Ryan Whalen
Northwestern University, Evanston, IL, USA &
Dalhousie University, Halifax, NS, Canada

Noshir Contractor
Northwestern University, Evanston, IL, USA

Citation Distance: Measuring Knowledge Translation, Integration, Diffusion, and Scope

Abstract: This paper introduces and demonstrates four measures of weighting forward and backward citations based on the distance between citing/cited documents in a document vector space model.

1. Introduction

Citation analysis has been an important tool in information science for decades (Garfield, 1979). Citations are used to assess scholarly productivity (Hirsch, 2005) and journal prestige (Garfield, 1972; Glänzel & Moed, 2002), to assist in funding allocation (Abramo, D'Angelo, & Caprasecca, 2009) and tenure and promotion considerations (Holden, Rosenberg, & Barker, 2005; Segalla, 2008) as well as to improve our understanding of how knowledge is generated (Lee, Walsh, & Wang, 2015; Uzzi, Mukherjee, Stringer, & Jones, 2013) and how it diffuses (Börner, Penumarthy, Meiss, & Ke, 2013; Singh, 2005).

Despite their importance to information science, citation metrics remain relatively coarse. For the most part, citations are a binary construct: they either exist, or do not. There have been attempts to categorize or weight citations based on the source journal or disciplinary norms (Leydesdorff & Bornmann, 2011; Moed, 2010), the network centrality of the citing publication (Chen, Xie, Maslov, & Redner, 2007; Leydesdorff, 2009) the intentions of the citing authors (Chubin & Moitra, 1975; Moravcsik & Murugesan, 1975), and the sentiment of the surrounding text (Catalini, Lacetera, & Oettl, 2015; Small, 2011). But these attempts have met with mixed success. There are currently no commonly used and scalable citation weighting methods that allow scholars to assess the nature of the relationship between the citing/cited publications, despite the fact that this type of measure would provide useful insight into the research and knowledge diffusion processes.

2. Knowledge Distance & Recombination

There are of course many ways that publications can be related, and many dimensions along which we could measure these relationships. In this project, we propose that the "topical distance" between publications is both salient to many important research questions, and a measurable trait that enables tractable metrics.

Distance is relevant to citation relations both because of the recombinatorial nature of research, and the knowledge search process that underlies it. Much of research relies on recombining existing knowledge in novel ways (Nelson & Winter, 1982; Schumpeter, 1939). By taking existing ideas and techniques, and assembling them into new combinations, researchers create new ideas and techniques or develop new applications for existing ideas. In seeking out ideas to recombine, researchers must search through existing knowledge. Much of this search is "local" or exploitative in nature, as it seeks to exploit expertise related to the research area (Rosenkopf & Nerkar, 2001; Stuart & Podolny, 1996). The remainder is explorative, as researchers seek out knowledge that is distant from their areas of expertise (March, 1991). These searches can lead researchers to make varied patterns of recombination, as their search strategies affect the knowledge they are exposed to (Fleming & Sorenson, 2004).

Evidence suggests that the pattern of recombination is highly important to the quality of the research outcome. For instance, Uzzi et al (2013) demonstrate that mixing atypical combinations of sources with relatively typical combinations creates knowledge that is much more likely to go on to be highly influential. Similarly, Foster and colleagues (2015) show that research making new or infrequently seen combinations of chemical compounds tends to have greater scientific impact, and Fleming shows that novel combinations are more varied in the degree of success they enjoy (Fleming, 2001). Measuring the distance between publications will help researchers assess patterns of knowledge recombination, and better understand how knowledge diffuses across disciplines.

3. Method

In this project we propose 4 citation distance metrics. To calculate our measures, we draw on patent data from the USPTO. Patent data provides a record of the evolution of the knowledge underlying technological development (Fleming & Sorenson, 2001), while the prior art citations they contain demonstrate relationships between inventions and suggest knowledge flows (Almeida & Kogut, 1999; Jaffe, Trajtenberg, & Henderson, 1993; Rosell & Agrawal, 2009). Because open publication is a condition of receiving a patent grant, patent data also allows access to the textual content of all granted patents. As such, patents provide a rich and accessible source of data with which to demonstrate citation distance measures, but our proposed measures would be equally applicable to analogous types of data including scientific articles.

