1 Introduction

Adam B. Jaffe and Manuel Trajtenberg

For more than two centuries now, the economies of the western world have grown at a pace that greatly exceeds anything previously known in the long sweep of human history. In the last few decades, we have experienced what have come to be called the "information age" and the "knowledge economy." Hype aside, these labels do reflect a very real transformation: it is now "knowledge"—not labor, machines, land or natural resources—that is the key economic asset that drives long-run economic performance.

At the heart of this phenomenon lies a complex, multifaceted process of continuous, widespread and far-reaching innovation and technical change. Yet "knowledge," "innovation," and "technical change" are elusive notions, difficult to conceptualize and even harder to measure in a consistent, systematic way. Thus, while economists from Adam Smith on have amply recognized their crucial role in shaping the process of economic growth, our ability to study these phenomena has been rather limited.

The last several decades have seen a number of pioneering efforts to overcome these measurement problems and gather data that can be used for the systematic empirical analysis of technological change. This volume describes our contribution to this tradition, based on the massive use of detailed patent data. Patent records contain a wealth of information on each patented invention, including the identity and location of the inventors and the inventors' employer, and the technological area of the invention. Moreover, they contain citations to previous patents, which open the possibility of tracing multiple links across inventions.

We have compiled a highly detailed dataset on all US patents granted between 1963 and 1999, and the citations they made. The

sheer volume of these data (3 million patents, 16 million citations), combined with the rich detail that they provide on each invention, make them indeed a promising "window" on the "knowledge economy."

Ever since we started along this path over a decade ago, our goal has been to throw that window wide open, so as to make empirical research on the economics of innovation and technical change a viable, exciting and fruitful enterprise. Together with a number of coauthors, we have used the patent data to get an empirical handle on quantifying the "importance" or "value" of innovations, measuring flows of technological knowledge, and characterizing the technological development and impact of particular institutions and countries. This volume lays out the conceptual and methodological foundations of this line of research, shows a range of interesting findings that give initial credence to the approach, and provides tools to tackle many of the thorny issues that arise along the way. Moreover, we include with this volume a CD with the complete data on patents and citations, encompassing both data items from the source records, and a range of novel measures that we have developed and computed. We hope that this will stimulate further work in this area, and provide a broader and deeper measurement base upon which research on the economics of technical change can flourish.

The rest of this introduction provides an overview of the patent data, some historical background on the origins of this line of research, and a summary of the main themes of the book. Part I contains three papers from the early 1990s that provide the conceptual foundations of our research approach. Part II describes the use of citations data to explore geographic patterns of spillovers across regions and countries. Part III deploys patent-based data to look at some policy-motivated issues regarding recent changes in the performance and technological impact of universities and government research labs, and to characterize the evolution of innovation in a particular high-tech economy, that of Israel. The final part of the book contains two chapters focused on the patent data themselves. One is a survey of inventors that provides a first-hand perspective on the inferences that can be drawn from citation data, and the other is a detailed overview and "users' guide" to the complete patent database itself, including discussion of statistical problems that arise in the use of the data.

1 Overview of the Data

Patents have long been recognized as a very rich and potentially fruitful source of data for the study of innovation and technical change.[1] The number of patents is very large: the "stock" of patents is currently in excess of 6 million, and the flow is over 150,000 patents per year (as of 1999–2000). Our database contains the approximately 3 million patents granted between 1963 and 1999. Each patent granted produces a highly structured public document containing detailed information on the innovation itself, the technological area to which it belongs, the inventors (e.g., their geographical location), and the organization (if any) to which the inventors assign the patent property right. The data used in this book are the computerized data items that appear on the "front page" of a granted patent, an example of which is reproduced in the appendix.

Patent data include references or citations to previous patents and to the scientific literature. Unlike bibliographic citations, patent citations perform an important legal function, in helping to delimit the patent grant by identifying "prior art" that is not covered by a given patent grant. Our data base contains the 16.5 million citations made by patents granted between 1975 and 1999. These citations open up the possibility of tracing multiple linkages between inventions, inventors, scientists, firms, locations, etc. In particular, patent citations allow one to study spillovers, and to create indicators of the "importance" or technological impact of individual patents, thus introducing a way of capturing the enormous heterogeneity in the value of patents. In addition to the "raw" patent and citations information, our data include a number of citations-based measures that are meant to capture different aspects of the patented innovations, such as "generality," "originality," and citations time lags. The measures are explored in several of the chapters, and described in detail for the complete dataset in chapter 13.

There are, of course, important limitations to the use of patent data, the most glaring being the fact that not all inventions are patented. First, not all inventions meet the patentability criteria set by the USPTO, the United States Patent and Trademark Office (the invention has to be novel and nontrivial, and has to have commercial

1. For a broader and more detailed survey of the use of patent statistics in empirical research prior to 1990, see Griliches (1990).

application). Second, the inventor has to make a strategic decision to patent, as opposed to rely on secrecy or other means of appropriability. Exploring the extent to which patents are indeed representative of the wider universe of inventions is an important, wide-open area for future research.

