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

firms' profits and sales among publicly traded pharmaceutical companies.

Trajtenberg (1990b) calculated the social welfare gains associated with successive generations of Computed Tomography (CT) scanners by estimating hedonic demand functions for the attributes. He then showed that the number of citation-weighted patents associated with each generation was statistically predictive of the magnitude of welfare gains, while the raw or unweighted count of patents was not correlated with surplus (sample of about 500 patents). This suggests that the gross social value of these inventions is associated with the citation intensity of the associated patents. Interestingly, the unweighted patent counts were correlated with the level of R&D expenditure. He interpreted these findings as suggesting that the number of patents is associated with the magnitude of research effort, but not indicative of research success. Counting citation-weighted patents then combines the scale of effort with a measure of such success and yields a measure of effective research output.

Moser, Ohmstedt, and Rhode (2014) identified specific improvements in hybrid corn and gathered data on the magnitude of the yield improvement they allowed. They interpret this as measuring the "inventive step" associated with the patent, but as the measurement is in the use domain rather than strictly in the technology domain it seems more closely related to social value than to inventive step, *per se*. They found that there is, indeed, a strong correlation between yield improvements and citation intensities. Interestingly, they find that there are a small number of early patents that are routinely cited in almost all patents in the field. Excluding these citations enhances the correlation between yield and citation frequency.

Hall, Jaffe, and Trajtenberg (2005) consider the relationship between citation intensity and the private value of patents by relating citation-weighted patents to the market value of the firm. They confirm that citation weighting greatly improves the information content of patent counts in terms of predicting market value. In addition, they find that citations from future patents assigned to the same firm as the original patentee have a larger associated market value than citations from others.[11] They also find that a disproportionate share of the value associated with patents is associated with a very small number of highly cited patents. Finally, they find that forward citations are associated with increases in market value at the time a patent is initially granted, suggesting that to a significant extent market participants can anticipate the eventual value of inventions at this early stage, and those expectations are (on average) then confirmed by subsequent citations.

Lanjouw and Schankerman (2001) provide indirect evidence of the relationship between citations and value, by assuming that patents that are litigated are, on average, more valuable than those that are not, and comparing the citation patterns of litigated patents with a control sample of nonlitigated patents. They find that the probability of litigation rises with the number of claims and the number of forward citations per claim, whereas declining with the number of backward citations per claim. Allison, Lemley, Moore, and Trunkey (2003) undertake a similar approach. Consistent with expectations, they find that litigated patents are more highly cited. Interestingly, they find that litigated patents also have more backward citations.

Harhoff, Scherer, and Vopel (2003) obtained estimates from patent holders of the private value of 772 patents with a 1977 German priority date, and that were maintained to full term. They then examined how that reported value correlated with publicly observable indicia of patent value, including patent citations (and also the number of four-digit IPC codes and family size). They found that both the number of forward citations and the number of backward references to the patent literature are significantly correlated with patent value (see also Harhoff, Narin, Scherer, & Vopel, 1999). Interestingly, they also found that the number of citations made to nonpatent literature was predictive of value, particularly in drug and chemical patents. They note that the predictive value of backward citations (both patent and nonpatent) is quite useful, as this information is available at time of patent grant, while forward citations must be awaited.[12] It is unclear theoretically why backward citations are predictive of value. For nonpatent references, it is plausible that in some fields inventions linked to science are less incremental and hence more valuable. For backward patent citations, it may reflect some tendency for bigger, more complex patents to make more backward citations and also be more valuable on average. In addition, the positive correlation between the number of backward citations and value may simply arise from the fact that applicants have stronger incentives to search for prior art for more important patents (Sampat, 2010).

Gambardella, Harhoff, and Verspagen (2008) undertook a similar survey of inventors listed in patent applications at the EPO. They found that the number of forward citations is by far the best predictor of reported value, but that the fraction of the variance in reported value explained by any or all of the metrics was relatively low, consistent with a view of citation-weighted patents as an indicator of value, but one with substantial noise.

Nicholas (2008) looked at patents granted to U.S. corporations between 1910 and 1939, and identified the citations to those historical patents from the period 1976–1999. He found that about 15% of the patents from the 1910s received at least one citation from the recent patents, rising to almost 30% for those from the 1930s. He then goes on to show that citation-weighted patents constructed in this way are correlated with firm market value. Thus, patents that are still cited after 40 to 60 years are more valuable than those that are not. What we cannot know from this exercise (since early citations have not been captured) is the extent to which valuable patents are simply more highly cited at all lag durations, or whether there is greater persistence in the sense that the rate of obsolescence is lower.

Bessen (2008) related the value of patents, as indicated by both renewal information and firm financial data, to a

DOI: 10.1002/asi

number of patent characteristics, including forward citations received. He estimated that each additional citation is associated, on average, with an increase in value of about 1%. Nonetheless, the relationship is very noisy, so that even among very highly cited patents, a significant fraction appears to be of little value; 37% of the patents in the top decile in citation intensity from 1991 were not renewed.

Recent work by Abrams, Akcigit, and Popadak (2013) also suggest an overall positive correlation between forward citations and patent value, but with an inverted-U-shaped relationship in which value falls at high citation rates. This finding is provocative, but it is unclear how robust it is, given the highly selected nature of the sample and the fact that the value of individual patents was estimated as the value of patent portfolios divided by the number of patents in the portfolio.

The next section moves away from work focused on citations as indicators of invention characteristics, and discusses the use of citation data to capture geographic and temporal dimensions of the innovation process.

Citations as an Indicator of Knowledge Flows

Geographic Dimension of Knowledge Flows

Jaffe, Henderson, and Trajtenberg (1993) took on the challenge identified by Krugman (1991) on the invisibility of knowledge flows. They suggested that patent citations could be used as a kind of "article trail" that could allow knowledge flows to be measured and tracked. They took a sample of patents from universities, large firms and other firms, and identified all of their citations. They then found, for every citing patent, a corresponding "control" patent, issued at the same time and in the same primary U.S. patent class as the citing patent, and compared the frequency with which citing patents were geographically proximate to the cited patents with the frequency with which the control patents were proximate. Looking at metropolitan statistical areas, states and countries, and eliminating citations that are "self-citations" from the same firm, they showed that citations are indeed more likely to be proximate. For example, at the level of metropolitan areas, 7–9% of citations (depending on the nature of the cited patents) were from the same area, while only 1–4% of the control patents were, and the differences were highly significant statistically.

Thompson and Fox-Kean (2005) criticize the Jaffe, Henderson, and Trajtenberg methodology. They argue that selecting control patents based on the primary patent class of the citing patents is too rudimentary to capture the heterogeneity of technology. Patents in the same main patent class may be in different subclasses with inherently different technologies, and patents are assigned to multiple classes, again introducing heterogeneity not captured by the main patent classification. In response, Henderson, Jaffe, and Trajtenberg (2005) agree that it is possible that finer technological controls might be appropriate, but they point out that slicing things too finely minimizes the possibility for identifying knowledge flows across subclasses. Ultimately, the question comes down to the robustness of the localization effect under different identifying assumptions.

