Patent Citation Data in Social Science Research: Overview and Best Practices

Adam B. Jaffe
Motu Economic and Public Policy Research, Wellington 6011 New Zealand; Queensland University of Technology, and Te Pūnaha Matatini Centre of Research Excellence.
E-mail: adam.jaffe@motu.org.nz

Gaétan de Rassenfosse
Ecole polytechnique fédérale de Lausanne, College of Management of Technology, CH-1015 Lausanne, Switzerland. E-mail: gaetan.derassenfosse@epfl.ch

The last 2 decades have witnessed a dramatic increase in the use of patent citation data in social science research. Facilitated by digitization of the patent data and increasing computing power, a community of practice has grown up that has developed methods for using these data to: measure attributes of innovations such as impact and originality; to trace flows of knowledge across individuals, institutions and regions; and to map innovation networks. The objective of this article is threefold. First, it takes stock of these main uses. Second, it discusses 4 pitfalls associated with patent citation data, related to office, time and technology, examiner, and strategic effects. Third, it highlights gaps in our understanding and offers directions for future research.

"Knowledge flows [. . .] are invisible; they leave no paper trail by which they may be measured and tracked, and there is nothing to prevent the theorist from assuming anything about them that she likes."

Paul Krugman (1991)

Introduction

Eugene Garfield is one of the pioneers of the study of citation data. In his 1955 article, Garfield proposes to build a citation index for scientific articles in order to make it possible for "the conscientious scholar to be aware of criticisms of earlier articles." He further explains, "even if there were no other use for a citation index than that of minimizing the citation of poor data, the index would be well worth the effort required to compile it" (p. 108). It turns out that citation indices have been used in a variety of ways and for a variety of purposes. Two of the most notable uses are to assess the attributes of the idea embedded in a scientific article and to track its diffusion through time, space and technology domains. In fact, Garfield (1955) foresaw these two uses as he described the citation index as an "association-of-ideas index" (p. 108) and as he explained that the citation index may "help the historian to measure the influence of the article—that is, its 'impact factor'" (p. 111).

Although the analogy with the broader field of bibliometrics may seem obvious, patent citations differ from scientific citations in substantial ways. Citations in patents are the results of a highly mediated process that involves multiple parties: the inventor, the patent attorney, and the patent examiner (Meyer, 2000). These parties have different incentives for citing publications and may do so at different times and in different sections of the patent document (Cotropia, Lemley, & Sampat, 2013). Much of the empirical research relies on U.S. citations, but there are important differences across jurisdictions in citation rules and practice.[1] This creates interesting opportunities for research on non-U.S. data, but also suggests a degree of caution in thinking about the global implications of results based solely on U.S. data.

The widespread use of patent citations in social science research can be traced to the availability of patent statistics in digitally readable form in the late 1970s.[2] Zvi Griliches (1979), in his important manifesto for research on R&D and productivity growth, suggested that the frequency with

Received August 14, 2015; revised January 4, 2016; accepted January 31, 2016

 DOI: 10.1002/asi.23731

which patents from different industries cite each other could be used as a measure of the technological proximity of industries. An early strand of research on patent citations was the work of Francis Narin and his associates at CHI Research, Inc. (Carpenter & Narin, 1983; Carpenter, Narin, & Woolf, 1981; Narin & Noma, 1985; Narin, Noma, & Perry, 1987). An influential early demonstration of the potential utility of patent citation data in economic research was the PhD research of Griliches's student Manuel Trajtenberg (Trajtenberg, 1990a, 1990b). The use of patent citation data has grown dramatically over the last two decades, as illustrated in Appendix A.

What makes citations potentially useful is that they convey information about the cumulative nature of the research process, as well as information about the consequences. Although some inventors and research organizations pursue patents for motives of prestige or internal tracking of research success, most patent applications are made with the goal of securing commercial advantage, or at least preserving options for pursuit of commercial advantage. Another virtue of patent data for social science research is that patents reside in a nonmarket-based technological classification system, allowing one to place patents, inventors, and organizations in technology space in a way that is not derived from sales or other economic data that one may be trying to relate to invention.[3] Furthermore, the classification scheme is hierarchical so that technology categories can be very fine or relatively broad as desired. This feature, and others, has been combined with patent citation data to provide powerful indicators.