To measure the distance between patents, we first perform latent semantic analysis dimensional reduction (Deerwester, Dumais, Landauer, Furnas, & Harshman, 1990) on the full text of all U.S. utility patents granted between 1976 and 2014. We then calculate the cosine distance between patents in this latent semantic space, weighting over 52 million citations by the topical distance between the citing and cited references, as well as weighting the relationships between random samples of co-citing and co-cited references. We propose and demonstrate four citation distance measures:

Knowledge translation: Defined as the backward citation distance between a focal patent and its prior art references, this measure provides insight into the degree to which researchers have translated distant knowledge to their own discipline.

Knowledge integration: Defined as the size of the minimum spanning tree of a distance-weighted fully-connected graph consisting of all the co-cited references of a focal patent, this measure provides insight into the variety of knowledge researchers integrate into a single invention.

Knowledge diffusion: Defined as the forward citation distance between a focal patent and the future patents that cite it, this measure provides insight into the degree to which the knowledge within a patent diffuses to topically distant fields.

Knowledge scope: Defined as the size of the minimum spanning tree of a distance-weighted fully connected graph consisting of of all the co-citing references of a focal patent, this measure provides insight into the degree to which an invention goes on to influence varied technical fields.

INSERT TABLE 1 HERE

4. Results

The general level of distance between any two patents is very high. Figure 1 plots the distribution of distance scores between 100,000 randomly selected patent pairs.

INSERT FIGURE 1 HERE

We see a starkly different story when we plot the distribution of distances between patents that share a citation relationship. Figure 2 plots the distribution of distances between 100,000 randomly selected patent pairs that share a citation relationship.

INSERT FIGURE 2 HERE

Analyzing the knowledge translation scores across time shows clear changes in the types of prior art citations that we observe over time. Figure 3 plots the mean citation distance score by year from 1980 to 2014, showing that citations have tended to come from increasingly distant publications as time has gone by.

INSERT FIGURE 3 HERE

Figure 4 demonstrates that this trend towards increased distance in citations is consistent across technical fields. When we classify each patent into one of 6 broad categorizations we see that in each category distance has steadily increased. Furthermore, citation distance is not only rising across categories, but the different technical fields appear to be converging.

INSERT FIGURE 4 HERE

Figure 5 graphs the mean distance and standard deviation in the co-cited minimum spanning tree. We see that the mean distance between co-cited references has decreased in recent years, meaning that on average patents have tended to co-cite more similar prior art. However, when we look to the standard deviation we see there has been a steady increase in the variance of co-cited distances.

INSERT FIGURE 5 HERE

When we flip the focus from backward to forward citations and look to knowledge diffusion, we again see significant changes over time. Let us first look to how an average patent's knowledge diffuses over time. We do this by measuring the time elapsed between when cited and citing patent pairs are published. We then plot the mean distance scores by week after publication. Figure 6 shows the knowledge diffusion curve showing that as time passes citations come from further and further afield.

INSERT FIGURE 6 HERE

Plotting the mean distance between co-citing patents demonstrates steady changes to knowledge flows. Figure 7 shows that the mean distance between co-citing patents has decreased, meaning that the level of similarity between patents that cite the same prior art reference has increased in recent decades. However, we also see that the standard deviation of co-citing distance has increased during the same period. This suggests that, while the average distance between co-citing patents is decreasing, there has been a concomitant increase in the variety of co-citing prior art.

INSERT FIGURE 7 HERE

5. Discussion

In the interest of brevity, we have presented an abbreviated introduction to and demonstration of these measures in this abstract. Our four measures of knowledge translation, integration, diffusion and scope provide novel empirical insight into the research environment that traditional citation analysis measures do not. By accounting for the structure of knowledge space, we demonstrate changes in the way researchers search for and combine knowledge, and how their research goes on to influence future work. We observe changing tendencies both in our measures of backward citation distance and forward citation distance. By measuring these changes over time, citation distance measures provide insight into the way the innovation system has evolved in recent decades. More nuanced citation metrics, like those proposed here, offer great potential in improving impact measures, and research studying knowledge diffusion and the research process.

Tables:

	Forward Citations	**Backward Citations**
Citing/Cited Distance	Knowledge translation	Knowledge diffusion
Co-Cited or Co-Citing Distance	Knowledge integration	Knowledge scope

Table 1. Four citation distance measures.

