Measuring Innovation in Science*

Wei Yang Tham

October 10, 2017

Introduction

The importance of scientific advancement to economic growth and quality of life is rarely questioned, but concretely defining and measuring innovative science is a challenging task. Yet doing so is an important step towards understanding the inputs into and properties of High Impact and Transformative Science (HITS), which will in turn hopefully lead to insights that can be applied to better policy.[1]

Although a better understanding of HITS is desirable across all fields of science, the issue has taken on a particular urgency in the biomedical research community. The combination of stagnant growth in government funding of basic research and the heavy reliance of biomedical researchers on those funds has led to concerns that researchers have become too hesitant to take on risky projects that could lead to scientific breakthroughs, because they perceive that "safe" but incremental projects are more likely to get funded (Stephan 2015; Alberts et al. 2014).

Researchers interested in HITS have started to move beyond publication and citation counts by analyzing citation networks and text analysis. In studies of the biomedical field, researchers have also been able to make use of the MeSH taxonomy instead of or in conjunction with

*Research reported in this paper was supported by the National Institute on Aging of the National Institutes of Health under Award Number R24AG048059 to the National Bureau of Economic Research. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institutes of Health or the NBER.

[1]This term is from Staudt et al. (2017)

publication text.[2] This has to led to new measures that can capture different aspects of HITS.

I review the literature that has developed these measures and the insights from using these measures in studies. I will focus on the efforts to move beyond publication and citation counts and quantify *impact*, *riskiness*, *novelty*, and *distance* (between fields, journals, papers, or researchers) using citation networks and text analysis, or combinations of both.

Impact

A straightforward way to measure impact is through citation counts. For the researcher is interested in "high impact" work, a common approach is to define a paper as "high impact" when it is in the N^{th} percentile for forward citations received for papers in the same cohort (i.e. published in the same year). The choice of N is admittedly ad-hoc, and typically different values of N are attempted for robustness checks.

Azoulay, Graff-Zivin, and Manso (2011) use a variation of this approach where they look at a paper's citation performance relative to an author's own previous publications. Specifically, Azoulay, Graff-Zivin, and Manso (2011) compare the citation quantile of an article (for articles in the same cohort) to citation quantiles for an author's previous publication. If the article outperforms an author's previous highest citation quantile, then it is defined as a "hit"[3]. This "within-person" measure is useful for capturing information about an individual scientist's behavior, even though what may be a hit paper for a scientist might not be a hit for the overall scientific community.

These measures do not capture all the dimensions of impact that we are interested in. Staudt et al. (2017) point out that we may be interested in the breadth of an article's impact (number and diversity of fields impacted) or whether it has growing impact (how the article's impact unfolds over time), and propose ways to measure those outcomes.

[2] A description of MeSH is available in the last section of this paper.
[3] A similar definition is used to define "flops".

Breadth of Impact. Staudt et al. (2017) measure breadth of impact by the concentration of fields that cite a paper. Fields are defined as MeSH terms at the 4-digit level (referred to as "MeSH4" in their paper). More detailed MeSH terms (i.e. beyond the 4-digit level) are aggregated to the MeSH4 terms they are associated with. Since lower-level terms can be associated with more than one higher-level term, each lower-level term is evenly distributed across its associated higher level terms.

Having defined fields, Staudt et al. (2017) propose a citation-based measure of breadth. Their measure is a Herfindahl index of fields that have cited the focal article.[4]

Staudt et al. (2017) also propose a text-based measure of breadth. They process the titles and abstracts of all articles from MEDLINE, extracting n-grams for $n \in \{1, 2, 3\}$.[5] Defining an n-gram's vintage year as the year it first appeared in an article, they further restrict the set of n-grams to those ranked in the top 0.01% of their vintage year.

They then measure the breadth of impact by a similar approach to the citation measure. They count the number of times a concept mentioned in other fields, and compute the Herfindahl index for shares of mentions by field.

Timing of Impact. Staudt et al. (2017) calculate the age of an article's forward citations as the difference between the publication year of the citing article and the focal/cited article, and then find the average of the article's forward citation ages. The longer an article takes to have an impact, the higher the average forward citation age. They find that average forward citation age is negatively correlated with their other HITS measures, suggesting that the impact of transformative work is largest early on.

