw when they are first born, they need the attention and revision by many scientists to mature into useful advances. But such work is risky, and without explicit incentives for novel work, too many choose to pursue well-trodden research paths in areas where many other scientists also work and thus the prospects of receiving citations are better. A potential consequence of excessive focus on impact is thus that science becomes stagnant, a phenomenon which appears to have already occurred in biomedicine (Rzhetsky et al., 2015).

Of course, given the absence of metrics that capture novelty, the obsession with impact has been inevitable. The edge factor is a valuable tool for this very reason. By providing an approach that can be used to measure the novelty of scientific work of a nation, a journal, an institution, or even an individual scientist, the edge factor allows university administrators and funding agencies to re-structure their reward systems so that scientists are rewarded based on not just the impact of their work but also the novelty of their work. When scientists are rewarded based on *both* the impact and novelty of their work, more scientists can be expected to pursue novel research paths, leading to healthier, less stagnant science.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements:

I thank Jay Bhattacharya, Bruce Weinberg, Partha Bhattacharyya, Richard Freeman, Horatiu Rus, Joel Blit, David Autor, Larry Smith and Peter Tu for discussions. I acknowledge financial support from the National Institute on Aging grant P01-AG039347.

References

- Agrawal Ajay and Goldfarb Avi, 2008, Restructuring Research: Communication Costs and the Democratization of University Innovation, *American Economic Review* 98, 1578–1590.
- Alberts Bruce, 2013, Impact Factor Distortions, *Science* 340, 6134.
- Besancenot Damien, and Vranceanu Radu, 2015, Fear of Novelty: A Model of Strategic Discovery with Strategic Uncertainty, *Economic Inquiry* 53, 1132–1139.
- Bornmann Lutz, Wagner Caroline, and Leydesdorff Loet, 2017, The geography of references in elite articles: What countries contribute to the archives of knowledge, Unpublished manuscript available at <https://arxiv.org/pdf/1709.06479.pdf>.
- Boudreau Kevin J., Guinan Eva C., Lakhari Karim R. and Riedl Christoph, 2016, Looking Across and Looking Beyond the Knowledge Frontier: Intellectual Distance, Novelty, and Resource Allocation in Science, *Management Science* 62, 2765–2783. [PubMed: 27746512]
- Brezis Elise S., Krugman Paul R. and Tsiddon Daniel, 1993, Leapfrogging in International Competition: A Theory of Cycles in National Technological Leadership, *American Economic Review* 83, 1211–1219.
- Ding Waverly, Levin Sharon, Stephan Paula and Winkler Anne, 2010, The Impact of Information Technology on Academic Scientists' Productivity and Collaboration Patterns, *Management Science* 56, 1439–1461.
- Foster Jacob G. and Rzhetsky Andrey and Evans James A., 2015, Tradition and Innovation in Scientists' Research Strategies, *American Sociological Review* 80, 875–908.