Another problem that used to be a serious hindrance stemmed from the fact that the patent file was not entirely computerized. Furthermore, until not long ago it was extremely difficult to handle those "chunks" that were computerized, because of the very large size of the data. The practical significance of the difficulty of dealing with large datasets was exacerbated by what we call the "inversion" problem. This refers to the fact that, in order to count the number of citations *received* by any given patent, one has to look at the citations made by *all* subsequent patents. Thus, any study using citations received, however small the sample of patents is, requires in fact access to the whole citations data, in a way that permits efficient search and extraction of citations. Indeed, when we started in the late 1980s, it was only possible to analyze relatively small samples of the data, and the feasibility of economic analyses routinely incorporating all patents was dubious. Today, however, rapid progress in computer technology has virtually eliminated these difficulties. Our complete patent data reside in personal computers and can be analyzed with the aid of standard PC software.

2 The Broad Shoulders on Which We Stand[2]

Our conception of the role of patent citations is predicated on a cumulative view of the process of technological development, by which each inventor benefits from the work of those before, and in turn contributes to the base of knowledge upon which future inventors build. Analogously, this book builds upon a broad and deep foundation of previous work. The origins of the quantitative analysis of technological change lie in the immediate post–WW II period. The

2. We confine the discussion in this brief subsection to a small number of key research programs that we perceive as *direct* antecedents to our own. The work presented in this volume obviously "stands (also) on the shoulders" of prominent scholars in this field such as Bob Evenson, Edwin Mansfield, Nathan Rosenberg, and Mike Scherer, to mention just the most notable omissions. However, this is not intended as a survey of the literature, and hence we chose not to expand on them simply because the methodological connection between their work and the current volume is somewhat less direct.

path-breaking findings of Abramovitz (1956) and Solow (1957) that there was a large "residual" of aggregate productivity growth that could not be explained by capital accumulation opened up a whole new and exciting research frontier. In parallel, and responding to the challenge posed by the productivity black box identified in those studies, empirical microeconomic analysis of the underlying phenomena of invention and innovation were also undertaken. A conference held in the spring of 1960 brought together these early lines of inquiry, and set the agenda for future work in this area. The resulting volume, edited by Richard Nelson (1962), *The Rate and Direction of Inventive Activity*, went on to became a landmark, and constitutes to this day a source of inspiration and guidance.[3]

Nelson's volume, best known perhaps for the classic paper by Kenneth Arrow that formalized the market failure inherent in research, contains also a less-cited but visionary paper by Simon Kuznets on the difficulties of measuring the results of the inventive process. Kuznets's paper raised many of the issues that permeate later research, including many of the papers in the present volume. He discussed the problems of defining and measuring the magnitude of inventions; the relationship between the technological and economic significance of an invention; the distinction between the cost of producing an invention and the value it creates; and the consequences of the highly skewed distribution of inventions values. Kuznets also considered the benefits and drawbacks of patent statistics, and included a plea—to which the current volume is certainly responsive—to go beyond merely counting patents, and utilize the rich and detailed information about the inventive process itself that is revealed in patent documents.

A clear antecedent of the present volume can be seen in Jacob Schmookler's 1966 book, *Invention and Economic Growth*, as well as the posthumous volume of Schmookler's work published in 1972. Schmookler methodically went through (non-computerized) patent records to compile hundreds of time series of patent totals by industry, going back over a century. He also gave careful attention to the methodological issues arising from the use of these data, particularly the difficulty of identifying patents with particular industries based on their technological classification by the patent office. Using

3. Richard R. Nelson's Introduction to the 1962 volume contains an interesting account of the developments in the 1950s that led up to the conference.

these data, Schmookler provided strong evidence for the role of market forces in shaping the rate and direction of inventive activity. More important in the long run, he demonstrated that patent statistics, though perhaps cumbersome to accumulate and subject to issues of interpretation, provide a unique source of systematic information about the inventive process. In this volume we carry on Schmookler's work, both by publishing updated and expanded patent statistics for other researchers to use, and by analyzing in detail the methodological and interpretational issues that arise in the use of patent statistics.

In the late 1970s, Zvi Griliches took advantage of the computerization of the USPTO, as well as the availability of other micro-data in computerized form (such as Compustat), to launch a major new research initiative on the innovation process, that relied on the merger and joint use of these distinct data sources. His students and colleagues at the NBER, led by Bronwyn Hall, compiled a large firm-level panel dataset that combined patent totals with R&D and other financial information from firms' 10-K financial reports (Bound et al. 1984; Hall et al. 1988). Armed with the first plentiful crop from this research program, Griliches organized a conference on R&D, Patents and Productivity in the fall of 1981, and published the proceedings in a volume that echoed the 1962 Nelson's volume mentioned above (Griliches 1984). Over the ensuing decade, a large amount of research was done based on the NBER R&D panel dataset and its descendants, and established many of what we now think of as "stylized facts" about R&D and patents at the firm level:[4]

- In the cross-section, patents are roughly proportional to R&D, with the ratio varying by industry and being higher for small firms (Bound et al. 1984).
- For particular firms over time (the "within" panel dimension), patents are correlated with R&D, typically with decreasing returns to R&D, with the strongest relationship being simultaneous and contemporaneous between R&D and patent applications (Hall, Griliches, and Hausman 1986).
- In multivariate models including R&D, patents and performance measures (e.g. productivity growth, profitability, market value), most

4. We do not discuss here the even larger literature derived from these data relating to R&D and productivity, but not necessarily to patents; see for example Griliches, 1994.

of the information is in the correlation between R&D and performance. Patent counts are more weakly correlated with performance, and often do not have incremental explanatory power once R&D is included (Pakes 1985; Cockburn and Griliches 1988; Griliches, Hall, and Pakes 1991).