Wang, Veugelers, and Stephan (2017) take a slightly different approach and look at how long it takes for a paper to become a hit (top 1% in citations). After developing a measure of novelty (discussed further in the Novelty section), they find that highly novel papers suffer from delayed recognition, taking at least 3 years to be as likely as non-novel papers to become a hit. They find that this delayed recognition occurs in both a paper's home and foreign fields,

[4] That is, $1 - \sum_f s_f^2$, where f denotes a field and s_f is the share of forward citations from that field.

[5] A 1-gram is a single word, a 2-gram is a 2 word sequence, and so on.

but novel papers catch up with (for home fields) or surpass (for foreign fields) non-novel papers over a sufficently long period of time. These results do not necessarily contradict Staudt et al. (2017) since they compare the *relative performance* of novel papers to non-novel papers.

Novelty

I discuss two approaches to measuring novelty. The first makes use of the fact that citations carry information about what areas of knowledge a paper is drawing on and combining. The second uses text analysis to track the appearance and performance of concepts in the literature.

Citations

The citation approach is motivated by a view of creativity as the unique combination of existing ideas. Uzzi et al. (2013) translate this view into a concrete measure based on the occurrence of co-citations relative to chance.[6] First, they count the co-citation frequency of journal pairs that appear in a paper's references, treating journals as a proxy for knowledge domains. Next, they implement a Markov Chain Monte Carlo algorithm that randomly reassigns citations *at the paper-level*. Their algorithm preserves the number of citations to and from a paper, as well as the timing of the citations.

Using this algorithm, Uzzi et al. (2013) create 10 randomized citation networks, counting the frequency of journal co-citations for each randomized network. This creates a distribution of co-citation frequency for each journal pair, which they use to calculate a z-score for each observed journal pair:

$$z = \frac{freq - \mu}{\sigma}$$

For each journal pair, $freq$ is the observed frequency of co-citations, while μ and σ are the

[6] A co-citation is two articles being cited together in the same article

empirical mean and standard deviation of the 10 simulated networks respectively. Each paper now has a distribution of z-scores. Uzzi et al. (2013) focus on two summary statistics for each paper - the median z-score and the 10th percentile z-score. They find that papers with a conventional core and high tail novelty - that is, a high median z-score and low 10th percentile z-score - are twice as likely to be a "hit" than the average paper, where a hit is defined by a citation percentile cutoff. They interpret these results as suggesting that work is more likely to be high impact when it is both grounded in conventional knowledge but also makes use of unusual combinations of knowledge.

Wang, Veugelers, and Stephan (2017) draw on the recombination model of creativity as well. Compared to Uzzi et al. (2013), who measure the atypicality or unusualness of a combination, Wang, Veugelers, and Stephan (2017) focus on whether a combination is *new*. Just as in Uzzi et al. (2013), they look at co-citations of journal pairs. A journal pair is considered "new" or "novel" if the journals are being cited together for the first time in the literature.[7]

Wang, Veugelers, and Stephan (2017) use articles published in 2001 and indexed in the Web of Science Core Collection. They introduce a "difficulty" score intended to capture how hard it is to combine two journals. They begin by constructing a symmetric co-citation matrix where the i, j-th entry is the number of times journals i and j have been co-cited in years 1998, 1999, and 2000. The difficulty score for ij is calculated as the cosine similarity between each journal's co-citation vector.[8]

They then define the novelty score for a paper as the sum of novelty scores over all journal pairs that appear in the paper.[9]

$$Novelty = \sum_{ij} \mathbf{1}(New)_{ij} Cosine_{ij}$$

[7]Due to data limitations, Wang, Veugelers, and Stephan (2017) actually observe whether a journal pair is new *since 1980*

[8]Cosine similarity is a measure of distance between vectors. I discuss it in more detail in the Intellectual Similarity and Multidisciplinarity section.

[9]Wang, Veugelers, and Stephan (2017) construct two alternative versions of this score, one using the maximum novelty score of the paper and another that essentially normalizes for the number of references.