Figures:

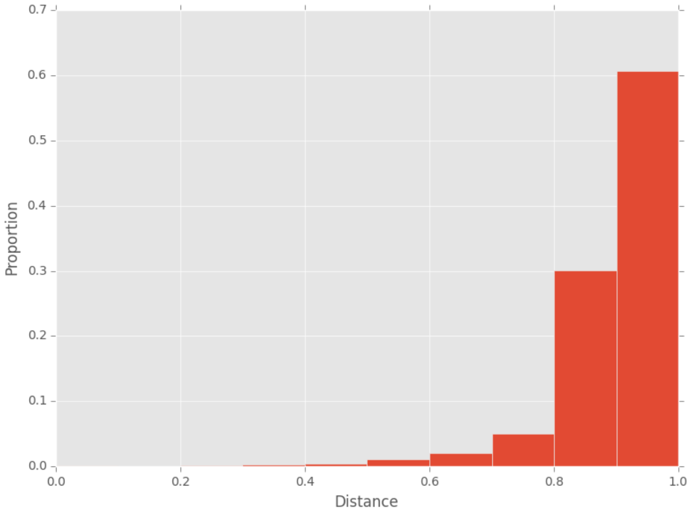

Figure 1: The distribution of distances between randomly paired patents (mean = 0.87).

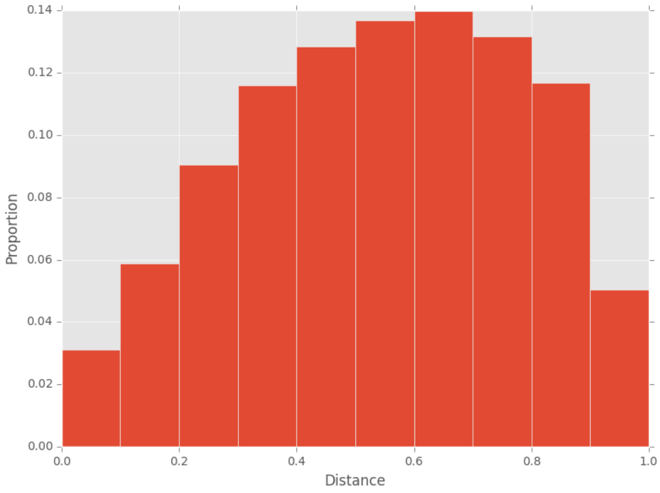

Figure 2: Distribution of citing/cited distances (mean = 0.54).

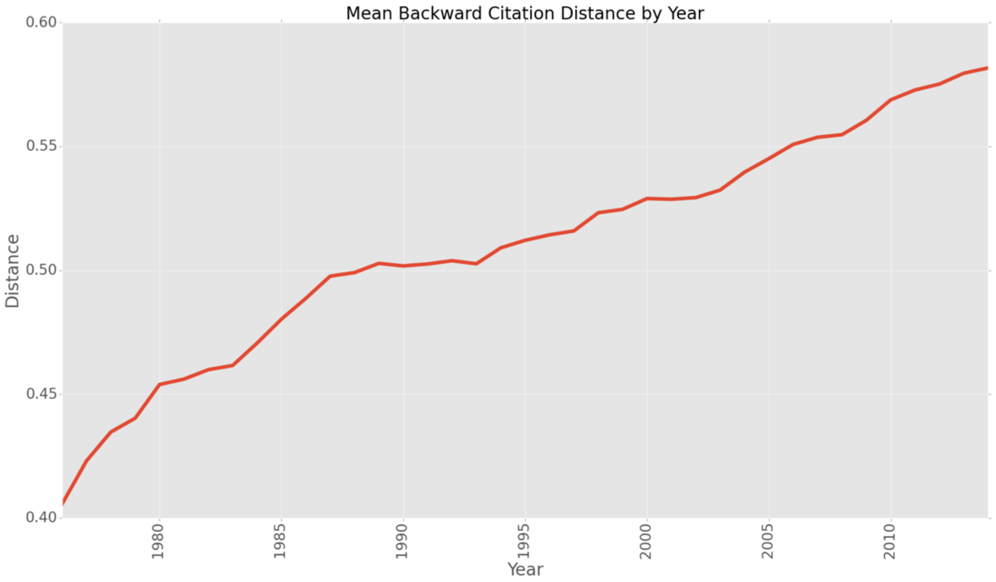

Figure 3: Mean backward citation distance from 1976 to 2014.