A number of other authors have similarly used citation data to measure knowledge flows. Almeida and Kogut (1997) compare the patent citations of small and large semiconductor firms, and find that the citations made by small firms are more geographically localized. Hicks, Breitzman, Olivastro, and Hamilton (2001) show that U.S. companies' citations to university patents exhibit geographic localization, particularly to patents of nearby public universities. Almeida and Kogut (1999) examine citation patterns among semiconductor firms in the United States, including data on both the firms and the inventors. They show that a significant fraction of the geographic localization of the citations can be traced to specific engineers who move among firms, but are more likely to move to another nearby firm than to one that is farther away. Sonn and Storper (2008) show that, despite improvements in communications technologies, geographical localization has been increasing over time.

Thompson (2006) compares the extent of localization in citations listed by the inventor to those added by the examiner. He finds localization at both the metropolitan area and state levels in both the examiner and inventor citations. Inventor citations are found to be about 20% more likely to match the country of origin of the citing patent than are examiner citations. In a similar vein, Alcácer and Gittelman (2006) estimate the probability that a citation is generated by an examiner or an inventor, conditional on a set of variables that are frequently employed in the knowledge spillover literature. They find that examiner citations introduce bias for some variables only e.g., self-citations). They find no evidence that the degree of geographic proximity between citing and cited patents differs for inventor and examiner citations.

A subtler pitfall in the use of citations to track knowledge flows relates to the intervention of law firms in the drafting of the patent document. Wagner, Hoisl, and Thoma (2014) show that patents by firms who rely on external agents are more likely to cite documents that are part of the law firm's knowledge repository. They take this result as evidence that law firms help overcome localization. However, a blunter interpretation is that external agents include citations that the firms were not aware of, further increasing the noise in patent citation data.

Maurseth and Verspagen (2002) used data on citations among European patents to construct a region-by-region citation frequency matrix. They then looked at numerous variables to explain these frequencies. Geographical distance has a negative and substantial impact on knowledge flows. Controlling for distance, knowledge flows are greater between regions located within one country than between regions located in separate countries. The country effect remains even if regions share the same language, though sharing a language increases the amount of knowledge flows between two regions by up to 28%. The study also suggests that knowledge flows are industry specific, and regions'

DOI: 10.1002/asi

technological specialization is an important determinant of their technological interaction.

Temporal Dimension of Knowledge Flows

Caballero and Jaffe (1993) and Jaffe and Trajtenberg (1999) developed a structured model of knowledge diffusion across space and time. They postulate that two competing forces dominate the citation process. Over time, knowledge gradually diffuses, so that the number of people potentially citing a given patent increases exponentially with time. But the relevance or usefulness of a bit of knowledge becomes obsolete, leading to a countervailing exponential depreciation in the likelihood of citations. The parameters of these two exponential functions can be estimated econometrically. If allowed to vary across different technologies, different kinds of research organizations, and different geographic locations, they then capture the rates of diffusion in different areas across organizations and across space. Jaffe and Trajtenberg show, for example, that the geographic localization of citations diminishes as time passes, and also that obsolescence (as captured by declining citation rates) is more rapid in electronic technologies than in chemical and mechanical technologies.

Bacchiocchi and Montobbio (2009) used this double-exponential function to look at knowledge flows from universities and public research organizations compared to flows from corporate patents in six countries: France, Germany, Italy, Japan, the United Kingdom, and the United States. They found that technology embodied in patents from universities and public research organizations diffuses more rapidly than that of firms. The diffusion rates are relatively homogenous across technological fields, but vary across countries: rapid in the United States and Germany, less so in France and Japan.

Mehta, Rysman, and Simcoe (2010) have criticized this diffusion model on the ground that the age of a citation is computed as the citation year minus the application year, leading to an identification problem. Because citations received by a patent are rare before it is issued, the authors propose to use the lag between application year and grant year as a source of exogenous variation. They find that the citation peak occurs earlier than suggested by the double-exponential function. However, their method does not alter differences in the mean citation ages across industries. They conclude that the double-exponential function provides a good approximation to the nonparametric age distribution.

Validation Studies

Jaffe, Trajtenberg, and Fogarty (2000) report a survey of inventors to test the extent to which citations in those inventors' patents correspond to the inventors' perceptions of how their inventions depended on earlier knowledge, and how the rate of citation relates to inventors' own perceptions of impact or importance. They find that citations are a valid but noisy indicator of knowledge flows: The likelihood of reported knowledge impact is significantly higher (both quantitatively and statistically) when a citation link exists, but a significant fraction of citations (perhaps as high as one half) do not correspond to any reported knowledge link.

Duguet and MacGarvie (2005) tested the validity of patent citations as a measure of knowledge flows using data from French firms on their patents and citations, combined with survey responses regarding sources of knowledge. The total number of backward citations was correlated with survey answers about R&D and innovation, but this correlation was weakened by controlling for the number of patents held by the citing firm. Backward citation rates of French firms can reflect their R&D activities (if the technology is obtained from firms located in the EU), or purchases of equipment goods (if the source is located outside the EU). In general it can be understood that backward citations are correlated with learning through R&D collaboration, licensing of foreign technology, mergers and acquisitions and equipment purchases.

In their analysis, Roach and Cohen (2013, discussed earlier) found evidence of both "errors of omission" (reported knowledge flows with no corresponding citations) and "errors of commission" (observed citations with no corresponding reported knowledge flows). They conclude that despite these sources of measurement error, patent citations are likely to reflect meaningful aspects of knowledge flows from public research. Interestingly, they found that references in patents to nonpatent publications (primarily scientific literature) are a better indicator of knowledge flow than are citations in commercial patents to the patents of universities and other public labs (cf. Tijssen, 2002).

The next section discusses a third category of citation data research, in which the focus shifts to using citation links to understand and characterize networks.

Citations as Links in Knowledge or Innovation Networks

A natural way of representing citation data is in the form of a network. Researchers have used concepts from network theory to grasp the way the innovation system is structured and the way knowledge is formed. A first group of studies seek to map key components of the innovation system (patents, individuals, institutions, and regions). A second group of studies use the network of citations to map technological trajectories. We review these two applications in turn.