This article provides an overview of the major uses of such data and the issues that arise in such research. Other authors have previously discussed the use of patent statistics in social science research (e.g., Griliches, 1990; Lerner & Seru, 2015), and Gay and Le Bas (2005) provide a brief overview of the use of patent citations to measure invention value and knowledge flows. However, we are not aware of a broad survey on the use of patent citation data.[4] In order to identify the articles to include in this survey, we started from a limited number of references that we were aware of and complemented those using a keyword-based search on Google Scholar. We then expanded this core of references by looking at cited and citing references. Ultimately, we kept the most influential articles, either in terms of the number of citations received or in terms of relevance of the findings. The majority of articles are published in economics, management, and information science journals.

Conceptually, we classify research using patent citations into two broad groups. One research line uses a variety of citation-based statistics to characterize the inventions, in terms of the magnitude and nature of their impact, as well as the nature and magnitude of the departure that they represent relative to the existing pool of knowledge. This work is discussed in the next section. The other research line focuses on the citations themselves, using them as proxies for knowledge linkages across inventors in order to explore the nature of knowledge flows and the factors that affect those flows. This research is discussed first with regard to relatively simple metrics of knowledge flow, and then with respect to attempts to map interactions in a more complex network framework. We then provide some brief comments on practical difficulties and pitfalls in using citation data. The last section concludes with opportunities for future research.

Citations as an Indicator of Invention Attributes

There is no agreed-upon model of inventions and the inventive process, which leads to some ambiguity in how citation metrics are interpreted. Nonetheless it is possible to identify two broad aspects of the process that underlie citation-based inferences. First, we can think of all possible technologies as mapping onto a high-dimensional technology space, such that a given invention can be located in that space, and a patent represents the right to exclude others from marketing products that impinge upon a specified region (or regions) of that space. Second, the invention process is cumulative, that is, inventions build on those that came before and, in turn, facilitate those that come after. In this "geometric" interpretation, the patent claims delineate the metes and bounds of the region of technology space over which exclusivity is being granted, whereas the citations indicate previously marked-off areas that are in some sense built upon by or connected to the invention being granted.

Thus the citations that appear in a patent (its "backward" citations) inform us about the technological antecedents of the patented invention. A patent that contains many citations corresponds to an invention with many antecedents; a patent whose citations are to technologically diverse previous patents has diverse antecedents; a patent whose citations are to old patents corresponds to an invention with old antecedents, and so forth. Conversely, the citations received by a patent from subsequent patents ("forward" citations) inform us about the technological descendants of the patented invention. A patent that is never cited was a technological dead end. A patent with many or technologically diverse forward citations corresponds to an invention that was followed by many or technologically diverse descendants.

Note that the discussion so far is entirely definitional. We have said nothing about the possibility of causal connections between these different attributes of inventions, or between any of these attributes and the private or social value of the invention. Ultimately, we are interested in whether, for example, patents with relatively few technological antecedents are more or less likely to spawn multiple lines of research or whether patents that generate many or diverse technological descendants correspond to inventions that generate large social benefits. It is in large part to be able to say something about these questions that citation metrics have been developed. In a very broad sense, citation analysis is predicated on an expectation that the extent and nature of an invention's antecedents tells us something about the novelty or "radicalness" of the invention, and the extent and nature of its descendants tell us something about both its

DOI: 10.1002/asi

technological impact and its economic value. But different authors propose or use different characterizations of citation information to elucidate these ideas.

In practice, writers are not always clear on the underlying concept that a given metric is intended to measure, and given metrics are used in different contexts as proxies or indicators for different concepts. In some cases, researchers *postulate* a relationship between a given citation metric and an underlying concept, and then test hypotheses about the concept taking that relationship as a given. In other cases researchers attempt explicitly to *validate* the extent to which a given metric reflects a particular underlying conceptual attribute of inventions. We will consider these different approaches below in the context of specific articles, but for expositional purposes it is useful to consider five broad categories of approaches:

- Counts of forward citations as an indicator of subsequent technological impact;
- Counts of backward citations as an indicator of the extent of reliance on previous technology;
- Characterization of both backward and forward citations in terms of technological diversity and technological distance;
- Examination of references to nonpatent literature as an indicator of science linkage; and
- Use of citations as an indicator for private and social value.

We consider each category in turn.