- Freeman Richard B., 2013, "One ring to rule them all?" Globalization of knowledge and Knowledge Creation. *Nordic Economic Policy Review* 1, 11–44.
- Freeman Richard B. and Huang Wei, 2015, China's "Great Leap Forward" in Science and Engineering; in Geuna Aldo, ed., *Global Mobility of Research Scientists: The Economics of Who Goes Where and Why*; Elsevier.
- Garfield Eugene, 1955, Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas, *Science* 15, 108–111.
- Garfield Eugene, 1972, Citation Analyses as a Tool in Journal Evaluation, *Science* 178, 471–478. [PubMed: 5079701]
- Hidalgo Cesar E. and Hausmann Ricardo, 2009, The Building Blocks of Economic Complexity, *Proceedings of the National Academy of Sciences* 106, 10570–10575.
- Jones Benjamin F., 2010, Age and Great Invention, *Review of Economics and Statistics* 92, 1–14.
- Jones Benjamin F., Wuchty Stefan and Uzzi Brian, 2008, Multi-University Research Teams: Shifting Impact, Geography, and Stratification in Science, *Science* 322, 1259–1262. [PubMed: 18845711]
- Lee You-Na, Walsh John P. and Wang Jian, 2015, Creativity in scientific teams: Unpacking novelty and impact, *Research Policy* 44, 684–697.
- Lucas Robert E. Jr., 2004, *Lectures on Economic Growth*. Harvard Univ. Press, Cambridge, MA.
- Lucas Robert E. Jr. and Moll Benjamin, 2014, Knowledge Growth and Allocation of Time, *Journal of Political Economy* 122, 1–55.
- Kuhn Thomas S., 1962, *The Structure of Scientific Revolutions*. Chicago University Press, Chicago.
- Kuhn Thomas S., 1977, Objectivity, Value Judgment and Theory Choice; in Kuhn Thomas S., ed., *The Essential Tension*, University of Chicago Press, Chicago, pp. 320–339.
- Marshall Albert, 1920) *Principles of Economics*. Macmillan and Co., London.
- Mokyr Joel, 1994, Cardwell's Law and the Political Economy of Technological Progress, *Research Policy* 23, 561–574.
- National Science Board, 2016, *Science and Engineering Indicators*.
- Osterloh Margit and Frey Bruno, S., 2015, Ranking Games, *Evaluation Review* 32, 102–129.
- Packalen Mikko and Bhattacharya Jay, 2015, *Cities and Ideas*, National Bureau of Economic Research working paper No. 20921.
- Packalen Mikko and Bhattacharya Jay, 2017, Neophilia Ranking of Scientific Journals, *Scientometrics* 110, 43–64. [PubMed: 28713181]
- Packalen Mikko and Bhattacharya Jay, forthcoming, Age and the Trying Out of New Ideas, *Journal of Human Capital*.
- Rzhetsky Andrey, Foster Jacob G., Foster Ian T., and Evans James A., 2015, Choosing experiments to accelerate collective discovery, *Proceedings of the National Academy of Sciences* 112, 14569–74.
- Usher Abbott P., 1929, *A History of Mechanical Inventions*. McGraw-Hill, New York..
- Wang Jian, Veugelers Reinhilde and Stephan Paula, 2016, Bias Against Novelty in Science: A Cautionary Tale for Users of Bibliometric Indicators, National Bureau of Economic Research Working Paper No. 22180.
- Weber Griffin M., 2013, Identifying Translational Science Within the Triangle of Biomedicine, *Journal of Translational Medicine* 11, e126.
- Yu Xie, Zhang Chunni and Lai Qing, 2014, China's rise as a major contributor to science and technology, *Proceedings of the National Academy of Sciences*, 11, 9437–9442.

Scientific Frontier Position of Nations

Figure 1.

Overall Scientific Frontier Position by Location. Edge factors are calculated using text analysis and data on biomedical research papers published during 2015–2016. Scatter points are colored to indicate edge factors that are well above average (red), about average (grey), and well below average (blue). An edge factor above 100 indicates an above average tendency for work that builds on relatively new ideas (a contribution is considered novel if it is in the top 5% by the age of the newest idea it builds upon; the comparison group for each contribution is all other papers published in the same year and linked to the same (idea category, research area) pair).

Figure 2.

Change in Scientific Frontier Position by Location from 1990s to present. Edge factors are calculated using text analysis and data on biomedical research papers published during 1990–1999 and 2015–2016. Smaller scatter points indicate the edge factors for 1990s, larger points for 2015–6. Red (blue) arrows indicate edge factors that increased (decreased) from 1990s to 2015–2016. An edge factor above 100 indicates an above average tendency for work that builds on relatively new ideas (a contribution is considered novel if it is in the top 5% by the age of the newest idea it builds upon; the comparison group for each contribution is all other papers published in the same year and linked to the same (idea category, research area) pair).

Table 1.

Edge Factors by Idea Category Type and by Research Area Type.