Mapping Patents, Individuals, Institutions, and Regions

Huang, Chiang, and Chen (2003) rely on patent citation data to map Taiwan's electronic industry. The researchers identify USPTO patents belonging to 58 relevant Taiwanese companies as well as the citations made by these patents. They identify the strength of the relationship between companies by looking at the strength of bibliographic coupling. Bibliographic coupling is a method proposed by Kessler (1963) that involves identifying related documents through common cited references. The researchers then applied cluster analysis on the data produced to identify major sectors of

DOI: 10.1002/asi

the Taiwanese electronic industry. Although bibliographic coupling provides rich insights on the relatedness of patent documents, more recent studies make better use of network analysis theory and tools.

Chen and Hicks (2004) study the citation "degree" distribution of 16 million citations made to the 3 million USPTO patents granted in the period from 1963 to 1999. The degree of a "node" (patent) is simply the number of "connections" (citations) received by the node. They estimate that the distribution follows a power law with an exponent of 2.89, which is very similar to the parameter obtained for scientific articles by Dorogovtsev and Mendes (2002).[13] The fact that the degree distribution of the patent citation network follows a power-law is indicative of so-called scale-free networks, which can be seen as networks characterized by large hubs through which knowledge flows.

Li, Chen, Huang, and Roco (2007) use a patent citation network to study the knowledge transfer process between entities. In particular they study the efficiency with which knowledge transfers within the network compared to a random network. Their measure of efficiency is the average path length between any pair of patents in the network. They focus on USPTO nanotechnology patents in the period from 1976 to 2004. They find that knowledge transfer across assignees in the citation network is more efficient than knowledge transfer that would occur in a random network. Knowledge flow across (assignee) countries is as efficient as a random network. However knowledge flow across technology fields is less efficient than knowledge flow that would occur in a random network. In other words, technological distance is a greater barrier to knowledge flows than geographic distance.

Hung and Wang (2010) examine the characteristics of the citation network formed by RFID patents. They find that the network can be characterized as a "small-world" network, that is, a network in which most nodes can be reached from every other by a small number of steps. They also find that the network has a power-law connectivity distribution and exhibits preferential connectivity behavior. That is, a few key patents have a very large number of connections and the majority of patents have few connections. The authors conclude that only a limited number of patents play a key role in diffusing RFID technology. This approach provides a more system-based way of thinking about knowledge flows than simply counting citations: Key patents are not only highly cited patents, but also connect and integrate different technological trajectories.[14] More detailed analyses of technological trajectories using citation network are described in the next section.

Mapping of Technological Trajectories

Scholars have recently used citation networks to identify technological trajectories that led to the advent of major technological breakthroughs. The main trajectory, or search path, is the sequence of links and nodes that is central to the development of a technology. It represents the main flow of ideas in the development of a technology. The method was pioneered by Hummon and Doreian (1989) on a citation network of scientific articles describing the development of DNA theory. This approach shifts the focus from the nodes of the network (looking at individual patents) to the connections that these nodes form. It allows identifying key patents through their structural connectivity in the network. Technologically important patents should belong to the main paths of the citation network and/or locate at particularly critical junctions within those paths.

Mina, Ramlogan, Tampubolon, and Metcalfe (2007), Verspagen (2007), and Fontana, Nuvolari, and Verspagen (2009) applied the method to patent citation networks. Mina et al. (2007) use it to understand how medical knowledge emerges, grows, and evolves. They argue that the approach provides a dynamic view of innovation that recognizes the long-term, path-dependent, and complex nature of technology. Their case study is based on treatment for coronary artery disease and covers 5,136 USPTO patent documents granted between 1976 and 2003. The authors seek to identify the main path and "islands" of the network. Islands are small clusters of inventions whose internal connectedness is relatively superior to the strength of their outward connections within the global network. The authors argue that islands allow accounting for the variety of complementary and competing areas of technical expertise that contributed to the advancement of the technology. They report that the results form a consistent map of the major scientific and technological trajectories in the domain.

Fontana et al. (2009) study the structural connectivity of the citation network formed by patents related to local area networks (LAN) technology. Innovation in such a systemic technology has three main features. First, innovation is distributed: it takes place at the level of individual components but these components all have to work together. Second, innovations in systems tend to be incremental and to occur around well-established technical designs. Third, innovations also tend to occur continuously. The authors argue that the classical approach of assessing the importance of patents by counting the number of citations they have received may have drawbacks in such systemic technologies. It may fail to identify concepts and principles that could act as "focusing devices" for a sequence of inventive activities. By contrast a structural analysis of the citation network would allow the identification of inventions that have played a major role in the evolution of LAN technology. They find that the main path they have identified displays a coherent economic and engineering logic, consistent with qualitative accounts of the evolution of the Ethernet standard.

One of the most interesting insights of the article comes from the analysis of companies owning patents that lie on the main path. No company is "dominant" in the sense of claiming ownership of the majority of patents on the main path, which the authors take as evidence that no company is strategically placed along the main path of knowledge flow. Verspagen (2007) performs a similar analysis for citations among fuel cell patents. He finds that there are dominant

DOI: 10.1002/asi

companies: A small number of organizations hold patents belonging to the main path. Study of the ownership structure of technologies on the main path provides a novel way of characterizing technology dominance. It is a promising avenue for research in industrial economics and strategic management.

Pitfalls and Best Practices in Use of Citation-Based Indicators

We take the opportunity of this review to discuss potential pitfalls associated with patent data. We focus on four key challenges.

Office Effects

Institutional differences across jurisdictions induce differences in citation practices across offices. We briefly summarize two main differences in citation practices between the EPO/Japan Patent Office (JPO) and the USPTO for illustrative purposes. More generally, researchers should get a clear understanding of citation practices in the office of interest before using citation-based indicators.[15]

A first difference is the "duty of candor" in U.S. patent law. Failure to report known relevant prior art may lead to subsequent revocation of the patent (inequitable conduct doctrine). There is no duty of candor in European patent law, and applicants do not have to submit a list of prior art. It follows that search reports at the EPO usually contain many fewer references than USPTO search reports. In fact, according to EPO philosophy, "a good search report contains all the technically relevant information within a minimum number of citations" (Michel & Bettels, 2001, p. 189). In addition, since applicants at the EPO do not bear the same responsibility to disclose prior art as applicants at the USPTO, the citations come mostly from the examiner. This does not undermine their interpretation as indicators of impact or value; for example Harhoff et al. (2003, discussed earlier) find EPO citations to be predictive of value. It does suggest that EPO citations might be less indicative of knowledge flows; although we are not aware of any empirical analysis of this question comparable to the survey work of Jaffe et al. (2000). In Japan, the patent law was revised in 2002 and imposed on applicants the obligation to disclose prior art. Although not complying with the disclosure requirement bears less severe consequence than in the United States, the reform led to a substantial increase in prior art disclosure by applicants. Takahiro, Nagaoka, and Naito (2015) find that about 8% of citations came from applicants in the years following the reform, compared to around 4–5% before the reform.