Forward Citations and Technological Impact

Using the number of forward citations as a measure of technological impact of a patented invention can be motivated by direct analogy to the larger and pre-existing bibliometric literature starting with Garfield (1955). Nonetheless, Trajtenberg, Henderson, and Jaffe (1997) undertook to demonstrate the validity of this (and other) metrics by comparing the citation rate to university patents and corporate patents, based on a maintained assumption that university patents are more "basic" and hence have, on average, greater technological impact. To incorporate the cumulative nature of invention into the metric, they proposed that the importance of an invention be characterized by the number of forward citations received, plus a fractional weight multiplied by the number of citations received by those citing patents. That is, important patents are those that are cited a lot, and are cited by patents that are themselves relatively highly cited.[5] The authors showed that importance by this definition is, indeed, higher for university patents than for corporate patents, using a sample of patents assigned to U.S. corporations, matched by patent class and grant date to patents assigned to U.S. universities. In addition, they discuss qualitatively the highest-importance patents in their sample, and argue that the citing patents can be seen as technological descendants, and these highly "important" patents are, indeed, subjectively very important in their respective fields.

More recently, taking advantage of improvements in computing power, scholars have taken into account the *whole stream* of citations. For example, Lukach and Lukach (2007) have proposed computing importance by the PageRank score of patents. This method is directly inspired from Google's "random surfer" model and takes into account the fact that different citations weigh differently depending on the importance of the citing documents (Brin & Page, 1998). However, the authors are not able to validate their ranking using external measures such that the conditions under which the PageRank method is more appropriate than a straightforward citation count are unclear. This approach is a natural extension of earlier work, and begins to move this line of analysis towards the "innovation network" formulation discussed later in the text.

Albert, Avery, Narin, and McAllister (1991) provide a validation study of the use of forward citations as an indicator of impact. They reported a strong correlation between the citation intensities of 77 Kodak silver halide patents and expert evaluations of technical impact and importance of the patents. Narin (1995) showed that patents that have attained the legal status of pioneering patents in the United States, as well as other prominent patents appearing in such patent office publications as "Hall of Fame" patents, are very highly cited. Czarnitzki, Hussinger, and Schneider (2011) relate a group of "wacky" patents to control groups and test the extent to which commonly used metrics are able to identify wacky patents from patents in the control group. Wacky patents are selected by an employee of the World Intellectual Property Organization "for their futile nature, as they do not involve a high-inventive step or only marginally satisfy the 'non obviousness' criterion" (p. 131). They find that the number of forward citations is a good predictor of importance. However, other measures such as originality and generality (discussed below) were higher for wacky patents. Another interesting confirmation of patent citations as indicative of technological impact is Benson and Magee (2015). They identify 28 "technological domains" (e.g., "Solar Photovoltaics" or "Genome Sequencing") in which it is possible to identify a specific metric of the technological state of the domain (e.g., watts/$ for Solar Photovoltaics). They take the exponential rate of improvement of these metrics across domains and across time as the dependent variable in regressions on various citation metrics of patents in the technology domain. They find that forward citations are positively related, and the average age of backward citations negatively related, to the rate of improvement of the technology over the subsequent 10-year period.

Backward Citations and Reliance on Previous Technology

Although it seems clear that important inventions generate more forward citations, the opposite may hold for backward citations. That is, more trivial inventions are more extensively rooted in what has come before, whereas more basic inventions are less incremental in nature and thus have fewer identifiable antecedents (Trajtenberg et al., 1997). Another way to think of this is that a patent will, to some

DOI: 10.1002/asi

extent, tend to cite other patents all the way back along the inventive trajectory upon which it lies. Patents that are near the beginning of a trajectory are in this sense more basic, and may be expected to make fewer backward citations because they have less historical background.

Empirical evidence is rather inconclusive. Trajtenberg et al. (1997) find that university patents (presumably more important than the average patent) do make fewer citations and cite patents that are themselves less highly cited. However, von Wartburg, Teichert, and Rost (2005) provide a different view. They correlate a measure of backward citations with expert ratings on the technological value added (in the form of technical scoring tables) of 107 patents related to four strokes internal combustion engines. Their backward citations measure counts first and second-generation's citations received. They obtain a statistically significant correlation coefficient of 0.38, implying that patents with higher technological value added build on more references. Liu et al. (2011) propose a more in-depth analysis of backward references and patent value. They correlate the number of backward references with the probability that a patent will stand up in court and find a statistically strong positive association. Overall, it is unclear whether the number of backward citations captures patent importance.