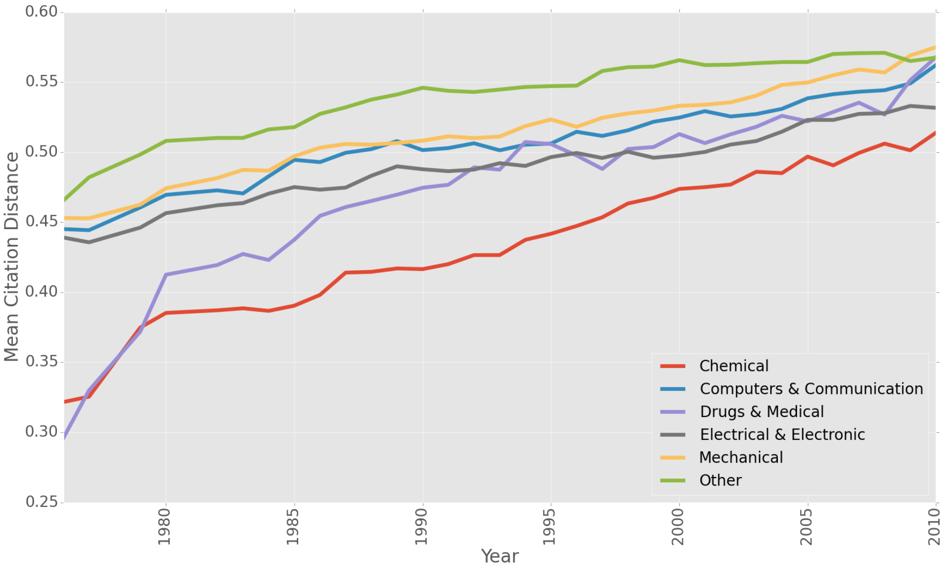

Figure 4: Mean backward citation distance by research field.

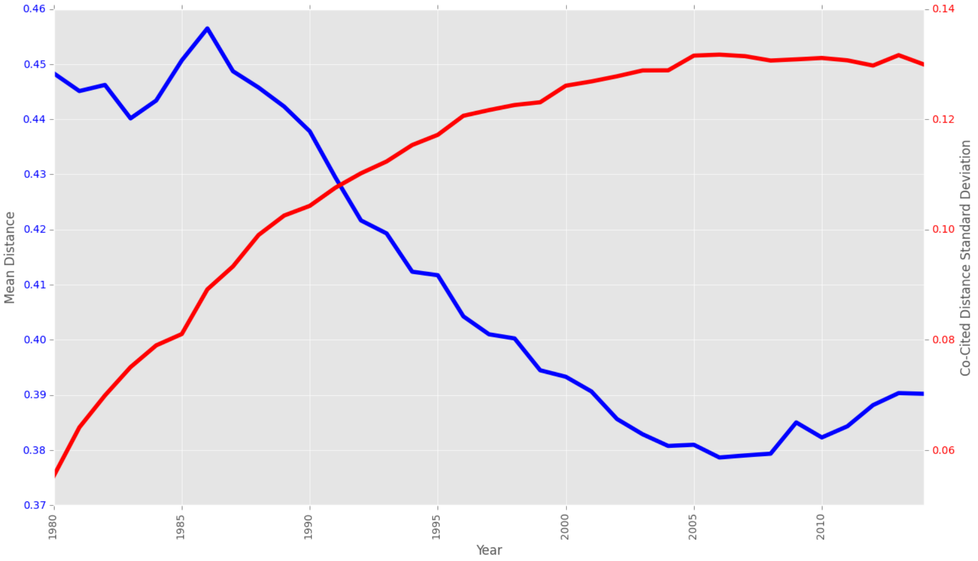

Figure 5: Knowledge integration trends.