A replication by Boyack and Klavans (2014) finds that Uzzi et al. (2013) may not have sufficiently accounted for differences across fields and journals. Their findings suggest that after accounting for field differences in atypicality of combinations, the results from Uzzi et al. (2013) hold up but are quantitatively less pronounced. In addition, they warn that journals are a poor proxy for knowledge domains, especially multidisciplinary journals like *Science* and *Nature*. Wang, Veugelers, and Stephan (2017) is subject to these same concerns and try to address them with a variety of robustness checks. Boyack and Klavans (2014) note that these issues do not invalidate the underlying ideas of Uzzi et al. (2013).

Text: Birth and Age of Concepts

A prominent use of text to measure novelty has been to track the "birth" of concepts, as illustrated in Packalen and Bhattacharya (2015a) and related papers (Packalen and Bhattacharya 2017; Packalen and Bhattacharya 2015b). Packalen and Bhattacharya (2015a) extract 2-grams and 3-grams from the titles and abstracts of articles in the MEDLINE database. These n-grams are regarded as idea inputs into a paper. The authors find the earliest year that an n-gram appears in a publication in the MEDLINE database. For a given article, the ages of its idea inputs are the difference between the article's publication year and the cohort years of the concepts.[10]

In Packalen and Bhattacharya (2017), the authors repeat the same exercise by searching the article text for terms obtained from the United Medical Language System (UMLS) metathesaurus, rather than directly extracting word sequences. An advantage of using a thesaurus is that it can be used to identify synonyms and prevent misattribution of cohort years, but not all fields will have a thesaurus.

Other examples that use the birth and age of concepts to measure novelty include Azoulay, Graff-Zivin, and Manso (2011), which uses the age of MeSH terms, and Staudt et al. (2017),

[10]Packalen and Bhattacharya (2015a) note that the first article that a concept appeares in may not necessarily be the originating article for the concept. For example, the concept may have originated outside of the biomedical literature.

which focus on the birth of n-grams and timing of their usage in articles.

The textual approach offers the following advantages over the combinatorial approach (Packalen and Bhattacharya 2017). First, counting combinations of idea inputs (such as papers or chemicals) quickly becomes computationally costly. Second, Packalen and Bhattacharya (2017) argue that the generation and development of new ideas may be more important than new combinations of old ideas. Presumably these phenomena are related to some extent. Understanding the interaction between these two types of novelty metrics could lead to a better understanding of how new ideas are generated.

Intellectual distance/similarity

Azoulay, Graff-Zivin, and Manso (2011) test whether a scientist has deviated from their previous field of research. This can be thought of as measuring the intellectual distance of the scientist's work in the pre- and post-periods. They count the number of common unique MeSH terms, normalized by the number of unique keywords in the post-period.

Another option is to use a *cosine similarity* measure. Cosine similarity is a measure of the distance between two vectors, the dot product of the vectors normalized by their magnitudes. A geometric interpretation is that it is a measure of the angle between two vectors.

$$Cosine_{ij} = \frac{v_i \cdot v_j}{||v_i||\,||v_j||}$$

Catalini (2017) uses cosine similarity in citation space and keyword space as an outcome to determine the effect of colocation on collaboration patterns, but also as a proxy for search costs between knowledge domains. Wang, Veugelers, and Stephan (2017) use the cosine similarity of journals in citation space as a proxy for the difficulty of combining two fields. These similar but still distinct uses for cosine similarity illustrate that a direct economic interpretation of cosine similarity can be difficult. Nonetheless, its computational simplicity

and its wide use in other fields (e.g. information retrieval) make it an appealing tool.[11]

Cosine similarity can be calculated with weights instead of raw counts in order to account for words that are common are therefore less helpful in distinguishing content. This can be done with *tf-idf* and its variants.[12] The implicit assumption that words are orthogonal to each other can also be relaxed in different ways, such as through a generalized vector space model that takes into account the relatedness of different word pairs (Wong, Ziarko, and Wong 1985).

For researchers interested in using more sophisticated measures to measure intellectual distance or delineate fields but are deterred by their complexity, Azoulay, Fons-Rosen, and Zivin (2015) illustrates how it can be possible to use such measures without having to generate them. The authors use the PubMed Related Citations Algorithm (PMRA), which powers PubMed's search feature, to delineate research fields. PMRA scores can also be scraped from PubMed.

Riskiness

Work that is novel or high impact is also thought to be riskier. Measuring risk and relating it to measures of impact and novelty can give us better insight into the factors that affect scientists' propensity for risk-taking.