Technological Distance and Diversity

As noted, one of the basic virtues of patent data is that they provide a nonmarket-based technological classification system for inventions. Looking at the way in which citations span the technology space defined by the classification scheme is a natural way to characterize the technological complexion of both an invention's roots and its impacts. Broadly speaking, there are two major aspects to be considered, whether looking forward or backward. One is pure distance: how technologically different are the patents connected by a citation link. For example, does a drug patent cite other patents for compounds in the same chemical class, or patents on other chemicals, or mechanical or electronic patents? The other is breadth or diversity: independent of whether that drug patent generally cites other patents that are close to or far from *itself*, are they all bunched together in technology space, or are they dispersed far from each other?

Trajtenberg et al. (1997) implement a measure of technological distance using a three-level representation of the USPTO patent classification system. The lowest level used is the three-digit original patent class (e.g., Electric lamp and discharge devices); the next level is the set of two-digit categories (e.g., Electrical Lighting); the highest level is six very broad fields (e.g., Electrical and Electronic). The authors axiomatically set two patents in the same patent class at distance 0; two that are in different classes but the same category at distance 0.33; two that are in different categories but the same broad field as distance 0.66; and two that are not even in the same field as distance 1. They then calculate the average distance over both forward and backward citations for each patent in the university and corporate samples. As expected, they found that the forward citations received by university patents came, on average, from farther away in technology space, although the difference was small and not always statistically significant. For backward citations, there was no consistent pattern, that is, university patents did not systematically cite earlier patents that were, on average, technologically more distant by this metric.

To measure technological dispersion or diversity, Trajtenberg et al. (1997) proposed 1 minus the Herfindahl-Hirschman Index (HHI) of concentration of the citations across patent classes, that is, 1 minus the sum of squared shares of citations in each class. This metric is equal to zero if all citations are in the same class, and it approaches unity as the citations are spread thinly across all classes. The authors dubbed this metric of diversity "generality" when applied to forward citations, and "originality" when applied to backward citations.[6,7] They conjectured that both measures should be larger for more basic inventions, and therefore expected to be larger for university patents than for corporate patents. This hypothesis was borne out in the data for generality measure, but not for originality.

A concept related to generality is that of "General Purpose Technology" or GPT. GPTs are conceived as technologies that subsequently connect to many different application or development technologies to allow multiple lines of technology innovation and diffusion. Frequently mentioned examples are the electric motor in the late 19th and early 20th centuries, and digital information technology in the late 20th century. Hall and Trajtenberg (2006) use data from a selected sample of 780 most highly cited patents that were granted by the USPTO in the years 1967–1999 to construct generality, number of citations, and patent class growth, for both cited and citing patents, intended to identify GPTs in their early stages. The article finds that highly cited patents differ in almost all respects from the population of all patents (they take longer to be issued; have twice as many claims; are more likely to have a U.S. origin; are more likely to be assigned to a U.S. corporation; are more likely to have multiple assignees; have on average higher citation lags; have a higher generality; are in patent classes that are growing faster than average). The article concludes that the identified measures, although promising, give contradictory messages when taken separately and that it is not obvious how to combine those measures to choose a sample of GPT patents.[8] The fundamental difficulty is that we don't have measures of how general-purpose a technology is other than broad conceptions of GPT technologies. Thus, although it seems plausible that general-purposeness would be reflected in citation patterns, it is hard to pin such patterns down or test their validity.[9]

Youtie, Iacopetta, and Graham (2008) found that nanotechnology patents from 1990–1993 were more general than computer patents and much more general than drug patents, and interpret this result as evidence that nanotechnology is an emerging GPT. Moser and Nicholas (2004), however, found that electricity patents from the 1920s were less

DOI: 10.1002/asi

general and less highly cited than chemical and mechanical patents from the same period, suggesting that the relationship between the characteristics that make a technology a GPT and other characteristics of inventions is complex.