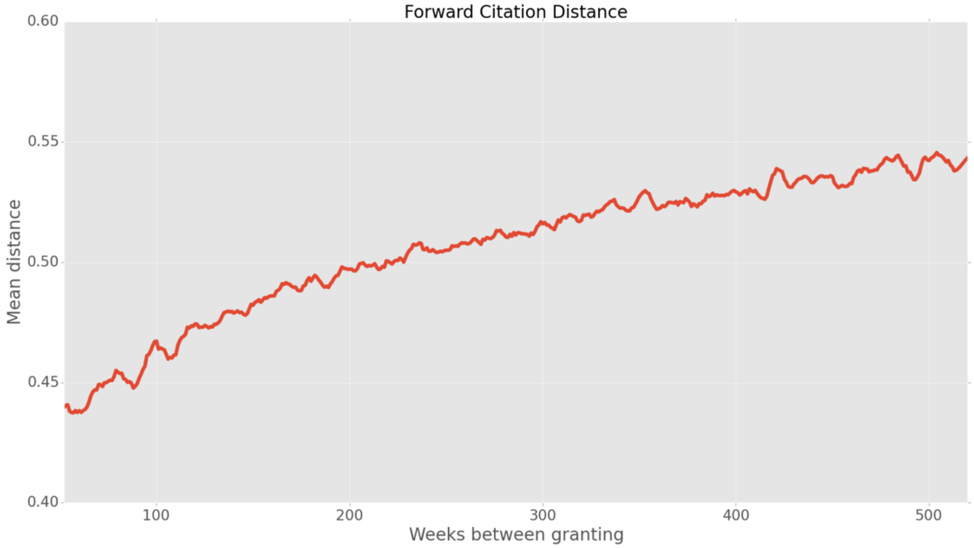

Figure 6: Mean forward citation distance by time between publications of citing and cited patent.

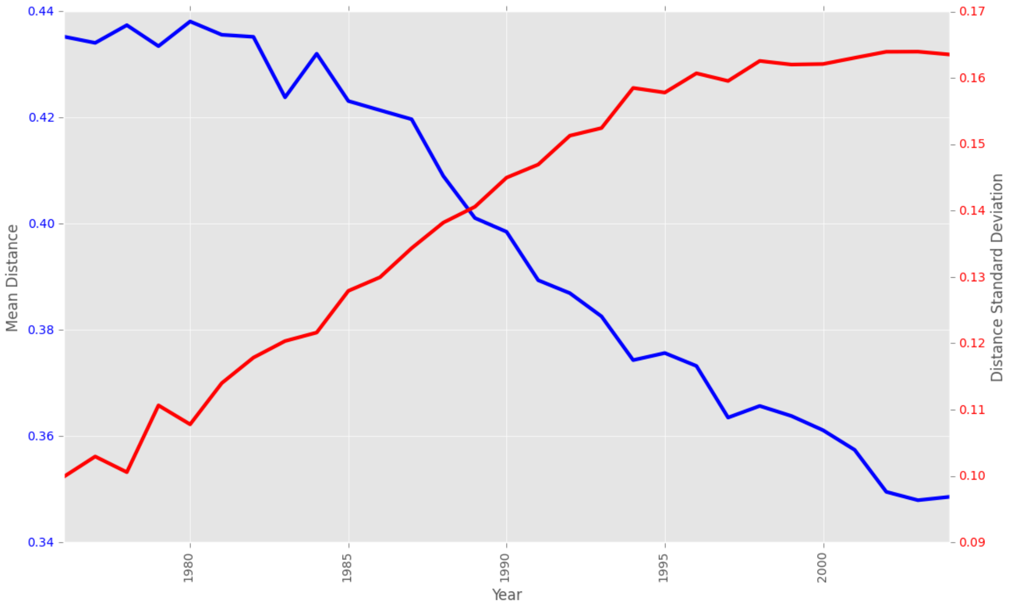

Figure 7: Mean co-citing distance, and the patent-wise standard deviation of co-citing distance by year for citations received within the first 10 years post-grant.

Reference List:

Abramo, G., D'Angelo, C. A., & Caprasecca, A. (2009). Allocative efficiency in public research funding: Can bibliometrics help? *Research Policy*, *38*(1), 206–215. doi:10.1016/j.respol.2008.11.001

Almeida, P., & Kogut, B. (1999). Localization of Knowledge and the Mobility of Engineers in Regional Networks. *Management Science*, *45*(7), 905–917. doi:10.1287/mnsc.45.7.905

Börner, K., Penumarthy, S., Meiss, M., & Ke, W. (2013). Mapping the diffusion of scholarly knowledge among major U.S. research institutions. *Scientometrics*, *68*(3), 415–426. doi:10.1007/s11192-006-0120-2

Catalini, C., Lacetera, N., & Oettl, A. (2015). The incidence and role of negative citations in science. *Proceedings of the National Academy of Sciences*, *112*(45), 13823–13826. doi:10.1073/pnas.1502280112

Chen, P., Xie, H., Maslov, S., & Redner, S. (2007). Finding scientific gems with Google's PageRank algorithm. *Journal of Informetrics*, *1*(1), 8–15. doi:10.1016/j.joi.2006.06.001

Chubin, D. E., & Moitra, S. D. (1975). Content Analysis of References: Adjunct or Alternative to Citation Counting? *Social Studies of Science*, *5*(4), 423–441.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990). Indexing by latent semantic analysis. *JAsIs*, *41*(6), 391–407.