A second important difference with the USPTO is that EPO patent examiners classify documents cited in particular citation categories (Schmoch, 1993). A document that shows essential features of the invention or questions the inventive step of these features if taken alone is marked with the letter "X." A document that questions the inventive step if combined with another document is marked with the letter "Y" (hence "Y" citations never occur singly). The letter "A" marks a document that shows the general state of the art. According to Schmoch (1993, p. 195) a patent document can be highly cited because it comprises "a good description of the prior art from a didactic point of view." The classification provides opportunities for finer analyses. One may want to exclude class "A" citations for assessing the inventive step of patents, but class "A" citations are relevant for measuring technological proximity of patents. Additional classification codes exist; see Webb, Dernis, Harhoff, and Hoisl (2005) for a discussion. Examiners at the JPO also classify citations into categories. In particular, they flag whether citations are used as ground for rejection, similar to "X" and "Y" citations at the EPO, or whether they are used for assessing the application but do not serve as a basis for rejection, similar to "A" citations at the EPO (Goto & Motohashi, 2007).

The classification into categories opens the door to original uses of citation data. For example, von Graevenitz, Wagner, and Harhoff (2011) identify patent thickets at the EPO using X and Y citations. Their measure identifies constellations in which three firms each own patents that block patent applications of the other two firms (so-called triples). The authors show that density of triples in *complex* technology areas has risen steadily since the early 1980s, whereas the density of triples has been constant in *discrete* technology areas. Guellec, Martinez, and Zuniga (2012) use "X" and "Y" citations together with administrative information on the patent examination process (withdrawal and grant events) to identify defensive patents, that is, patent applications used to pre-empt others from getting their patents granted. Palangkaraya, Webster, and Jensen (2011) posit that patents with a higher inventive step will generate more "X" and "Y" citations, and use this information to proxy for the probability of grant ex-ante.

Beyond institutional differences in the use of citations, researchers have also illustrated the presence of home bias in citation practices. Bacchiocchi and Montobbio (2010) analyze the geographic distribution of cited documents for a set of 657,151 equivalent patents filed at the EPO and the USPTO. In theory, distributions should be similar since they refer to the same invention. They find that the frequency of U.S.-cited patents at the USPTO exceeds 65%, while the frequency at the EPO is less than 40%. That examiners have a tendency to cite local documents does not come as a surprise.[16] However, it illustrates an important limitation of the use of citations for assessing cross-border knowledge flows.

Time and Technology Field Effects

The number of citations received by a patent increases as time passes such that there are strong cohort effects. This issue can be dealt with in a straightforward manner by counting citations received in a fixed time interval (e.g., citations received up to 5 years after grant). A more serious concern is the increase over time of citations made *per patent*. Hall, Jaffe, and Trajtenberg (2001) report that the average

DOI: 10.1002/asi

patent issued in 1999 made over twice as many citations as the average patent issued in 1975 (10.7 vs. 4.7 citations). Although this issue does not affect the comparison of patents within a cohort, citation inflation makes it challenging to compare patents across cohorts. Analogously, citation practices and the intensity of activity vary by technology fields, so that what constitutes a high citation rate in one field may be modest or small for another field.[17] The authors discuss two econometric techniques to deal with citation inflation and varying intensities by field: scaling citation counts by "dividing them by the average citation count for a group of patents to which the patent of interest belongs"; and identifying the multiple biases on citation rates via econometric estimation. Marco (2007) provides a recent illustration of the latter technique. He argues that by estimating a hazard rate based only on factors that are correlated to citation inflation rather than value, residuals can be used to measure latent patent value. For example, the ratio of observed citations to predicted citations may represent a proxy of patent value. Such an approach is an important step forward, although it is difficult to identify factors that are truly exogenous to value.

A broader question, which has received little coverage in the literature, relates to differences in patenting and citation practices across technology fields. We know that the propensity to patent differs across fields (Cohen, Nelson, & Walsh, 2000) and that the relevance of patent data as innovation indicator therefore also varies across fields (e.g., Danguy, de Rassenfosse, & van Pottelsberghe, 2014). However, to the best of our knowledge, no study has investigated in a systematic manner how differences across fields affect the relevance of patent citation data.

Examiner Effects

Cockburn, Kortum, and Stern (2002) show that there is substantial examiner heterogeneity, for example, in terms of variations in tenure at the USPTO and in the average approval time per issued patent. Such heterogeneity translates into variations in outcomes of the examination process — such as in the volume and pattern of citations made.[18] Lemley and Sampat (2012) demonstrate the presence of an examiner effect, in the sense that more experienced examiners cite less prior art. Alcacer, Gittelman, and Sampat (2009) painted a picture of examiner-added citations across key strata of patent data. They report that the proportion of citations added by examiners is higher for patents: by foreign applicants to the USPTO; by applicants with a large patent portfolio; and by applicants in electronics, communications, and computer-related fields. Criscuolo and Verspagen (2008) perform a similar analysis for EPO patent data. They show that the share of inventor citations has been declining from about 14% in 1985 to 9% in 2000. In addition, there is also substantial variation across fields. More than 20% of citations in organic chemistry patents were added by the inventor, while for information technology patents this share is 4%.[19]

Examiner intervention may bias the information content of citations. It may undermine the use of citations as a measure of knowledge flow, since the inventors may not have even been aware of the patents cited by examiners at the time of invention. However, examiner citations may be taken as a valid reflection of technological and economic value. In this spirit, Hegde and Sampat (2009) show that examiner citations have a much stronger relationship with renewal probability (a measure of private value) than the number of applicant citations.

Strategic Effects

Variations in the number of examiner-added citations may also come from differences in applicants' incentives to search for or disclose prior art. Recent research suggests that citing prior art (or not) is a strategic decision. Atal and Bar (2010) study firms' incentives to *search for unknown prior art*. Although applicants at the USPTO have a duty to disclose what they know, they have no duty to search for prior art and may be better off by remaining ignorant. The authors show theoretically that firms search more when R&D investment (a proxy for innovation quality) and patenting costs are higher. Sampat (2010) provides empirical data on when applicants search for prior art. He shows that applicants contribute more prior art for their more important inventions. He also shows that applicants are more likely to search for prior art in fields where individual patents are important for appropriating returns from R&D (chemicals and drugs) and less likely to do so in industries where firms tend to accumulate patent portfolios for other strategic reasons (computers and communications, electronics and electrical, and mechanical).

Lampe (2012) focuses on applicants' decision to *disclose known prior art*. He identifies "voluntary withholding" of citations to prior art material by looking only at citations that were present on prior patents issued to the same firm. He estimates that applicants withhold between 21% and 33% of relevant citations. The rate is higher for firms applying for computer and electronic patents (25 to 42%) and lower for firms applying for drug and chemical patents (8–22%). More generally, Lampe finds that the likelihood of citation is positively correlated with proxies of patent value (number of claims and forward citations) and negatively correlated with the size of applicant patent portfolios.