Azoulay, Graff-Zivin, and Manso (2011) define risk as the frequency of tail outcomes, or "hits" and "flops". Just as hits are defined as papers that hit a certain threshold in citation performance, flops can be defined symmetrically. The key here is that both hits *and* flops must be taken into account. Staudt et al. (2017) take a similar approach, but rather than setting a threshold for hits and flops, they measure the variance of citation counts.

Wang, Veugelers, and Stephan (2017) look at both variance and tail outcomes. They follow

[11] For uses of cosine similarity in other areas of economics, see Gentzkow, Kelly, and Taddy (2017).
[12] See https://en.wikipedia.org/wiki/Tf-idf

Fleming (2001) by estimating the variance of citation counts in a Negative Binomial Count model. Since the variance is estimated within a statistical model, they can control for factors such as field effects and number of authors. Using their measure of novelty, they find that "highly novel" papers have a greater dispersion than "moderately novel" papers, relative to non-novel papers. They also find that highly novel papers are more likely to be in both the top and bottom 10% of the citation distribution than non-novel papers, but moderately novel papers are more likely to be in the upper tail but not the lower tail.

A problem that is common to any measure of risk based on publications is that outcomes on the left-tail are censored, as there are projects that may be abandoned without even being sent to publication(Catalini 2017). The extent of this bias may vary by field if there are differences in the pressure to publish, but in general the measures discussed so far may understate the variance of outcomes.

In fields such as the biomedical sciences where scientists are heavily reliant on funding, another relevant dimension of risk is the probability of being funded and amount of funding a scientist anticipates getting (Alberts et al. 2014). Azoulay, Graff-Zivin, and Manso (2011) explores this dimension by comparing scientists in different incentive schemes, but does not explicitly measure risk.

Applications

In this section I present a few different examples that illustrate the wide range of questions that can be answered with these metrics.

Comparing Incentive Systems

Azoulay, Graff-Zivin, and Manso (2011) compare the research behavior of scientists under different incentive schemes. They study scientists under the Howard Hughes Medical Institute (HHMI) investigator program, which is a system that is more tolerant of early failure than

the typical R01 grant. Using a variety of measures including citation outcomes and MeSH terms, they find that high-achieving scientists are more likely to explore new areas and take more risks under the HHMI program.

Age-Innovation Relationship

Packalen and Bhattacharya (2015a) use the age of ideas to study the relationship between scientist (career) age and trying out of new ideas in biomedical research. They determine whether an article tries out new ideas by the age of its idea inputs relative to other papers in its cohort. They find that younger researchers do try out new ideas more, and that teams with a young first author and experienced last author are the most likely combination to try out new ideas.

Effect of Colocation

Catalini (2017) uses a quasi-experiment to study the effect of colocation on collaboration patterns between labs. They find that colocation increases the probability of collaboration, especially for labs that have higher search costs (as proxied by intellectual distance). They also find that colocated labs become more similar in both citation and keyword space, again measured by cosine similarity. Finally, there is some evidence that by increasing collaboration between distant fields, colocation leads to more left-tail and right-tail outcomes.

Data

MEDLINE

MEDLINE is a bibliographic database of the U.S. National Library of Medicine. It can be downloaded by anyone, free of charge. The data can be accessed through the Entrez API. For R users, a useful package for interacting with Entrez is the `rentrez` package.

Author-ity

Although MEDLINE is free and relatively convenient to access, identifying individuals in the dataset can be problematic. The same individual may have multiple variants of their name, and multiple individuals may have the same name. Author-ity (Torvik and Smalheiser 2009) is a database of disambiguated authors names based on a snapshot of PubMed (which includes MEDLINE) and can be found at this link, along with many other potentially useful tools.

Citations: Web of Science and Scopus

The Clarivate Analytics Web of Science and Scopus are citation-indexing databases. They have the advantage of covering multiple disciplines, but are not free.

NIH ExPORTER

The NIH makes available data on grants awarded through ExPORTER. A rich set of variables is provided, including information about the principle investigators of the grant and funding amounts. Perhaps most importantly in the context of this white paper, they also provide a research abstract, opening the door to the application of the text measures discussed in this paper.

