# Citation Distance: Measuring Changes in Scientific Search Strategies

**Ryan Whalen**
Northwestern University
Evanston, USA
r-whalen@northwestern.edu

**Yun Huang**
Northwestern University
Evanston, USA
yun@northwestern.edu

**Craig Tanis**
University of Tennessee, Chattanooga
Chattanooga, USA
craig-tanis@utc.edu

**Anup Sawant**
Northwestern University
Evanston, USA
anup.sawant@northwestern.edu

**Brian Uzzi**
Northwestern University
Evanston, USA
uzzi@kellogg.northwestern.edu

**Noshir Contractor**
Northwestern University
Evanston, USA
nosh@northwestern.edu

## ABSTRACT

Using latent semantic analysis on the full text of scientific articles, we measure the distance between 36 million citing/cited article pairs and chart changes in citation proximity over time. The analysis shows that the mean distance between citing and cited articles has steadily increased since 1990. This demonstrates that current scholars are more likely to cite distantly related research than their peers of 20 years ago who tended to cite more proximate work. These changes coincide with the introduction of new information technologies like the Internet, and the increasing popularity of interdisciplinary and multidisciplinary research. The "citation distance" measure shows promise in improving our understanding of the evolution of knowledge. It also offers a method to add nuance to scholarly impact measures by assessing the extent to which an article influences proximate or distant future work.

## Author Keywords

Citation analysis; scholarly impact; citation distance; science of science

## 1. INTRODUCTION

The introduction of powerful information technologies like the Internet has transformed the accessibility of scientific information. Articles that once reached scholars via trips to libraries and subscriptions to academic journals are now available online in large searchable databases. Digitalization and search engines have the power to change the way scientists search for and combine information, yet thus far we have little insight into how research strategies have changed in recent decades.

By measuring the "knowledge distance" between citing and cited articles, we explore how research strategies have changed as information technologies have altered the way scholars perform their work. We measure the distance traversed by over 36 million citations and show that the mean distance between citing and cited articles has steadily increased since 1990.

## 2. INTERNET & RESEARCH

Improvements in information technology have had countervailing effects for the production of scholarly research. On the one hand, increased ease of writing, and publication venues have led to increased scholarly output [6]. This leads to a "burden of knowledge" leaving scholars less able to master broad expanses of knowledge and more prone to specialization [7]. On the other hand, improved search technologies and increasingly expansive databases of scholarly research have improved our ability to seek out and find diverse pieces of knowledge [16], while also perhaps reducing our probability of serendipitously encountering useful information [11]. Improved search capabilities potentially mitigate the challenges created by the burden of knowledge. By enabling researchers to more easily navigate the vast amount of knowledge created, improvements in information technology help scholars filter through the information available to them.

Despite these countervailing effects, we know relatively little about how the introduction of new technologies have influenced researchers' search and citation strategies. There is some evidence to suggest that in recent years research has become more interdisciplinary in nature [12], and that this interdisciplinarity is rewarded with somewhat higher impact [1,18]. However, this body of research relies extensively on metadata to infer the content of cited articles. As such, it is an example of "metaknowledge" research, capitalizing on the growing amount of metadata about science available to researchers [3]. Metaknowledge research is by definition

Copyright is held by the International World Wide Web Conference Committee IW3C2). IW3C2 reserves the right to provide a hyperlink to the author's site if the Material is used in electronic media.
WWW 2016 Companion, April 11-15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2890515.

somewhat abstract. It relies on assumptions about article content by categorizing articles based on the journal of publication, or by clustering them based on citation patterns.

As data availability improves, the abstractions inherent in metaknowledge research become less essential. Just as the increased availability of computation and scientific metadata ushered in a boom in metaknowledge research, further improvements in computational power along with the increasing availability of large corpuses of textual material suggest that the next stage of "science of science" research will be able to peel back some of the layers of abstraction inherent in metaknowledge studies by focusing more closely on the actual content in publications. Here, we contribute to this data-driven science of science research by using the full text of scientific articles as we assess the evolution of scientific knowledge search and citation styles.