Conclusion

The use of patent citation data in social science research has exploded in the last two decades. As just one indication, the frequency of appearance of the term "patent citation" in scientific documents listed in Google Scholar increased 10-fold between 2000 and 2014 (Appendix A). As is often the case, this increase reflects increases in both supply and demand. On the supply side, the digitization of the patent office records, combined with the increased power of computers to analyze them, makes analyses possible today that simply could not have been undertaken 25 years ago. The

DOI: 10.1002/asi

number of scientific documents referencing the National Bureau of Economic Research (NBER) patent citation data file is likewise continuously increasing (Appendix A). On the demand side, intangible assets are increasingly seen as a source—some would argue the dominant source—of economic returns. By definition, intangible assets are hard to track and measure, and so researchers interested in diverse questions about knowledge accumulation and diffusion, innovation, firm strategy and regional economic growth seek measures that convey information about the sources and consequences of these assets.

Neither of these trends is likely to reverse, so interest in measures of this kind is likely to continue to grow. Recent developments in computational linguistics may allow for construction of measures that are conceptually related to citations but use all of the information contained in the patent text rather than relying solely on the links between patents that are explicitly identified via citation. It is now possible, for example, to identify connections between a patent and its antecedents by measuring the frequency with which important words are used in both patents, to measure novelty by identifying patents that use a certain technical term or combination of words in a particular phrase for the first time, and to measure impact by counting the number of subsequent patents that use such a phrase (e.g., Packalen & Bhattacharya, 2015). Younge and Khun (2015) use more advanced techniques to develop a text-based pairwise similarity comparison of any and every two patents at the USPTO. These new approaches have not yet been subjected to the kind of validation that has demonstrated the economic significance of citations, but because they utilize more information, they offer the promise of a valuable broadening and deepening of the research possibilities.

A more mundane, but equally important task, is to further validate citation indicators. This applies to both established and novel indicators, at both the USPTO and other offices. For example it is unclear whether the count of backward citations proxies for patent importance. Even the link between forward citations and economic value, one of the most established and used indicators, is not well understood. In a similar vein, little research exists on technology field differences on the relevance of patent citation data. The need for validation studies will grow more pressing as new indicators are being developed and more patent offices make their data available. Similarly, legislative changes affect citation practices in nontrivial ways, and conclusions drawn using data from one-time period are not necessarily valid in another time period. This calls for a continuous assessment of the validity of citation indicators.

Another exciting area of research is the further application of network theory and analysis tools to the patent citation network. For example, the identification of key technologies and actors on the main knowledge path promises to greatly improve our understanding of industry dynamics and the knowledge creation process. A limitation of current research in the area is the insularity of two communities of scholars. Studies by scholars using advanced network analysis tools offer little practical implications, whereas studies by scholars looking at real-world implications use quite basic network analysis tools. A promising way forward is to better integrate the technical and the practical aspects of network analysis.

Finally, researchers realize that the patent citation generation process is complex but more work needs to be done to understand it. The complexity of the patent citation generation process is a blessing and a curse. Whereas it may distort the reality in an undesirable fashion, it may also provide a window into the incentives faced by inventors, patent attorneys and examiners and serve as a source of econometric identification. The example of examiner-added citations is a case in point. Whereas citations made by examiners arguably weaken the measurement of knowledge flows, they also strengthen the measurement of patent value.

Acknowledgments

We thank David Schwartz for suggesting this project. We are grateful to two anonymous referees for valuable comments. We also received helpful comments from Bronwyn Hall, Dietmar Harhoff, Sadao Nagaoka, and Beth Webster, as well as participants at the 2015 Workshop on the Economics of Intellectual Property in Northwestern University. Jan Kozak provided valuable research assistance. We are solely responsible for all opinions or errors.

Endnotes

[1]The present survey discusses evidence on citations at the European Patent Office whenever available.

[2]The earliest reference that we found is Clark (1976). It presents statistics on the obsolescence of United States Patent and Trademark Office (USPTO) patents using citation data. Garfield (1966) discusses the use of patent citation searches to say something about the significance of a patent, but it does not present any systematic analyses or statistics. Kuznets (1962) did not specifically discuss citations, but did emphasize that patent documents are a rich and deep source of information on the inventive process, and urged that this richness be exploited in addition to researchers' simply counting patents

[3]Jacob Schmookler pointed out that in a patent subclass "Dispensing of semi-solid materials," he found a patent for a manure spreader and another for a toothpaste tube (Schmookler, 1966).

[4]Jaffe and Trajtenberg (2002) reprints 12 of the key articles on patent citations by them and their co-authors.

[5]The authors report "forward importance" as the number of citations received plus .5 times the number of citations received by the citing patents, and undertook sensitivity analysis varying this weight between 0.25 and 0.75. Extending this throughout the citation tree involves a geometrically declining weight, for example, if patent E cites patent D which cites patent C which cites patent B which cites patent A, we might consider patent B to contribute 1 to the importance of A, patent C 0.5, patent D 0.25 and patent E 0.125.

[6]For a small number of citations, it is clear that this measure is heavily influenced by the number of citations, for example, a patent receiving only two citations cannot possibly have generality greater than 0.5. Whether or not this is a problem is largely a matter of interpretation; in some sense it is meaningful to say that a patent receiving only two citations cannot have a very diverse impact. A different interpretation is that every invention has a latent or unobserved generality that is

DOI: 10.1002/asi

randomly realized in the citations it happens to receive. Under this formulation, the distribution of citations across patent classes is multinomial, and the observed generality or originality is a biased estimator of the true parameter. Bronwyn Hall has derived a formula to correct for this bias (Hall et al., 2001); it produces a significant correction for patents with just a few citations.

[7]Ziedonis (2004) has built on this idea to construct a measure of the fragmentation of ownership rights to a firm's complementary patents. Backward citations are stratified by assignee instead of technology class.

[8]Hall and Trajtenberg (2006) explain that the generality measures suffer from the fact that they treat citations from patents in patent classes different from the cited patents in the same way, although some patent classes are very different and some are closely related. They suggest that the future research could construct a weighted generality measure, with weights inversely related to the overall probability that one class cites another class. To the best of our knowledge no one has implemented such an approach.

[9]Hall and Trajtenberg (2006) also show that a disproportionate share of the patents in the extreme upper tail of the distribution for generality and total forward citations in the period 1967–1999 are information technology (IT) patents, suggesting that these metrics may be indicative of a GPT.

[10]Lemley and Sampat (2012, footnote 12) find that the vast majority of references to nonpatent prior art at the USPTO come from applicants, not examiners, potentially making these a relevant measure of science dependence.