• Detailed information on the technological composition of firms' patents can be used to locate firms' research programs in "technology space;" variations across this space in technological opportunity and "spillovers" of R&D have measurable effects on research performance (Jaffe 1986).

In addition to this econometric work, the late 1970s and early 1980s saw important conceptual developments in modeling the research process and the role of patents in that process. Griliches (1979) and Griliches and Pakes (1984) extended and refined the concept of the "knowledge production function," a stochastic relationship in which current R&D investment, the firm's existing stock of knowledge, and knowledge from other sources combine to produce new knowledge. Patent applications can be viewed as a noisy indicator of the success of this stochastic knowledge production process, with the "propensity to patent"—the ratio of patents to the unobservable knowledge production—possibly varying over time and institutions. Griliches (1979) also suggested that the possibility of excess social returns in research should be explicitly modeled in relationship to flows of knowledge between and among different economic agents.

At about the same time, Schankerman and Pakes (1985, 1986) took another original track, using information on fees paid for the renewal of patents in European countries. These data allowed them to estimate the distribution of (private) patent values, as induced by the frequencies of renewal and the magnitude of the renewal fees at every stage. This line of research provided firm empirical evidence on the extent of heterogeneity in patent values, and also a great deal of stimuli for further research using novel aspects of patent data (Pakes and Simpson 1989).

The current volume is a direct outgrowth of and response to this research trajectory. First, we develop the use of patent citations to trace flows of technological knowledge from one inventor to another, thereby implementing Griliches' original suggestion. Second, we use the number and character of citations ultimately received by a given patent to characterize the technological and economic impact of a

given invention. In so doing, we provide an empirically meaningful way of examining the issue of the magnitude of inventions raised by Kuznets. This approach also provides a way to deal with the apparently low "information content" of patent counts found in much of the 1980s econometric work: weighting patents by the number of citations that they later receive produces a much more meaningful measure of inventive output than simple patent counts. Finally, we continue with the tradition of Schmookler and Griliches by providing extensive analysis of how the process that generates the data affects their interpretation, and putting forward econometric techniques for dealing with some of these issues.

3 Overview of the Volume

The volume is organized in four major parts (following this introduction): Part I lays out the conceptual and methodological foundations that underlie the subsequent work. Part II focuses on the use of citations data to explore the geography of knowledge spillovers. Part III contains applications and analysis of particular institutions (e.g., universities), countries and policies. Finally, part IV returns to the patent data themselves, and will be of particular interest to readers planning to use the data in future research: it describes the data in detail, and offers tools to deal with some thorny issues of interpretation that are created by the way the patent and citations data are generated and collected. It also discusses a survey of inventors that sheds light on the meaning of citations.

The three chapters in part I, written in the late 1980s and early 1990s,[5] start with the premise that a patent citation constitutes a (probably noisy) signal of a technological relationship between the citing and cited inventions. Based on this premise, we formulate hypotheses about how the cumulative process of technological development ought to manifest itself in the citations data. Chapter 2 is both chronologically and conceptually the "first" paper in the book. Stemming from Trajtenberg's 1983 Ph.D. thesis, it constitutes the first systematic use (of which we are aware) of patent citation data, using information on patent citations related to a particular line of inventions (CT scanners) collected from paper patent records. It showed

5. Chapter 3, published in journal form in 1997, originally appeared as Trajtenberg, Jaffe, and Henderson 1992.

that the citation-weighted patent count received by CT patented inventions is highly correlated with the social surplus generated by these innovations. Thus, the frequency of subsequent citations turned out to be indicative of the "importance" of the underlying innovations, as measured by the ensuing welfare gains computed on the basis of a discrete choice demand model.

At a basic level, the "success" of this initial empirical exercise spawned the research trajectory reflected in this book, but it also anticipated many of the difficulties and limitations of citations data. First, one had to confront the truncation problem: patents receive citations from subsequent ones over a long period of time (up to several decades), and therefore at any given point in time, when the data are collected, we observe only a fraction of the citations that they will eventually receive. Clearly, older patents would have received a higher fraction of the total number of eventual citations, whereas for more recent patents the truncation problem is more acute. Chapter 2 deals with this issue in a straightforward way, foreshadowing the more systematic statistical approaches that we developed later.