Broadly speaking, there are two types of search strategies, each with its own implications for the resulting research product. The predominant search strategy is "local" or exploitative in nature, as it seeks to exploit expertise related to one's research area [13,14]. The remainder is explorative, seeking out knowledge that is distant from one's area of expertise [10]. Exploitative search focuses on improving existing technologies and refining existing ideas, whereas explorative search seeks to generate new groundbreaking technologies and ideas.

While theory & empirical studies suggest that most research tends to be exploitative in nature—improving and refining existing technologies and ideas—we know very little about if and how improvements in information technology like the Internet have altered these tendencies. Empirically testing whether or not research tendencies have responded to changes in information technology capabilities, and how scholars' search strategies have responded to the burden of knowledge has important implications for science & technologies studies, and research policy. Thus, our research question is:

**RQ**: Have scholars' research strategies changed along with changes in information technology capabilities and access rates?

In order to answer our research question, we use citation records to measure changes in research strategies over time. The citations between research articles provide a record of the knowledge that scholars have drawn upon [5]. We can use these trace records of scholarly research to provide insight into the effects that changes in information technologies and research norms have had on search strategies in recent decades. The challenge lies in measuring the meaning inherent in citations. Most citation research treats citations as binary relationship, providing little insight into the search strategies underlying a research project.

To improve on comparatively coarse binary citation measures, we measure the "distance" between citing and cited articles, providing insight into how far from their own field scholars searched for inspiration. This "knowledge distance" is an important element in understanding the nature of the idea recombination that underlies scholarly output. It relies on the assumption that some ideas are more closely related to one another (e.g. two articles exploring the historical economic development of Southeast Asia) while other articles are more distantly related to one another (e.g. an article on Southeast Asian economic development, and an article on dolphin social networks). These varied relationships provide clues as to which search strategies the authors engaged in. If their citations are to predominantly closely-related articles, we can infer their search strategy was predominantly exploitative. On the other hand, if their work features many distant citations, we can presume their search strategy was more exploratory in nature.

## 3. DATA & METHODS

In order to examine how scientific search and citation strategies have changed over time, we first construct a corpus of Elsevier's publications and identify the journal articles indexed by the Thomson Reuters Web of Science. The full text of 5,172,578 articles is extracted from the Science Direct database and the citation relationships among them are extracted based on the Web of Science.[1]

In order to measure citation distance, we use latent semantic analysis (LSA) [2] to reduce term-document relations in the corpus into a 400-dimension latent space. Each article is represented as an LSA vector in the space and the distance between articles is measured by calculating the cosine distance between their vectors [9]. We compute citation distance for all 36,864,014 citation relations in our dataset.

We assign each article an OECD topic categorization [19] corresponding the Revised Field of Science and Technology (FOS) Classification of the Frascati Manual 2002, determine the year of publication, and chart the mean citation distance over time.

## 4. RESULTS

Figure 1 shows the distribution of the citation distances between randomly selected articles. This distribution, skewed towards high distances, is not surprising since two randomly selected articles are likely to have little in common.

[1] We use this dual database approach because of the qualities of the data available to us in each. We have full text of articles from Science Direct, but these do not contain unique identifiers for citations. So, we reconcile the Science Direct articles with their Web of Science entries and extract the citations from the Web of Science.

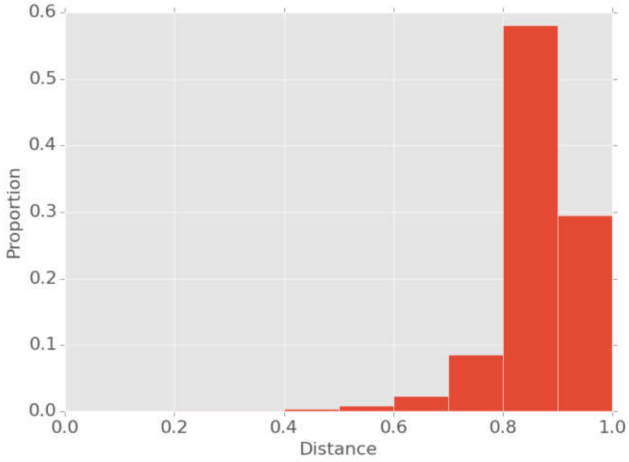

**Figure 1:** *The distribution of distances between randomly paired articles.*

Examining the distance relationships between actual citing/cited pairs of articles in the dataset shows a wide range of citation tendencies and paints a starkly different picture. Figure 2 plots the distribution of citation distance scores, showing a strong tendency towards proximate citations. The mean citation distance is 0.35, with a standard deviation of 0.21.