[11]Since Trajtenberg (1990b) showed that total citations are correlated with social returns, the finding that self-citations have a stronger effect on market value than other citations suggest that self-citation is associated with the extent of appropriation of the social returns by the original patenting firm.

[12]Similarly, international family size is a measure predictive of value that is knowable soon after patent application.

[13]Cf. Huang, Huang, Chang, Chen, and Lin (2014) who provide evidence that the distribution of patent citations is more concentrated than the distribution of citations in scientific articles.

[14]Hu, Rousseau, and Chen (2012) provide another study on the importance of patents using their positions in the citation network. Other applications include, for example, Liu and Shih (2011) who use the network formed by patents to improve patent classification.

[15]For example, researchers interested in EPO citations should read the "Guidelines for Examination in the European Patent Office" available on the EPO website.

[16]For example, there is a substantial cost to including non-English references at the USPTO. When using a foreign language reference in a rejection, examiners should provide a translation of the entire document.

[17]Technology fields are tracked using the patent office classification systems. Historically, the United States has maintained its own classification (USPC), while other offices use the International Patent Classification (IPC). The USPTO has recently introduced a Cooperative Patent Classification (CPC) based on the IPC, and is phasing out the USPC.

[18]Alcácer and Gittelman (2006) estimate that examiners insert two thirds of citations on the average patent, and 40% of all patents have all citations added by examiners.

[19]There are few country-specific studies. See Azagra-Caro, Mattsson, and Perruchas (2011) for Spanish evidence.

References

Abrams, D., Akcigit, U., & Popadak, J. (2013). Patent value and citations: Creative destruction or strategic disruption? National Bureau of Economic Research Working Paper No. 19647.

Ahuja, G., & Lampert, C.M. (2001). Entrepreneurship in the large corporation: A longitudinal study of how established firms create breakthrough inventions. Strategic Management Journal, 22(6–7), 521–543.

Albert, M.B., Avery, D., Narin, F., & McAllister, P. (1991). Direct validation of citation counts as indicators of industrially important patents. Research Policy, 20(3), 251–259.

Alcácer, J., & Gittelman, M. (2006). Patent citations as a measure of knowledge flows: The influence of examiner citations. The Review of Economics and Statistics, 88(4), 774–779.

Alcacer, J., Gittelman, M., & Sampat, B. (2009). Applicant and examiner citations in US patents: An overview and analysis. Research Policy, 38(2), 415–427.

Allison, J.R., Lemley, M.A., Moore, K.A., & Trunkey, R.D. (2003). Valuable patents. Georgetown Law Journal, 92, 435.

Almeida, P., & Kogut, B. (1997). The exploration of technological diversity and the geographic localization of innovation. Small Business Economics, 9, 21–31.

Almeida, P., & Kogut, B. (1999). Localization of knowledge and the mobility of engineers in regional networks. Management Science, 45(7), 905–917.

Atal, V., & Bar, T. (2010). Prior art: To search or not to search. International Journal of Industrial Organization, 28(5), 507–521.

Azagra-Caro, J.M., Mattsson, P., & Perruchas, F. (2011). Smoothing the lies: The distinctive effects of patent characteristics on examiner and applicant citations. Journal of the American Society for Information Science and Technology, 62(9), 1727–1740.

Bacchiocchi, E., & Montobbio, F. (2009). Knowledge diffusion from university and public research. A comparison between US, Japan and Europe using patent citations. The Journal of Technology Transfer, 34(2), 169–181.

Bacchiocchi, E., & Montobbio, F. (2010). International knowledge diffusion and home-bias effect: Do USPTO and EPO patent citations tell the same story?. The Scandinavian Journal of Economics, 112(3), 441–470.

Benson, C.L., & Magee, C.L. (2015). Quantitative determination of technological improvement from patent data. PLoS One, 10(4), e0121635.

Bessen, J. (2008). The value of US patents by owner and patent characteristics. Research Policy, 37(5), 932–945.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30, 107–117.

Caballero, R.J., & Jaffe, A.B. (1993). How high are the giants' shoulders: An empirical assessment of knowledge spillovers and creative destruction in a model of economic growth. In O.Blanchard and S. Fischer (Eds), NBER Macroeconomics Annual 1993 (Vol. 8, pp. 15–86). Cambridge: MIT press.

Callaert, J., Van Looy, B., Verbeek, A., Debackere, K., & Thijs, B. (2006). Traces of prior art: An analysis of non-patent references found in patent documents. Scientometrics, 69(1), 3–20.

Carpenter, M.P., & Narin, F. (1983). Validation study: Patent citations as indicators of science and foreign dependence. World Patent Information, 5(3), 180–185.

Carpenter, M.P., Narin, F., & Woolf, P. (1981). Citation rates to technologically important patents. World Patent Information, 3(4), 160–163.

Chen, C., & Hicks, D. (2004). Tracing knowledge diffusion. Scientometrics, 59(2), 199–211.

Clark, C.V. (1976). Obsolescence of the patent literature. Journal of Documentation, 32(1), 32–52.

Cockburn, I.M., Kortum, S., & Stern, S. (2002). Are all patent examiners equal?: The impact of characteristics on patent statistics and litigation outcomes. *National Bureau of Economic Research Working Paper* 8980.

Cohen, W.M., Nelson, R.R., & Walsh, J.P. (2000). Protecting their intellectual assets: Appropriability conditions and why U.S. manufacturing firms patent or not. *National Bureau of Economic Research Working Paper* 7552.

Collins, P., & Wyatt, S. (1988). Citations in patents to the basic research literature. Research Policy, 17(2), 65–74.

Cotropia, C.A., Lemley, M.A., & Sampat, B. (2013). Do applicant patent citations matter? Research Policy, 42(4), 844–854.

Criscuolo, P., & Verspagen, B. (2008). Does it matter where patent citations come from? Inventor vs. examiner citations in European patents. Research Policy, 37(10), 1892–1908.

Czarnitzki, D., Hussinger, K., & Schneider, C. (2011). "Wacky" patents meet economic indicators. Economics Letters, 113(2), 131–134.

DOI: 10.1002/asi

Dahlin, K.B., & Behrens, D.M. (2005). When is an invention really radical?: Defining and measuring technological radicalness. Research Policy, 34(5), 717–737.

Danguy, J., de Rassenfosse, G., & van Pottelsberghe de la Potterie, B. (2014). On the origins of the worldwide surge in patenting: An industry perspective on the R&D-patent relationship. Industrial and Corporate Change, 23(2), 535–572.

Dorogovtsev, S.N., & Mendes, J.F. (2002). Evolution of networks. Advances in Physics, 51(4), 1079–1187.

Duguet, E., & MacGarvie, M. (2005). How well do patent citations measure flows of technology? Evidence from French innovation surveys. Economics of Innovation and New Technology, 14(5), 375–393.