Chapter 2 also illustrates both the value and the difficulty of "external" versus "internal" validation of patent-based measures.[6] By internal validation we mean attempts to substantiate the hypothesized role of patent and citations-based measures as indicators of technological impact by examining patterns and relationships wholly within the patent data themselves. By contrast, external validation substantiates the meaning of the patent-related data by correlating patent-based measures with independent technological or economic indicators whose meaning is more self-evident (e.g., the market value of firms). Construction of the estimates of social surplus associated with innovations in CT scanners (an example of such an independent indicator suitable for external validation) was in itself a major data and econometric task, even though it applied only to one specific case. Application of this method to anything like a comprehensive analysis of industrial innovation would be extremely difficult if not altogether impossible. Nonetheless, the fact that citations were found to be related to a well grounded, independent measure

6. Chapter 12 provides "validation" of a wholly different sort, by asking to what extent citation patterns are consistent with inventors' subjective assessments of spillovers and technological impact.

of importance made the results compelling in a way that is much harder to achieve using just internal validation methods.

Intrigued by the findings in the CT scanners case study, and interested in exploring the use of citations as indicators of knowledge spillovers, we constructed in the early 1990s, together with Rebecca Henderson, a dataset that consisted of all patents granted to universities in 1975 and 1980, and a matched sample of corporate patents.[7] Chapter 3 develops a series of novel citation-based measures, and explores how informative they are of the varied nature of inventions. It is based on the maintained hypothesis that university patents are, on average, likely to be more basic or fundamental in a technological sense than corporate patents, because universities engage in more basic research. And indeed, we found that citation-based measures related to the "basicness" of inventions exhibit higher scores for university patents than for corporate patents. Note that this tests the *joint* hypotheses that (i) citations are a proxy for technological impact, and (ii) university inventions have greater technological impact. Thus, positive findings in this kind of tests may be seen as providing both validation of citations as a proxy for the underlying phenomena, and information on the substantive questions of interest.

A number of themes emerged in this paper that have turned out to be of enduring significance. Thus for example, even if citations are indicative of importance, they do not themselves have any natural "calibration." That is, there is no way to say a priori whether 2 or 10 or 20 citations is "a lot." This is particularly true in the face of truncation, which causes more recent patents to have fewer citations. For this reason, all of our work uses "reference" or "control" groups, in which citation-based measures for a given institution, region or country are compared to the same measures calculated for some appropriate comparison group. Because patent and citation practices vary across technological areas and time, and because of the truncation problem, construction of these control groups must give careful attention to these dimensions of the data. Used first in chapter 3, this "matched sample" approach was refined and extended in several subsequent papers. Chapter 13 provides a systematic analysis of different ways to "benchmark" citation-based measures.

7. This dataset of a few thousand patents was purchased "retail" from a commercial data service. It forms the basis of chapters 3, 5, and 8. The apparent fruitfulness of these data convinced us to undertake the NSF-funded data construction effort encompassing the complete USPTO database.

Chapter 3 introduced for the first time a number of the "constructed" measures of the technological character and significance of innovations that have been applied in later work and which are contained in the attached CD. "Generality" is defined as a measure of the extent to which the citations received by a patent are widely dispersed across technology classes. Holding constant the number of citations, we suggest that higher dispersion of those citations across technologies indicates a wider technological impact, and hence potentially higher social returns. "Originality" is the analogous statistic calculated on the basis of citations made, rather than citations received. It is predicated on the notion that "original" research tends to be synthetic, drawing on previous research from a number of different fields. This chapter also introduced the measurement of rates of "self-citation" (that is, the proportion of citations made by the same assignee as the one owing the cited patent) and conjectured that these rates may reflect the degree of appropriation of potential spillovers from a given invention by the organization that owns it.[8]

Chapter 4 was conceived after Ricardo Caballero served as a discussant of a version of chapter 5 that was presented at an NBER conference in the spring of 1992.[9] It places the concept of knowledge spillovers, proxied by patent citations, within the context of an explicit general equilibrium model of knowledge-driven endogenous growth. It represents one of only a few attempts to link the endogenous growth literature to microeconomic empirical foundations. In so doing, it introduces several methodological innovations that are developed further in subsequent chapters. These include a "citations function" that models the citations generation process as the combined effects of gradual diffusion and of gradual obsolescence. The former causes citation rates to rise as time elapses after an invention, whereas gradual obsolescence causes citation rates to fall as time elapses. It also explores how multiple observations on citing and cited patent cohorts, across time and in different technological fields, can be used to identify empirically the extent to which observed patent rates and citation rates are affected by variations in the propensity to patent and the propensity to make patent citations.

8. Another measure developed there was the extent of reliance on science versus technology, as measured by the ratio of citations to the scientific literature ("non-patent citations") to all patent and non-patent citations appearing on the patent.

9. We thank Olivier Blanchard for brokering the Caballero/Jaffe research collaboration, which made an important contribution to the overall research trajectory.