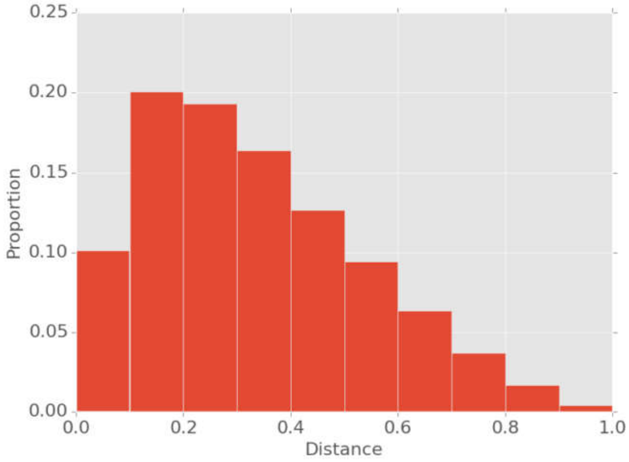

**Figure 2:** *Distribution of citation distances.*

These distributions show a strong tendency for researchers to cite proximate literature, with highly distant citations making up a very small proportion of total citations. While these measures combine all fields into a single view, similar analyses could be used in the future to assess how "insular" any given field, or organization is.

Plotting the mean citation distances over time shows a steady increase beginning in the early 1990s and continuing through to recent years. Over this time the mean citation distance rose from 0.32 to approximately 0.37 (see Figure 3), as researchers began to make citations to increasingly distant articles.

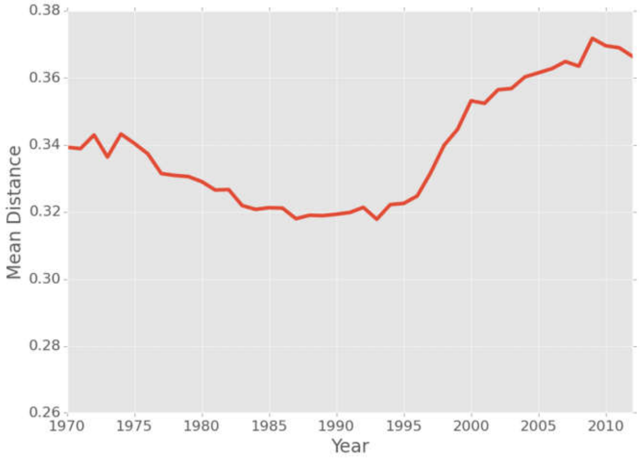

**Figure 3:** *Mean citation distance by year.*

Categorizing the distances by their OECD research categorization shows substantial variation between disciplines (see Figure 4). Research in the humanities and social sciences began the period with substantially lower levels of average citation distances than their hard science peers. Over the next four decades, much of this distinction vanishes as the citation norms in the social sciences and humanities have grown to become more similar to those of their peers.

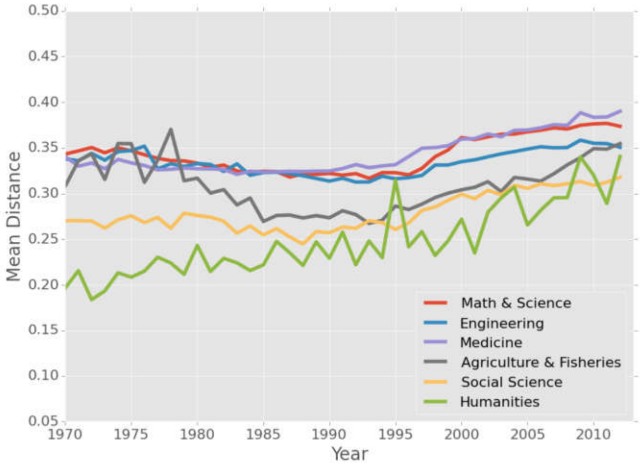

**Figure 4:** *Mean yearly citation distance by research discipline.*

By 2012—the latest publication date for articles in our dataset—the various disciplines appear to be converging towards a common mean citation distance.