Fontana, R., Nuvolari, A., & Verspagen, B. (2009). Mapping technological trajectories as patent citation networks. An application to data communication standards. Economics of Innovation and New Technology, 18(4), 311–336.

Gambardella, A., Harhoff, D., & Verspagen, B. (2008). The value of European patents. European Management Review, 5(2), 69–84.

Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of ideas. Science, 122(3159), 108–111.

Garfield, E. (1966). Patent citation indexing and the notions of novelty, similarity, and relevance. Journal of Chemical Documentation, 6(2), 63–65.

Gay, C., & Le Bas, C. (2005). Uses without too many abuses of patent citations or the simple economics of patent citations as a measure of value and flows of knowledge. Economics of Innovation and New Technology, 14(5), 333–338.

Goto, A., & Motohashi, K. (2007). Construction of a Japanese Patent Database and a first look at Japanese patenting activities. Research Policy, 36(9), 1431–1442.

Griliches, Z. (1979). Issues in assessing the contribution of research and development to productivity growth. The Bell Journal of Economics, 92–116.

Griliches, Z. (1990). Patent statistics as economic indicators: A survey. Journal of Economic Literature, 28(4), 1661–1707.

Guellec, D., Martinez, C., & Zuniga, P. (2012). Pre-emptive patenting: Securing market exclusion and freedom of operation. Economics of Innovation and New Technology, 21(1), 1–29.

Hall, B.H., Jaffe, A.B., & Trajtenberg, M. (2001). The NBER patent citation data file: Lessons, insights and methodological tools. *National Bureau of Economic Research Working Paper* 8498.

Hall, B.H., Jaffe, A., & Trajtenberg, M. (2005). Market value and patent citations. RAND Journal of Economics, 36(1), 16–38.

Hall, B.H. & Trajtenberg, M. (2006). Uncovering GPTs using patent data. In C. Antonelli, D. Foray, B.H. Hall, & W.E. Steinmuller (Eds.), New frontiers in the economics of innovation and new technology—Essays in honor of Paul A. David. Cheltenham: Edward Elgar Publishing.

Harhoff, D., Narin, F., Scherer, F.M., & Vopel, K. (1999). Citation frequency and the value of patented inventions. Review of Economics and Statistics, 81(3), 511–515.

Harhoff, D., Scherer, F.M., & Vopel, K. (2003). Citations, family size, opposition and the value of patent rights. Research Policy, 32(8), 1343–1363.

Hegde, D., & Sampat, B. (2009). Examiner citations, applicant citations, and the private value of patents. Economics Letters, 105(3), 287–289.

Henderson, R., Jaffe, A., & Trajtenberg, M. (2005). Patent citations and the geography of knowledge spillovers: A reassessment: Comment. American Economic Review, 95(1) 461–464.

Hicks, D., Breitzman, T., Olivastro, D., & Hamilton, K. (2001). The changing composition of innovative activity in the US—A portrait based on patent analysis. Research Policy, 30(4), 681–703.

Hu, X., Rousseau, R., & Chen, J. (2012). A new approach for measuring the value of patents based on structural indicators for ego patent citation networks. Journal of the American Society for Information Science and Technology, 63(9), 1834–1842.

Huang, M.H., Chiang, L.Y., & Chen, D.Z. (2003). Constructing a patent citation map using bibliographic coupling: A study of Taiwan's high-tech companies. Scientometrics, 58(3), 489–506.

Huang, M.H., Huang, W.T., Chang, C.C., Chen, D.Z., & Lin, C.P. (2014). The greater scattering phenomenon beyond Bradford's law in patent citation. Journal of the Association for Information Science and Technology, 65(9), 1917–1928.

Hummon, N.P., & Doreian, P. (1989). Connectivity in a citation network: The development of DNA theory. Social Networks, 11(1), 39–63.

Hung, S.W., & Wang, A.P. (2010). Examining the small world phenomenon in the patent citation network: A case study of the radio frequency identification (RFID) network. Scientometrics, 82(1), 121–134.

Jaffe, A.B., & Trajtenberg, M. (1999). International knowledge flows: Evidence from patent citations. Economics of Innovation and New Technology, 8(1–2), 105–136.

Jaffe, A.B., & Trajtenberg, M. (2002). Patents, citations, and innovations: A window on the knowledge economy. Cambridge: MIT press.

Jaffe, A.B., Trajtenberg, M., & Fogarty, M.S. (2000). Knowledge spillovers and patent citations: Evidence from a survey of inventors. *American Economic Review*, 90(2), 215–218; also published with additional detail as "The Meaning of Patent Citations: Report on the NBER/Case-Western Reserve Survey of Patentees," in Jaffe and Trajtenberg (2002), *op cit*.

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.

Kessler, M.M. (1963). Bibliographic coupling between scientific papers. American Documentation, 14(1), 10–25.

Krugman, P.R. (1991). Geography and trade. Cambridge: MIT press.

Kuznets, S. (1962). Inventive activity: Problems of definition and measurement. In The rate and direction of inventive activity: Economic and social factors (pp. 19–52). Princeton, NJ: Princeton University Press.

Lampe, R. (2012). Strategic citation. Review of Economics and Statistics, 94(1), 320–333.

Lanjouw, J., & Schankerman, M. (2001). Characteristics of patent litigation: A window on competition. RAND Journal of Economics, 32(1), 129–151.

Lemley, M., & Sampat, B. (2012). Examiner characteristics and patent office outcomes. Review of Economics and Statistics, 94(3), 817–827.

Lerner, J., & Seru, A. (2015). The use and misuse of patent data: Issues for corporate finance and beyond. Harvard Business School Mimeo. Cambridge: MA.

Li, R., Chambers, T., Ding, Y., Zhang, G., & Meng, L. (2014). Patent citation analysis: Calculating science linkage based on citing motivation. Journal of the Association for Information Science and Technology, 65(5), 1007–1017.

Li, X., Chen, H., Huang, Z., & Roco, M.C. (2007). Patent citation network in nanotechnology (1976–2004). Journal of Nanoparticle Research, 9(3), 337–352.

Liu, Y., Hseuh, P.Y., Lawrence, R., Meliksetian, S., Perlich, C., & Veen, A. (2011). Latent graphical models for quantifying and predicting patent quality. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1145–1153). August 21–24, 2011, San Diego, California, USA.

Liu, D.R., & Shih, M.J. (2011). Hybrid-patent classification based on patent-network analysis. Journal of the American Society for Information Science and Technology, 62(2), 246–256.

Lukach, R., & Lukach, M. (2007). Ranking USPTO patent documents by importance using random surfer method (pagerank). Available at SSRN: http://ssrn.com/abstract=996595.

Marco, A.C. (2007). The dynamics of patent citations. Economics Letters, 94(2), 290–296.