The second part of this volume focuses on the use of citations data to explore the geography of knowledge spillovers. Prior to this, there had been extensive theoretical analysis of the implications of knowledge spillovers for economic growth, but little was done to give empirical content to this concept. A general skepticism about the difficulty of measuring spillovers is reflected in Paul Krugman's influential book on economic geography: "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" (Krugman 1991, p. 53). Our work has shown that patent citations do constitute indeed a "paper trail" of knowledge spillovers, though one that is incomplete and mixed in with a fair amount of noise. Still, the large volume and wide coverage of patent citations data make them extremely useful for studying the geography of innovation. Chapter 5 was the first paper to demonstrate statistically significant geographic localization of knowledge spillovers as captured by patent citations. Specifically, we found that citing patents were more likely to come from the same metropolitan area, the same state and the same country as the cited patents, relative to a "control" sample of patents that were carefully matched for similarity in time and technological focus to the citing patents. Moreover, we found that localization tends to "fade" over time, that is, as time elapses the geographic differences in citation rates decreases. This finding corresponded well with intuition, and hence gave further credence to the results: whatever initial advantage geographic proximity may offer in terms of knowledge transmission and as stimuli for further knowledge creation, the very "ethereal" nature of knowledge dictates that such advantage should diminish with time. Other dimensions of "proximity" across inventions—technological, institutional, etc.—also appeared to matter. Thus, this paper helped established the notion that knowledge spillovers could after all be traced empirically across geographic and other dimensions, and that the junction between geography and time does matter.

Chapters 6 and 7 examine in more detail the patterns of geographic localization by using much richer data in the context of the citations-function estimation method developed in chapter 4.[10] We were thus

10. Chapter 6 was based still on partial data that limited the amount of parameter flexibility that could be implemented. Chapter 7 is based on the complete data through 1994, allowing examination of more countries and a richer econometric specification in which more parameter variation is allowed.

able to quantify quite precisely differences in spillover flows across countries, and to uncover interesting idiosyncrasies, such as the tendency of Japanese inventors to draw from more recent innovations compared to inventors in other countries. Above all, these papers demonstrate the research potential of this kind of approach for the study of the complex web of knowledge spillovers as they flow across locations, technologies, and time.

Part III deploys patent data to the analysis of the innovative performance of major research institutions (such as universities and national labs), and of particular countries (Israel). In so doing it expands the scope of patents as a research tool both to issues that have clear policy implications, and to units of analysis other than traditional firms and sectors. Chapter 8 examines the changes in the patenting of universities that occurred after the passage of the Bayh-Dole Act in 1980, which was explicitly intended to foster commercialization of university-derived inventions. The rate of patenting by universities was rising even before the passage of the act, but exploded in the 1980s and 1990s. However, using our citations-based measures of importance and generality, we find that the average significance of university inventions actually declined after the early 1980s. Thus, while promoting the *quantity* of technology transfer from universities, the change in policy regime brought about by Bayh-Dole apparently did not improve its *quality*.

Chapter 9 examines patenting by NASA and the rest of the U.S. government. Patents per dollar of federal research expenditure fell from the mid-1960s until the late 1980s. After 1980, federal agencies, with the exception of NASA, reversed this trend and increased their rate of patenting relative to the amount of research performed. Unlike universities, there is no evidence that changes in the ratio of patents to government R&D have been associated with changes in average importance, as measured by patent citations. This paper also contains a qualitative comparison of citation-based indicators of importance and knowledge flows with the inventor's perceptions regarding the underlying inventions, which served as a pilot study for the broader survey of inventors reported in chapter 12. Extending this line of inquiry, chapter 10 looks in detail at the patents and broader commercialization efforts of the "National Laboratories," the relatively large research facilities under the U.S. Department of Energy. As with universities, a number of policy changes in the 1980s and 1990s have sought to encourage commercialization of

technology from the National Labs. We find that these efforts have had some success: patenting rates have risen with no apparent decline in patent quality. Moreover, labs that maintained their technological focus, and those managed by universities, seem to have had the most success.

Chapter 11 uses the patent data and citation measures to evaluate innovation in Israel. Israel ranks high in terms of patents per capita, compared to most of the G7, the Asian Tigers' and a group of countries with similar GDP per capita. Israeli patents are also of high quality in terms of citations received (and getting better over time). Moreover, Israeli inventors patent a great deal in the emerging fields of computers and communications and in biotechnology. On the other hand, Israel ranks low in terms of the percentage of patented innovations that are assigned to local corporations, casting doubt on the ability of the country to fully reap the benefits from those innovations.

The final part of this volume returns to the patent data themselves, and hence it will be particularly useful to readers interested in using these data in future research. Chapter 12 reports on the results of a survey of inventors/patentees, designed to elucidate the extent to which our underlying assumptions about the patent and citation processes conform with the participants' perceptions. Thus, this survey is another form of "validation" of the citations data as a proxy for knowledge flows and technological impact. We surveyed both inventors whose patents were cited, and patentees whose patents made the citations. We find that citing inventors report significant communication with cited inventors (statistically more than with a control sample of inventors), some of it in ways that suggest that a "spillover" took place. However, there is also a large amount of noise in the citations data: about half of all citations do not correspond to any perceived communication, or to a perceptible technological relationship between the two inventions. We also found a significant correlation between the number of citations a patent received and its importance (both economic and technological) as perceived by the inventor.