Fleming, L. (2001). Recombinant Uncertainty in Technological Search. *Management Science*, *47*(1), 117–132. doi:10.1287/mnsc.47.1.117.10671

Fleming, L., & Sorenson, O. (2001). Technology as a complex adaptive system: evidence from patent data. *Research Policy*, *30*(7), 1019–1039.

Fleming, L., & Sorenson, O. (2004). Science as a map in technological search. *Strategic Management Journal*, *25*(8-9), 909–928.

Foster, J. G., Rzhetsky, A., & Evans, J. A. (2015). Tradition and Innovation in Scientists' Research Strategies. *American Sociological Review*, *80*(5), 875–908. doi:10.1177/0003122415601618

Garfield, E. (1972). Citation Analysis as a Tool in Journal Evaluation. *Science*, *178*(4060), 471–479.

Garfield, E. (1979). Is citation analysis a legitimate evaluation tool? *Scientometrics*, *1*(4), 359–375. doi:10.1007/BF02019306

Glänzel, W., & Moed, H. F. (2002). Journal impact measures in bibliometric research. *Scientometrics*, *53*(2), 171–193. doi:10.1023/A:1014848323806

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(46), 16569–16572. doi:10.1073/pnas.0507655102

Holden, G., Rosenberg, G., & Barker, K. (2005). Bibliometrics: A potential decision making aid in hiring, reappointment, tenure and promotion decisions. *Social Work in Health Care*, *41*(3-4), 67–92. doi:10.1300/J010v41n03_03

Jaffe, A. B., Trajtenberg, M., & Henderson, R. (1993). Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations. *The Quarterly Journal of Economics*, *108*(3), 577–598.

Lee, Y.-N., Walsh, J. P., & Wang, J. (2015). Creativity in scientific teams: Unpacking novelty and impact. *Research Policy*, *44*(3), 684–697. doi:10.1016/j.respol.2014.10.007

Leydesdorff, L. (2009). How are new citation-based journal indicators adding to the bibliometric toolbox? *Journal of the American Society for Information Science and Technology*, *60*(7), 1327–1336. doi:10.1002/asi.21024

Leydesdorff, L., & Bornmann, L. (2011). How fractional counting of citations affects the impact factor: Normalization in terms of differences in citation potentials among fields of science. *Journal of the American Society for Information Science and Technology*, *62*(2), 217–229. doi:10.1002/asi.21450

March, J. G. (1991). Exploration and Exploitation in Organizational Learning. *Organization Science*, *2*(1), pp. 71–87.

Moed, H. F. (2010). Measuring contextual citation impact of scientific journals. *Journal of Informetrics*, *4*(3), 265–277. doi:10.1016/j.joi.2010.01.002

Moravcsik, M. J., & Murugesan, P. (1975). Some Results on the Function and Quality of Citations. *Social Studies of Science*, *5*(1), 86–92.

Nelson, R., & Winter, S. (1982). *An evolutionary theory of economic change*. Cambridge MA: Harvard University Press.

Rosell, C., & Agrawal, A. (2009). Have university knowledge flows narrowed?: Evidence from patent data. *Research Policy*, *38*(1), 1 – 13. doi:http://dx.doi.org/10.1016/j.respol.2008.07.014

Rosenkopf, L., & Nerkar, A. (2001). Beyond Local Search: Boundary-Spanning, Exploration, and Impact in the Optical Disk Industry. *Strategic Management Journal*, *22*(4), 287–306.

Schumpeter, J. A. (1939). *Business cycles* (Vol. 1). Cambridge Univ Press.

Segalla, M. (2008). Publishing in the right place or publishing the right thing: journal targeting and citations' strategies for promotion and tenure committees. *European Journal of International Management*, *2*(2), 122–127. doi:10.1504/EJIM.2008.017765

Singh, J. (2005). Collaborative Networks as Determinants of Knowledge Diffusion Patterns. *Management Science*, *51*(5), 756–770.

Small, H. (2011). Interpreting maps of science using citation context sentiments: a preliminary investigation. *Scientometrics*, *87*(2), 373–388. doi:10.1007/s11192-011-0349-2

Stuart, T. E., & Podolny, J. M. (1996). Local search and the evolution of technological capabilities. *Strategic Management Journal*, *17*(S1), 21–38.

Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical Combinations and Scientific Impact. *Science*, *342*(6157), 468–472.