(1a)	(1b)	(1c)		(2a)	(2b)	(2c)	(2d)		(3a)	(3b)	(3c)
Location	Number of Contributions	2015–6		Clinical and Anatomy	Drugs and Chemicals	Basic Science and Research Tools	Miscellaneous		Applied	Basic Science	Other (Both Applied and Basic Science)
UNITED STATES	2853661	108		105	121	110	105		106	108	108
SOUTH KOREA	374227	107		111	103	105	105		103	109	106
SINGAPORE	52541	105		109	106	115	108		108	110	112
TAIWAN	177229	104		99	100	105	105		100	101	107
IRELAND	39495	103		107	88	108	98		99	101	108
BELGIUM	95644	102		109	120	99	98		103	106	102
ITALY	384029	102		105	117	94	101		99	103	103
CHINA	1734035	101		95	88	113	101		102	100	103
CANADA	375846	101		99	94	102	105		101	99	105
JAPAN	554589	100		104	106	103	92		94	104	101
UNITED KINGDOM	494917	100		100	105	100	100		98	102	100
NETHERLANDS	233631	100		106	87	100	97		95	103	100
GERMANY	539888	100		95	112	104	97		96	102	100
SWITZERLAND	123779	100		97	118	105	93		94	102	102
SAUDI ARABIA	34855	99		96	84	91	96		90	92	98
FINLAND	59534	99		96	78	106	96		89	99	102
NORWAY	63699	98		99	88	103	98		97	97	104
SOUTH AFRICA	43179	98		107	81	76	98		100	92	
SPAIN	278504	98		98	96	96	99		92	99	101
CZECH REPUBLIC	44024	97		97	86	95	92		95	97	89
AUSTRALIA	320955	97		99	94	95	97		96	95	100
SWEDEN	138949	96		94	102	97	93		91	98	96
AUSTRIA	65039	96		94	103	100	94		91	100	96
DENMARK	105066	95		98	91	96	96		92	95	101
FRANCE	305065	95		96	101	95	93		91	97	95
POLAND	113074	93		95	85	83	91		95	89	84
THAILAND	40080	93		95	79	73	94		91	81	92
HUNGARY	28574	92		84	94	93	85		88	89	87
ISRAEL	76781	92		96	78	95	90		88	95	94
OTHER EUROPE	107712	90		91	79	82	84		85	83	88
NEW ZEALAND	38946	90		90	111	88	93		92	95	90

(1a)	(1b)	(1c)		(2a)	(2b)	(2c)	(2d)		(3a)	(3b)	(3c)
Location	Number of Contributions	2015–6		Clinical and Anatomy	Drugs and Chemicals	Basic Science and Research Tools	Miscellaneous		Applied	Basic Science	Other (Both Applied and Basic Science)
TURKEY	157825	90		101	79	69	93		87	85	91
RUSSIA	51759	89		77	91	93	85		86	85	84
CHILE	23794	89		98	76	75	91		95	86	83
GREECE	46646	89		95	88	76	86		86	87	86
MALAYSIA	37997	87		87	64	81	90		87	83	83
PORTUGAL	65523	86		92	84	85	93		92	91	85
OTHER ASIA	60973	86		87	84	73	81		85	78	82
INDIA	291215	83		83	70	73	95		86	83	80
BRAZIL	274896	83		83	74	71	96		83	82	82
PAKISTAN	27511	83		78	93	74	80		81	79	77
MEXICO	54997	81		81	72	72	86		82	76	80
IRAN	121035	78		87	62	68	82		77	76	81
OTHER AMERICAS	30787	77		82	76	64	96		87	78	78
ARGENTINA	40775	77		85	76	69	86		82	79	78
EGYPT	48649	75		85	75	58	81		76	75	74
OTHER AFRICA	90041	70		87	57	55	73		76	68	72

Notes to Table 1:

All numbers in columns 1b and 1c are calculated based on papers published during 2015–2016. Numbers in columns 2a–d and columns 3a–3c are also calculated based on papers published during 2015–2016 for all countries for which the number of contributions reported in column 1c is at least 200000. For countries that fall below this threshold, the numbers in columns 2a–d and columns 3a–3c are calculated based on papers published during 2010–2016 (in order to decrease the variability of the edge factors reported in these columns).

Column 1a: Location.

Column 1b: Number of contributions based on which the edge factor in column (1c) is calculated. A contribution is defined as a link from a paper to an (idea category, research area) pair. A paper can link to multiple (idea category, research area) pairs because a paper can mention UMLS terms from multiple idea categories, and because a UMLS term can be linked to multiple idea categories, and because a paper may be linked to multiple research areas.

Column 1c: Edge factor for the baseline specification.

Columns 2a–d: Edge factors for each of the four idea category groups “Clinical and Anatomy”, “Drugs and Chemicals”, “Basic Science and Research Tools”, and “Miscellaneous”. See Table S1 for which idea categories (as represented by UMLS categories for UMLS terms) are included in each idea category group.

Columns 3a–c: Edge factors for each of the three types of research areas: “Applied”, “Basic Science”, and Other (Both Applied and Basic Science)”. See Table S1 for which research areas (represented by journal categories) are included in each of these three research area types.