Chapter 13 provides the basic roadmap to the data, and constitutes a users' guide for the use of the data in research. It describes the patent process and the legal meaning of patent citations. It illustrates basic statistics and trends in the data across technologies and countries and over time. Most important, it considers the econometric

implications of the process that generates patent citations. In particular, it explores biases and interpretational issues that arise from (i) variations in the propensity to patent over time; (ii) truncation of the patent series because we observe only patents granted before some cutoff date; and (iii) truncation in the citations series because we observe citations for only a portion of the "life" of an invention, with the duration of that portion varying across patent cohorts. It discusses possible econometric solutions to these problems, the identifying assumptions required, and pending problems that arise in implementing these solutions. This chapter is a "must read" for anyone contemplating doing analysis of their own with the data.

4 Linking Out: Market Value and Patent Citations

As already mentioned, a great deal of the work presented in this volume relies exclusively on data contained within the patent records themselves. However, many of the data items in patents offer exciting opportunities for linking them out with external data, which can greatly enrich the analysis. Thus, for example, the location of inventors or of assignees can be linked with geo-economic characteristics of their SMSAs, states, or countries; patent classes can be linked to industrial sectors; non-patent citations can be linked to scientific sources; application or grant dates can be linked to any relevant economic time series, etc. One of the potentially most fruitful linkages is through the identity of the assignee: after all, patents are meant to ensure some degree of appropriability to the owner of the patent rights, and hence at least the private value of the patented innovations should somehow be revealed in the performance of the assignees; likewise, the impact of spillovers as traced by citations should also manifest itself in that context.

The big stumbling block was matching the assignee names as they appear on the patent records (currently over 175,000 of them), to any external list of corporations: those names are not entirely standardized, companies take patents under a variety of different names (including subsidiaries), and there are just plenty of errors in spelling. Responding to this formidable challenge, Bronwyn Hall undertook to match the assignee names with the company names in the Compustat database of financial reporting data, and to further consolidate them following their mergers and acquisitions history. The result is a matched set of almost 5,000 publicly traded firms, that

allows one to link between the majority of patents (about 2/3 of all patents assigned to U.S. inventors up to the early 1990s) and the corporations that own them.[11]

The first use of this invaluable resource was a recent project that we have done with Bronwyn Hall, on the relationship between patents, patents citations and the stock market valuation of firms (Hall, Jaffe, and Trajtenberg 2001). Previous work along those lines have found that patent counts add little to market value after R&D is included in a Tobin-q type equation. Here however we find a significant relationship between *citation-weighted patent stocks*, and the market value of firms. The market premium associated with citations appears to be due mostly to the high valuation of the upper tail of cited patents, as opposed to a smoother increase in value as citation intensity increases. After controlling for R&D and the unweighted stock of patents, there is no difference in value between firms whose patents have no citations, and those firms whose patent portfolio has approximately the median number of citations per patent. There is, however, a significant increase in value associated with having above-median citation intensity, and a substantial value premium associated with having a citation intensity in the upper quartile of the distribution. This confirms the finding in the CT scanners study reported in chapter 2, namely that value increases with citation intensity, apparently at an increasing rate. It is also consistent with the conventional wisdom about the innovation process more generally, that a large fraction of the value of the stream of innovations is associated with a small number of very important innovations.

This paper also extends our understanding of the relationship between citations and value by examining the differential impact of self-citations (i.e., citations from patents assigned to the same firm). On average, self-citations are associated with about *twice as much* market value as citations from others. This confirms the conjecture of chapter 3 that self-citations, because they represent subsequent building on the invention by the original firm, are indicative of the firm′s capturing a larger share of the overall social value of the invention. Thus, the evidence shows that both social and private values are increasing in the citations intensity, apparently with increasing returns, and that a high rate of self-citation is indicative of a larger fraction of social returns accruing to the innovating firm.

11. The match file is included in the attached CD; see also chapter 13.

5 What's Next? Thoughts on the Future Research Agenda

Much of the work in this volume was meant to develop the infrastructure in terms of data, concepts, and method to allow the massive use of patents and citations in run-of-the mill economic research, to "validate" the novel measures used, and to demonstrate the viability of the whole approach. In so doing we responded to the aforementioned challenge posed by Kuznets (1962), and, we hope, made a few strides forward along the path initiated by Schmookler and followed by Griliches' NBER program. This is then just "the end of the beginning," to paraphrase Churchill's prophetic dictum: We have barely scratched the surface in terms of what could be done with these data and approach, in order to further elucidate the economics of technical change. Indeed, we have still opened only a small "window" into the mostly impenetrable "black box" of technological change, which has grown enormously in importance as the wonders of innovation engulf more and more of the economy. Here we offer a few additional thoughts on what we hope will be a vibrant research agenda.

The interactions among research spillovers, firms' efforts to appropriate the returns to their inventions, and observed citations within and across firms, constitute a fruitful area for further study. Our work so far has been entirely non-strategic, taking citation patterns as exogenous evidence of spillovers and/or cumulative innovation that is internal to the firm. The next step would be to marry the citations data to a model of innovation and competition, in which firms and their competitors choose levels of R&D, of effort in learning about the work of others, and of follow-up development when initial innovations appear promising. The existence of detailed data on inventions and citations, and the link between these and detailed data on firms, potentially offers an unusual opportunity for empirical testing of a rich strategic model.