## 5. DISCUSSION

These results suggest that knowledge search and citation strategies have changed in recent decades, as information technologies have altered search capabilities. We do not have sufficient evidence to make causal claims, but we do note that the current trend towards increasing citation distance coincides with the introduction of open access to the Internet, and the development of the World Wide Web. Scholars working in recent years are more likely to cite to work that is dissimilar to their own than their peers would have been 20 years ago. Access to these information technologies may explain much of the increase in citation distance we have observed. Researchers are now able to easily search vast databases of scientific publications, making them more likely to encounter works in distant disciplines.

This general trend towards increased citation distance has also coincided with a rise in multidisciplinary and interdisciplinary research [8,12]. As scholars draw on research from other disciplines, their citations are more likely to reflect this by referencing highly distant articles. Both the rise in Internet access rates and rising interdisciplinarity may be related to the increase in citation distance that we observe. Indeed, these two phenomena may be related to one another as the Internet collapses boundaries between disciplines. While scholars once relied extensively on their own discipline's journals and conferences, the Internet has now enabled them to much more easily access the work of scholars in different fields.

Our analysis suggests that changes to research norms have not been uniform across scientific disciplines. Figure 4 shows that the "softer" sciences have experienced a stronger trend toward making distant citations, while other areas have experienced only a modest increase. This is perhaps due to the nature of each research area. In some disciplines, the barrier to integrating distant research is much lower. For instance, in the social sciences or humanities, which are not strictly bound by physical laws governing their topics of study, it is perhaps easier to seek out and draw on distant knowledge. On the other hand, the "harder" sciences may have a somewhat higher barrier to integrating distant knowledge as the subjects of their research are more strictly bounded. Put another way, there are many humanistic and social scientific works that draw on scientific theories (e.g. the theories of relativity or the theory of evolution) but many fewer hard science works that draw on humanistic and social scientific theories.

Our analysis shows a general increase in explorative type citations, suggesting that researchers now reach further into our collective "knowledge space" in search of inspiration than they once did. This is potentially a good news story, as explorative type research is often considered difficult to adequately incentivize [10]. This explorative type research has a greater tendency of leading to important scientific developments and high impact research [1,4,15,18].

### The Citation Distance Measure

In addition to the substantive results we report here, this work demonstrates the potential for improved citation analysis. As more and more scholarly communication moves online, and as researchers have increased access to full text databases, opportunities to apply natural language processing techniques to analyze scholarly communication increase. We show here that analyzing citation distance can provide significant information that is omitted from traditional binary citation analyses. This same method can be applied to forward citations, measuring not how researchers draw on the existing knowledge space, but rather the nature of the impact their work goes on to have. Some research is influential only within its own field, other works have broad relevance to many types of researchers. Previous impact measures largely ignore this facet of research impact. Applying our method of citation distance measurement, could help ameliorate this issue and improve impact measurement by assessing how broad it is.

Similar techniques could also be used to better understand the evolution of scientific disciplines over time. Measuring disciplinary tendencies to draw on distant knowledge will help us understand how disciplines form and change over time by examining how researchers integrate proximate and distant knowledge in their work. Analyzing these tendencies in conjunction with scientific impact offers the potential to further our understanding of how knowledge integration patterns relate to scientific success.

Our observed increase in citation distance also corresponds with an increase in the importance of scientific teamwork [17]. Our future work will combine citation distance metrics with team analysis to help further our understanding of how team size and composition relate to both patterns of knowledge integration, and scientific success.

