Paper to be presented at the DRUID 2012

on

June 19 to June 21

at

CBS, Copenhagen, Denmark,

Identifying Breakthroughs: Cognitive vs. Economic

Sarah Kaplan
University of Toronto
Rotman School
sarah.kaplan@rotman.utoronto.ca

Keyvan Vakili
University of Toronto
Rotman School
keyvan.vakili08@rotman.utoronto.ca

Abstract

Previous research on breakthrough innovations has used patent data to identify them and assess their impact. The main proxy for breakthroughs uses forward citation counts, where patents at the top of the distribution are considered breakthroughs. Scholars have found this metric correlates with the economic value of patents (i.e., stock market valuations), yet, it does not tell us much about their technological content. We propose a new methodology (topic modeling of patent texts) to distinguish cognitive from economic breakthroughs. In our test case analysis of 2,826 nanotechnology patents, we find that cognitive breakthroughs are more likely to be highly cited, yet the mechanisms that produce cognitive and economic breakthroughs are quite different. Moreover, patents that are cognitive as well as economic breakthroughs have a bigger and more enduring impact on future inventions. This approach gives us traction in understanding the emergence and evolution of technologies over time.

Jelcodes:O31,-

Identifying Breakthroughs: Using Topic Modeling to Distinguish the Cognitive from the Economic

Abstract:

Previous research on breakthrough innovations has used patent data to identify them and assess their impact. The main proxy for breakthroughs uses forward citation counts, where patents at the top of the distribution are considered breakthroughs. Scholars have found this metric correlates with the economic value of patents (i.e., stock market valuations), yet, it does not tell us much about their technological content. We propose a new methodology – topic modeling of patent texts – to distinguish cognitive from economic breakthroughs. In our test case analysis of 2,826 nanotechnology patents, we find that cognitive breakthroughs are more likely to be highly cited, yet the mechanisms that produce cognitive and economic breakthroughs are quite different. Moreover, patents that are cognitive as well as economic breakthroughs have a bigger and more enduring impact on future inventions. This approach gives us traction in understanding the emergence and evolution of technologies over time.

Keywords: breakthroughs; patents; invention; innovation; cognition; topic modeling; text analysis; nanotechnology

9326 words

Identifying Breakthroughs: Using Topic Modeling to Distinguish the Cognitive from the Economic

1. Introduction

Breakthrough innovations are important because they are the potential basis for new technological trajectories, new sources of societal impacts (both benefits and costs) and new bases for the competitive advantage of firms. Research on innovation and technologies has sought to identify breakthroughs and to determine their sources. In these studies, the use of patent statistics has been a primary means for answering these questions, where breakthrough innovations have typically been measured by the number of "forward citations" (prior art citations made to the focal patent by subsequent patents). Higher numbers of citations, often at the top 5 percent of the distribution, are said to indicate that a patent represents a breakthrough innovation (Trajtenberg 1990). Citations rates have been shown to correlate strongly with a number of economic value assessments (Griliches 1990), such as inventors' estimates of future economic value (Harhoff et al 1999), patent renewal fee payments (Harhoff et al 1999), and firms' stock market values (Hall, Jaffe & Trajtenberg 2005; Deng, Lev & Narin 1999).

Using these forward citation measures, scholars have aimed to determine the sources of innovative breakthroughs, primarily arguing that inventor- and invention-level mechanisms that allow for more diverse combinations of knowledge will produce more innovative outcomes. For example, Singh and Fleming (2010) found that teams of researchers are more likely to generate path-breaking inventions than individual researchers, and suggest that this is because that teams provide more diverse inputs. Fleming (2001) as well as Hall (Hall, Jaffe & Trajtenberg 2001, Hall 2002) argued that breakthrough innovations are those that are the product of more creative processes that allow researchers to recombine more diverse knowledge domains in new ways. Similarly, various scholars have found that firms that access distant knowledge are more likely to produce breakthrough innovations (Ahuja & Lampert 2001; Phene, Fladmoe-Linquist & Marsh 2006; Rosenkopf & Nerkar 2001). Gittelman and Kogut (2003) show that individual patents that have higher levels of citations to scientific papers, which they argue measures the degree to which patents incorporate a substantial amount of scientific knowledge, are also more likely to be cited.

Interestingly, this research bases its arguments on cognitive mechanisms. Each of the identified constructs – teams, the prior experience of inventors, and the breadth of a patent's citations to distant knowledge domains – have assumptions about cognitive processes – such as improving search, increasing diversity of knowledge inputs or breaking existing mindsets – implicitly (or explicitly) embedded within them. Yet, the measure of breakthroughs to which we have been constrained is one of citation counts to patents which is mainly associated with economic value.

This suggests opportunities to use patents to capture the emergence of new technological ideas (what we will call "cognitive breakthroughs") separate from identifying economic breakthroughs, and to examine the sources and impact of these cognitive breakthroughs. There are several reasons to believe that achieving such cognitive breakthroughs may not be wholly consistent with achieving economic breakthroughs. A novel idea embodied in a patent still depends on several other factors to achieve economic impact, including the reputation of the inventor(s), the distribution of the idea in the relevant network, the presence of the complementary technologies at the time, and appropriate articulation of the idea. In the absence of such factors, a patent that represents a cognitive breakthrough may not get economic traction. Similarly, not all highly cited patents, i.e. economic breakthroughs, are necessarily cognitive breakthroughs. Patents with a broad scope and vague claims, patents inside patent thickets (dense networks of patents with overlapping claims), and patents that make an original idea more understandable and usable and or that distribute it strategically in a relevant network, all may lead to a high level of citations regardless of whether the patent introduces a breakthrough idea or not.