Stock market valuation is, of course, only one dimension of value or importance. Harhoff et al. (1999) have confirmed the relationship between patent citations and "value" using survey-based measures of the value of specific important inventions. Recent work by Jenny Lanjouw and Mark Schankerman (1999) explores the information content of patent citations relative to other indicators also derived from patent data, and examines the relationship of these measures to other economic variables. They construct composite measures of patent "quality" based on the number of citations received, the number

of citations made by the patent, the number of claims in the patent, and the number of countries in which patent protection is sought ("family size"). They show that this measure is related to the likelihood of patent renewal and patent litigation, and to measures of the economic significance of a patent to its owner. Finally, they show that the quality-adjusted rate of patenting by firms exhibits a more stable relationship to firm's R&D expenditure than simple patent counts.

There is clearly room for further work on the meaning of and relationships among these different indicators of quality, importance, and value. An important issue is the inter-relationships among the technological significance of an invention, the spillovers that it generates for future inventors, and the value of the invention to its owner. It remains to be seen whether the different measures of patent quality can shed light on these issues (beyond the self-citation effect mentioned above). One aspect of this is variations is patent "size," in the sense of different uses or applications for a single idea, as distinct from the intrinsic significance of the idea.

We have shown that citations exhibit an interesting geographic pattern: initial localization that fades over time. But there is much more that could be done to further explore these patterns. How important are "border" effects (continents, countries, metropolitan areas) as distinct from physical distance? Does language matter systematically? How about historical, social and cultural connections? For example, recent work by Hu and Jaffe (2001) shows that Korea is "closer" (in terms of frequency of patent citations) to Japan than to the United States, and Korea is much closer to Japan than Taiwan is to Japan. These relationships appear to be consistent with patterns of institutional and historical connections in these pairs of economies. As shown in Chapter 7, the large size and great detail of the patent dataset makes exploration of these kinds of questions feasible even for relatively small countries.

In addition to further empirical explorations, much could be gained by refining the modeling of the underlying processes. In principle, spillovers from an innovating unit A (e.g., inventor, firm, country) to another unit B benefit the latter by facilitating invention in B. This means that the rate of innovation and hence of patenting in B, as well as the citations made by B, are endogenous to the spillover process itself. Our work to date has ignored this, taking the flow of citations *to* A as telling us something about the importance of inno-

vations in A, but not being themselves generated or stimulated *by* A. Modeling of this process would allow us to start thinking about connecting the empirical analysis of citations once again to the overall rate of productivity improvement and economic growth. The model in chapter 4 links R&D, innovation, citations and growth over time at an aggregate level within a single country. It would be very interesting to extend this approach to endogenous growth with spillovers among industries and countries.

Another idea in chapter 4 that has not been pursued is endogenous obsolescence. A patent that is highly cited is presumed "important," but it would also seem that the accumulation over time of many patents building on a given invention would eventually make it *less* valuable, at least in the private sense. In principle, it should be possible to implement a dynamic model of the process that might be able to shed light on the rate of private obsolescence of knowledge, and how that varies across different technologies or industries, as well as over time.

Finally, despite the potentially rich detail in the classification of patents by technology-based patent class, we have looked only in limited ways at spillovers across technologies. Just as with geography, there is significant localization, in the sense that citations are more likely to patents in the same class than in other classes. But does this fade over time? Are there particular classes that have unusually large spillovers, in the sense of greatly impacting "distant" technological areas as much as closely related ones? In principle, analysis of this kind offers the potential to test for and explore the significance of "general purpose technologies" (Bresnahan and Trajtenberg 1995).

It has been a major theme of the National Bureau of Economic Research since its inception that good economic research depends on the generation of appropriate and reliable economic data. It is generally agreed that the twenty-first century economy is one in which knowledge—particularly the technological knowledge that forms the foundation for industrial innovation—is an extremely important economic asset. The inherently abstract nature of knowledge makes this a significant measurement challenge. We believe that patents and patent citation data offer tremendous potential for giving empirical content to the role of knowledge in the modern economy. We hope that by constructing the NBER Patent Citations Data File, demonstrating some of the uses to which it can be put, and making it available to other researchers, we can provide a broader and deeper

measurement base on which research on the economics of technological change can prosper.

Appendix

United States Patent ***4,440,871***

Lok et al. ***Apr. 3, 1984***

Crystalline Silicoaluminophosphates

Abstract

A novel class of crystalline microporous silicoaluminophosphates is synthesized by hydrothermal crystallization of silicoaluminophosphate gels containing a molecular structure-forming templating agent. The class comprises a number of distinct species, each with a unique crystal structure. The compositions exhibit properties somewhat analogous to zeolitic molecular sieves which render them useful as adsorbent or catalysts in chemical reactions such as hydrocarbon conversions.

Inventors: Lok, Brent M. (New City, NY); Messina, Celeste A. (Ossining, NY); Patton, Robert L. (Katonah, NY); Gajek, Richard T. (New Fairfield, CT); Cannan, Thomas R. (Valley Cottage, NY); Flanigen, Edith M. (White Plains, NY).

Assignee: Union Carbide Corporation (Danbury, CT).