MeSH

MeSH (Medical Subject Headings) is a vocabulary controlled by the National Library of Medicine (NLM) that is used to classify articles in the MEDLINE/PubMed database. Articles are read and assigned MeSH terms by librarians at NLM. MeSH terms have a hierarchical structure that can be viewed at http://www.nlm.nih.gov/mesh/MBrowser.html. MeSH files can also be downloaded at https://www.nlm.nih.gov/mesh/filelist.html.

MeSH are a relatively convenient way to incorporate textual information into an analysis or to delineate fields (e.g. Staudt et al. (2017)), but may not perform as well as text for certain uses (e.g. Boyack et al. (2011)). The hierarchical structure of MeSH terms also means that distinct terms can be closely related to each other. In some use cases, it may be necessary to take into account these relationships (e.g. relaxing the orthogonality assumption for the standard cosine similarity measure).

References

Alberts, Bruce, Marc W Kirschner, Shirley Tilghman, and Harold Varmus. 2014. "Rescuing Us Biomedical Research from Its Systemic Flaws." *Proceedings of the National Academy of Sciences* 111 (16). National Acad Sciences: 5773–7.

Azoulay, Pierre, Christian Fons-Rosen, and Joshua S Graff Zivin. 2015. "Does Science Advance One Funeral at a Time?" National Bureau of Economic Research.

Azoulay, Pierre, Joshua S Graff-Zivin, and Gustavo Manso. 2011. "Incentives and Creativity: Evidence from the Academic Life Sciences." *The RAND Journal of Economics* 42 (3). Wiley Online Library: 527–54.

Boyack, Kevin W, and Richard Klavans. 2014. "Atypical Combinations Are Confounded by Disciplinary Effects." *STI 2014 Leiden* 64.

Boyack, Kevin W, David Newman, Russell J Duhon, Richard Klavans, Michael Patek, Joseph R Biberstine, Bob Schijvenaars, André Skupin, Nianli Ma, and Katy Börner. 2011. "Clustering More Than Two Million Biomedical Publications: Comparing the Accuracies of Nine Text-Based Similarity Approaches." *PloS One* 6 (3). Public Library of Science: e18029.

Catalini, Christian. 2017. "Microgeography and the Direction of Inventive Activity." *Management Science*. INFORMS.

Fleming, Lee. 2001. "Recombinant Uncertainty in Technological Search." *Management*

Science 47 (1). INFORMS: 117–32.

Gentzkow, Matthew, Bryan T Kelly, and Matt Taddy. 2017. "Text as Data." National Bureau of Economic Research.

Packalen, Mikko, and Jay Bhattacharya. 2015a. "Age and the Trying Out of New Ideas." National Bureau of Economic Research.

———. 2015b. "Cities and Ideas." National Bureau of Economic Research.

———. 2017. "Neophilia Ranking of Scientific Journals." *Scientometrics* 110 (1). Springer: 43–64.

Staudt, Joseph, Huifeng Yu, Robert P. Light, Gerald Marschke, Katy Börner, and Bruce A. Weinberg. 2017. "High-Impact and Transformative Science (Hits) Metrics: Definition, Exemplification, and Comparison."

Stephan, Paula. 2015. "The Endless Frontier: Reaping What Bush Sowed?" In *The Changing Frontier: Rethinking Science and Innovation Policy*, 321–66. National Bureau of Economic Research.

Torvik, Vetle I, and Neil R Smalheiser. 2009. "Author Name Disambiguation in Medline." *ACM Transactions on Knowledge Discovery from Data (TKDD)* 3 (3). ACM: 11.

Uzzi, Brian, Satyam Mukherjee, Michael Stringer, and Ben Jones. 2013. "Atypical Combinations and Scientific Impact." *Science* 342 (6157). American Association for the Advancement of Science: 468–72.

Wang, Jian, Reinhilde Veugelers, and Paula Stephan. 2017. "Bias Against Novelty in Science: A Cautionary Tale for Users of Bibliometric Indicators." *Research Policy*. Elsevier.

Wong, SK Michael, Wojciech Ziarko, and Patrick CN Wong. 1985. "Generalized Vector Spaces Model in Information Retrieval." In *Proceedings of the 8th Annual International Acm Sigir Conference on Research and Development in Information Retrieval*, 18–25. ACM.