## 6. CONCLUSION

We have shown that, in recent years, scholars have increased their tendency to cite distantly related articles. Since the 1990s, mean citation distance has steadily increased. This occurred as information technologies have eased access to diverse information sources, and as interdisciplinary and multidisciplinary teamwork has also steadily increased. This increase is common across scientific fields, but is most pronounced in the social sciences and humanities. Our method of measuring citation distance shows promise in furthering our understanding of how scientific research occurs and evolves, and in nuancing the way we measure scientific impact.

## 7. ACKNOWLEDGEMENTS

This material is based upon work supported by, or in part by, the U. S. Army Research Laboratory and the U. S. Army Research Office under grant numbers W911NF-14-10686 and W911NF-15-1-0577.

The views, opinions, and/or findings contained in this manuscript are those of the authors, and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documents.

This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1053575.

## 8. REFERENCES

1. Shiji Chen, Clément Arsenault, and Vincent Larivière. 2015. Are top-cited papers more interdisciplinary?

*Journal of Informetrics* 9, 4: 1034–1046. http://doi.org/10.1016/j.joi.2015.09.003

2. Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *JAsIs* 41, 6: 391–407.
3. James A. Evans and Jacob G. Foster. 2011. Metaknowledge. *Science* 331, 6018: 721–725. http://doi.org/10.1126/science.1201765
4. Jacob G. Foster, Andrey Rzhetsky, and James A. Evans. 2015. Tradition and Innovation in Scientists' Research Strategies. *American Sociological Review* 80, 5: 875–908. http://doi.org/10.1177/0003122415601618
5. Eugene Garfield. 1979. Is citation analysis a legitimate evaluation tool? *Scientometrics* 1, 4: 359–375. http://doi.org/10.1007/BF02019306
6. Arif E. Jinha. 2010. Article 50 million: an estimate of the number of scholarly articles in existence. *Learned Publishing* 23, 3: 258–263. http://doi.org/10.1087/20100308
7. Benjamin F Jones. 2009. The Burden of Knowledge and the "Death of the Renaissance Man": Is Innovation Getting Harder? *The Review of Economic Studies* 76, 1: 283–317.
8. Julie Thompson Klein. 1990. *Interdisciplinarity: History, theory, and practice*. Wayne State University Press.
9. Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. 1998. An introduction to latent semantic analysis. *Discourse Processes* 25, 2-3: 259–284. http://doi.org/10.1080/01638539809545028
10. James G. March. 1991. Exploration and Exploitation in Organizational Learning. *Organization Science* 2, 1: pp. 71–87.
11. Eli Pariser. 2011. *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. Penguin.
12. Alan L. Porter and Ismael Rafols. 2009. Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics* 81, 3: 719–745. http://doi.org/10.1007/s11192-008-2197-2
13. Lori Rosenkopf and Atul Nerkar. 2001. Beyond Local Search: Boundary-Spanning, Exploration, and Impact in the Optical Disk Industry. *Strategic Management Journal* 22, 4: 287–306.
14. Toby E Stuart and Joel M Podolny. 1996. Local search and the evolution of technological capabilities. *Strategic Management Journal* 17, S1: 21–38.
15. Brian Uzzi, Satyam Mukherjee, Michael Stringer, and Ben Jones. 2013. Atypical Combinations and Scientific Impact. *Science* 342, 6157: 468–472.
16. Lav R. Varshney. 2012. The Google effect in doctoral theses. *Scientometrics* 92, 3: 785–793. http://doi.org/10.1007/s11192-012-0654-4
17. Stefan Wuchty, Benjamin F Jones, and Brian Uzzi. 2007. The Increasing Dominance of Teams in Production of Knowledge. *Science* 316, 5827: 1036–1039.
18. Alfredo Yegros-Yegros, Ismael Rafols, and Pablo D'Este. 2015. Does Interdisciplinary Research Lead to Higher Citation Impact? The Different Effect of Proximal and Distal Interdisciplinarity. *PLoS ONE* 10, 8: e0135095. http://doi.org/10.1371/journal.pone.0135095
19. OECD Classification — Web of Science Subject Headings. Retrieved December 19, 2015 from http://incites.isiknowledge.com/common/help/h_field_category_oecd_wos.html