Filed: Jul. 26, 1982

Intl. Cl.: B01J 27/14

[Some of the] Current U.S. Cl.: 502/214; 208/114; 208/136; 208/138; 208/213; 208/254.H; 585/418; 585/467; 585/475; 585/481

Field of Search: 252/435, 437, 430, 455 R; 423/305; 501/80

[*Some of the*] *References Cited*

U.S. Patent Documents

4,158,621	Jun., 1979	Swift et al.	252/437 X
4,310,440	Jan., 1982	Wilson et al.	423/305 X
4,364,839	Dec., 1982	McDaniel	252/430

[*Some of the*] *Foreign Patent Documents*

984502 Feb., 1965 GB

Other References

"Phosphorus Substitution in Zeolite Frameworks," E. M. Flanigen et al. (1971), [Advances in Chemistry Series No. 101-ACS].

Primary Examiner: Wright, William G.

38 Claims, 3 Drawing Figures

References

Abramovitz, Moses (1956). "Resource and Output Trends in the United States Since 1870." *American Economic Review*, May 1956.

Bound, John, C. Cummins, Z. Griliches, B. Hall, and A. Jaffe (1984). "Who Does R&D and Who Patents?" In Griliches 1984.

Bresnahan, Timothy, and M. Trajtenberg, "General Purpose Technologies: 'Engines of Growth?'" (1995). *Journal of Econometrics, Annals of Econometrics* 65, no. 1: 83–108.

Cockburn, Iain, and Z. Griliches (1988). "Industry Effects and Appropriability Measures in the Stock market's Valuation of R&D and Patents." *American Economic Review* 78 (May): 419–423.

Griliches, Zvi, ed. (1984). *R&D, Patents and Productivity*. University of Chicago Press.

Griliches, Zvi (1990). "Patent Statistics as Economic Indicators." *Journal of Economic Literature* 28, no. 4: 1661–1707.

Griliches, Zvi (1994). "Productivity, R&D, and the Data Constraint." *American Economic Review* 84, no. 1: 1–23.

Griliches, Zvi, B. Hall, and A. Pakes (1991). "R&D, Patents and Market Value Revisited: Is There a Second (Technological Opportunity) Factor?" *Economics of Innovation and New Technology* 1: 183–202.

Hall, Bronwyn, C. Cummins, E. Laderman, and J. Mundy (1988). "The R&D Master File Documentation." NBER Technical Working Paper, no. 72.

Hall, Bronwyn, Z. Griliches, and J. Hausman (1986). "Patents and R&D: Is There a Lag?" *International Economic Review* 27, no. 2: 265–283.

Hall, Bronwyn, A. Jaffe, and M. Trajtenberg (2001). "Market Value and Patent Citations: A First Look." University of California, Berkeley, Department of Economics, Working Paper, no. E01-304.

Harhoff, Dietmar, F. Narin, F. M. Scherer, and K. Vopel (1999). "Citation Frequency and the Value of Patented Inventions." *Review of Economics and Statistics* 81, no. 3: 511–515.

Hu, Albert G., and A. Jaffe (2001). "Patent Citations and Knowledge Flow: Korea and Taiwan." NBER WP, no. 8528.

Jaffe, Adam (1986). "Technological Opportunity and Spillovers of R&D: Evidence from Firms' Patents, Profits and market Value." *American Economic Review* 76: 984–1001.

Krugman, Paul (1991). *Geography and Trade*. MIT Press.

Kuznets, Simon (1962). "Inventive Activity: Problems of Definition and Measurement." In Nelson 1962.

Lanjouw, Jean O., and M. Schankerman (1999). "The Quality of Ideas: Measuring Innovation with Multiple Indicators." NBER WP no. 7345.

Nelson, Richard R., ed. (1962). *The Rate and Direction of Inventive Activity: Economic and Social Factors*. Princeton University Press.

Pakes, Ariel (1985). "On Patents, R&D and the Stock Market Rate of Return. *Journal of Political Economy* 93: 390–409.

Pakes, Ariel, and M. Simpson (1989). "Patent Renewal Data." *Brookings Papers on Economic Activity (Micro)*, 331–410.

Schankerman, Mark, and A. Pakes (1985). "The Rate of Obsolescence and the Distribution of Patent Values: Some Evidence from European Patent Renewals." *Revue Economique* 36: 917–941.

Schankerman, Mark, and A. Pakes (1986). "Estimates of the Value of Patent Rights in European Countries During the Post-1950 Period." *Economic Journal* 96: 1052–1076.

Schmookler, Jacob (1966). *Invention and Economic Growth*. Harvard University Press.

Schmookler, Jacob (1972). *Patents, Invention, and Economic Change: Data and Selected Essays*. Edited by Zvi Griliches and Leonid Hurwicz. Harvard University Press.

Solow, Robert M. (1957). "Technical Change and the Aggregate Production Function." *Review of Economics and Statistics*, August 1957.

Trajtenberg, Manuel, A. Jaffe, and R. Henderson (1992). "Ivory Tower versus Corporate Lab: An Empirical Study of Basic Research and Appropriability." NBER WP, no. 4146.