Maurseth, P.B., & Verspagen, B. (2002). Knowledge spillovers in Europe: A patent citations analysis. The Scandinavian Journal of Economics, 104(4), 531–545.

Mehta, A., Rysman, M., & Simcoe, T. (2010). Identifying the age profile of patent citations: New estimates of knowledge diffusion. Journal of Applied Econometrics, 25(7), 1179–1204.

Meyer, M. (2000). What is special about patent citations? Differences between scientific and patent citations. Scientometrics, 49(1), 93–123.

Michel, J., & Bettels, B. (2001). Patent citation analysis. A closer look at the basic input data from patent search reports. Scientometrics, 51(1), 185–201.

Mina, A., Ramlogan, R., Tampubolon, G., & Metcalfe, J.S. (2007). Mapping evolutionary trajectories: Applications to the growth and transformation of medical knowledge. Research Policy, 36(5), 789–806.
Moser, P., & Nicholas, T. (2004). Was electricity a general purpose technology? The American Economic Review, 94(2), 388–394.
Moser, P., Ohmstedt, J., & Rhode, P.W. (2014). Patent citations and the size of patented inventions - Evidence from Hybrid Corn (p. 48). Available at SSRN: http://ssrn.com/abstract=1888191.
Narin, F. (1995). Patents as indicators for the evaluation of industrial research output. Scientometrics, 34(3), 489–496.
Narin, F., & Noma, E. (1985). Is technology becoming science? Scientometrics, 7(3–6), 369–381.
Narin, F., Noma, E., & Perry, R. (1987). Patents as indicators of corporate technological strength. Research Policy, 16(2), 143–155.
Nicholas, T. (2008). Does innovation cause stock market runups? Evidence from the great crash. The American Economic Review, 1370–1396.
Packalen, M., & Bhattacharya, J. (2015). New ideas in invention. *National Bureau of Economic Research Working Paper 20922*.
Palangkaraya, A., Webster, E., & Jensen, P.H. (2011). Misclassification between patent offices: Evidence from a matched sample of patent applications. Review of Economics and Statistics, 93(3), 1063–1075.
Roach, M., & Cohen, W.M. (2013). Lens or prism? Patent citations as a measure of knowledge flows from public research. Management Science, 59(2), 504–525.
Sampat, B.N. (2010). When do applicants search for prior art? Journal of Law and Economics, 53(2), 399–416.
Schmoch, U. (1993). Tracing the knowledge transfer from science to technology as reflected in patent indicators. Scientometrics, 26(1), 193–211.
Schmookler, J. (1966). Invention and economic growth. Cambridge: Harvard University Press.
Sonn, J.W., & Storper, M. (2008). The increasing importance of geographical proximity in knowledge production: An analysis of US patent citations, 1975–1997. Environment and Planning A, 40, 1020–1039.
Takahiro, M., Nagaoka, S., & Naito, Y. (2015). Effects of stronger disclosure rule on applicants' behavior and on examination efficiency: Evidence from Japan. Paper presented at EPIP 2015 conference , 2–3 September 2015, University of Glasgow, UK.
Thompson, P. (2006). Patent citations and the geography of knowledge spillovers: Evidence from inventor-and examiner-added citations. The Review of Economics and Statistics, 88(2), 383–388.
Thompson, P., & Fox-Kean, M. (2005). Patent citations and the geography of knowledge spillovers: A reassessment. American Economic Review, 95, 450–460.
Tijssen, R.J. (2002). Science dependence of technologies: Evidence from inventions and their inventors. Research Policy, 31(4), 509–526.
Trajtenberg, M. (1990a). Economic analysis of product innovation: The case of CT scanners (Vol. 160). Harvard University PressCambridge: MA.
Trajtenberg, M. (1990b). A penny for your quotes: Patent citations and the value of innovations. The Rand Journal of Economics, 21(1), 172–187.
Trajtenberg, M., Henderson, R., & Jaffe, A. (1997). University versus corporate patents: A window on the basicness of invention. Economics of Innovation and New Technology, 5(1), 19–50.
Verspagen, B. (2007). Mapping technological trajectories as patent citation networks: A study on the history of fuel cell research. Advances in Complex Systems, 10(1), 93–115.
von Graevenitz, G., Wagner, S., & Harhoff, D. (2011). How to measure patent thickets—A novel approach. Economics Letters, 111(1), 6–9.
von Wartburg, I., Teichert, T., & Rost, K. (2005). Inventive progress measured by multi-stage patent citation analysis. Research Policy, 34(10), 1591–1607.
Wagner, S., Hoisl, K., & Thoma, G. (2014). Overcoming localization of knowledge—The role of professional service firms. Strategic Management Journal, 35(11), 1671–1688.
Webb, C., Dernis, H., Harhoff, D., & Hoisl, K. (2005). Analysing European and international patent citations: A set of EPO patent database building blocks. *OECD STI Working Paper* 2005/09.
Youtie, J., Iacopetta, M., & Graham, S. (2008). Assessing the nature of nanotechnology: Can we uncover an emerging general purpose technology? The Journal of Technology Transfer, 33(3), 315–329.
Younge, K, & Kuhn, J. (2015). Patent similarity: A vector space model. Mimeo, Ecole polytechnique fédérale de Lausanne, Lausanne, Switzerland.
Ziedonis, R.H. (2004). Don't fence me in: Fragmented markets for technology and the patent acquisition strategies of firms. Management Science, 50(6), 804–820.

Appendix A

Figure A1 plots the yearly number of scientific articles listed in Google Scholar that contain the term "patent citation" (blue line), and the number of articles citing the NBER patent citation data file described in Hall et al. (2001) (red dashed line). One can reasonable assume that the latter group of articles forms a subset of the former group.

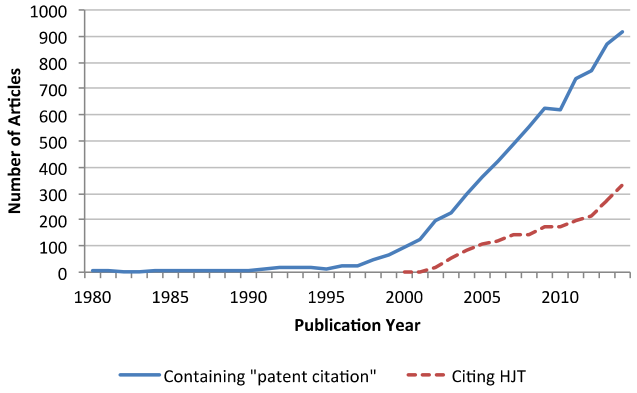

FIG. A1. Number of scientific articles listed in Google Scholar. Notes: HJT refers to Hall et al. (2001). [Color figure can be viewed at wileyonlinelibrary.com]