Another concept related to technological distance and diversity is that of a "radical" or "breakthrough" invention. Ahuja and Lampert (2001) propose that radical inventions are simply the top 1% of patents ranked on citations received in a given year. Dahlin and Behrens (2005) adopt a more sophisticated approach. They conceive a "radical" invention within a given technology domain (tennis rackets, in their application) to be one that recombines previous technology elements in a new and different way, but which is then imitated and so spawns subsequent patents that combine technology elements in a manner substantially similar to the radical invention. They construct a measure of the "overlap" in the respective sets of patents cited by two different patents, and show that the radical inventions (oversized and wide-body rackets, in their application) had little overlap with previous or contemporary patents, but significant overlap with patents that came after.

Linkage to Science

As discussed, patents contain references to nonpatent documents, the overwhelming majority of which are scientific articles. On this basis, the number of nonpatent backward citations made by a patent, or the fraction of backward citations that these nonpatent citations represent, has been explored as a metric of the closeness of linkage between an invention and scientific research.[10]

Collins and Wyatt (1988) looked at citations to scientific articles from 366 genetics patents granted from 1980 to 1985, in order to trace linkages from basic research to genetics technology. The United States had the highest number of articles cited in patents, followed by the United Kingdom, Japan, Germany, and France. These figures were compared to the total output of genetics articles for those countries, showing some differences, which were interpreted as indicating that the United Kingdom produced more articles that were useful in developing patented technology than Germany, France or Japan. The number of citations from patents received per article was highest for the United Kingdom, followed by the United States and Germany.

Callaert, Van Looy, Verbeek, Debackere, and Thijs (2006) characterizes nonpatent references in a sample of patents at the USPTO and the European Patent Office (EPO) from 1991–2001. Nonpatent references are found in 34% of USPTO patents and 38% of EPO patents, comprising about 17% of all references (patent and nonpatent combined). For both the USPTO and EPO, more than half of nonpatent references are journal references. Of the remaining nonpatent references, many can be considered scientific in the broader sense (as they consist of conference proceedings, books, databases or other nonjournal scientific publications), or technology related. The article reports that at the USPTO at least 42% of nonjournal nonpatent references can be considered scientific in broader sense, and 40% relate to technological information. For the EPO sample these figures are 77% and 20%, respectively.

Tijssen (2002) provides a note of caution on the use of nonpatent references. He found no relationship between the number of nonpatent references and the inventor-reported dependence on science in a small (<100) sample of Dutch patents from 1998–99. Li, Chambers, Ding, Zhang, and Meng (2014) qualify this finding. They argue that nonself-citations to scientific articles are a noisy measure of science linkage but that applicant self-citations to scientific articles are indeed informative of science linkage. Roach and Cohen (2013) matched patent citations to survey reports from R&D lab managers in the United States, with particular focus on the extent to which patent citations capture knowledge flows to commercial R&D from publicly funded research. They find that patent citations reflect codified knowledge. However, citations miss the reliance on private and contract-based science, as well as basic research. (The discussion in the section on citations as a measure of knowledge flows considers further whether nonpatent references are an indicator of science dependence.)

Economic Value

As noted earlier, the (public or private) economic value of an invention is a distinct concept from its technological impact. Citations are, first and foremost, an indicator of technological impact. But it turns out that forward citation intensity is, in fact, correlated with economic value. There are, however, several different concepts of economic value. First, we can in principle think of the (gross) social value of an invention, that is, the total producers' and consumers' surplus associated with its use. In some cases this gross social value may be much greater than the *net* value, for which we would subtract off the lost rents that may be suffered by previous technologies made wholly or partially obsolete. The gross social value is greater than the *private* value, that is, the value to the owner of a patented invention; the net social value may be either greater or less than the private value, depending on the magnitude of the "rent stealing" effect. For any of these concepts, we can distinguish the value of the *invention* and the value of the *patented invention*, which differ by the value of the legal protection afforded by the patent grant. In practice, these different value concepts may or may not be distinguishable, and proxies for value are often used whose mapping onto these different value concepts may be ambiguous.

An early strand of research on citations and economic value was the work of Francis Narin and his associates seeking to develop indicators based on patent data of companies' competitiveness or technological strength. Carpenter et al. (1981) showed that inventions identified in The Industrial Research Institute IR100 awards are much more highly cited than a random sample of matched patents. Narin et al. (1987) found that the average citation frequency of a company's patent portfolio was associated with increases in

DOI: 10.1002/asi

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.