In this paper, we propose a new approach to identify cognitive breakthroughs as a complement to citation-based measures, allowing us to separate out the cognitive from the economic. We introduce a computer science approach to text analysis called "topic modeling" to understand the ideas (topics) represented in patents by looking at the association of different words in the abstracts of the patent documents. In our analysis, the formation of a new topic in the patent data can be seen as the emergence of a new area of innovative activity (be it a new method, a new composition, or a new application). We measure cognitive breakthroughs as those first patents that introduce new topics to the field. This measure has the advantage of providing an *ex ante* perspective on the development of a technological field, while

citation-based measures are *ex post* assessments only assigned in subsequent years.

Our approach takes seriously the idea put forth by Griliches (1990) and pursued in recent studies (Alcácer & Gittelman 2006; see also Alcácer, Gittelman & Sampat, 2009; Benner & Waldfogel, 2008; Hegde & Sampat, 2009; Jaffe, Trajtenberg & Fogarty, 2000; Tan & Roberts, 2010) that patents, while useful as indicators of innovation, should also be evaluated as historical documents produced by inventors, prosecuted by patent attorneys and evaluated by patent examiners. An implication is that it should be useful to focus our attention on what the authors of the patents write in these documents. Doing so allows the patent texts to represent contemporaneous interpretations of the technology, which avoids problems of retrospection in tracking the emergence of new technological arenas. It also offers a complementary approach to the use of established patent classification systems because such classification systems lag the creation of new technological fields.

In order to develop and validate this approach, we examine the formation of new themes or topics in one technological field, that of Buckminsterfullerenes (and the related area of carbon nanotubes). This domain of nanotechnology is a useful setting because fullerenes can be seen as a "general purpose technology" (Bresnahan & Trajtenberg, 1995; Helpman, 1998) with potential applications in many domains. New understandings of the technology could therefore create new technological trajectories for research and development.

To examine the potential value of this alternative metric for breakthroughs, we replicate prior studies examining the sources of breakthroughs and show that cognitive breakthroughs significantly predict future citation rates even when controlling for the alternative explanations. Thus, cognitive breakthroughs are more likely to become economic breakthroughs in the future. On the other hand, factors that have been shown to predict economic breakthroughs appear less likely to be associated with cognitive breakthroughs, suggesting that these two types of breakthroughs are related but the mechanisms that produce them are quite different. Moreover, we find that patents that are both economic and cognitive breakthroughs have a significantly bigger impact on future inventions compared to those that are only considered economic breakthroughs. Cognitive breakthroughs also have a flatter citation age profile suggesting that they have a longer lasting impact on future technologies. This approach to the analysis of

patents gives us new traction in analyzing breakthroughs and in understanding the emergence and evolution of technologies over time.

2. Topic modeling of patent texts

2.1 Background on topic modeling

The methodological move made in the study reported here is to treat patents as historical documents written by particular human beings, in particular places, at particular times. Studying the language in the documents should provide a reading of the cognitive content of the patent. The idea behind the use of textual (or content) analysis – an idea that has recently been rearticulated by Duriau and colleagues (2007) – is that language is tightly connected with human cognition. This is the Whorf-Sapir hypothesis (Sapir, 1944; Whorf, 1956) from which many content analysis techniques are drawn. In management studies, this idea has been adapted methodologically to use groups of words to represent important themes (Abrahamson & Hambrick, 1997; Huff, 1990; Kaplan, Murray & Henderson 2003).

We measure interpretations using the text in the abstracts of patents to understand how different actors describe what the technology is and could be. Doing this analysis over large numbers of patents requires automated text analysis procedures. Where the concern is specifically about identifying themes and trends, newly developed computer science text analysis techniques such as topic modeling will be most appropriate (Blei, Griffiths, Jordan & Tenenbaum, 2004; Blei & Lafferty, 2007; Mimno, Wallach & McCallum, 2008; Ramage, Rosen, Chuang, Manning & McFarland, 2009). The goal of topic modeling techniques as developed in the computer sciences is unsupervised analysis of text designed both to generate a model that would predict future texts and to provide a representation of the topics in an existing corpus (Chang, Boyd-Graber, Wang, Gerrish & Blei, 2009; Hall, Jurafsky & Manning, 2008). For our purposes, we focus on this second goal – representing topics in a body of text – and we will use these data to track the emergence of new meanings over time and identify the patents that lead to these breakthroughs.

The topic modeling approach we use is based in the Bayesian statistical technique of latent

Dirichlet Allocation (LDA)[1] to determine the meanings of words by looking at co-presence with other words in the same document or block of text (after removing "stop words" such as "the," "and," "that," or "were") (Blei, Ng & Jordan 2003). The approach assumes that there is a latent set of topics within each document and that any word appearing in the document is attributable with some probability to one of these topics. The same word may have different meanings depending on its association with other words in a document. Given a particular corpus of texts, topic models infer a set of topics and identify the words associated with each topic (each weighted by its importance to the topic). The output is a list of topics and vectors of the weight of each topic in each document (in our case, each patent abstract). For our analysis, we used the publicly available "Stanford Topic Modeling Toolbox" developed by the Stanford Natural Language Processing Group and made available in 2009 (Ramage et al 2009).[2] While topic modeling was initially developed for computer science applications, e.g., for improving Internet search algorithms, the Stanford toolbox was adapted specifically with the needs of social scientists in mind.

This method allows the researcher to quantify meaning over large numbers of texts and to identify the emergence of new topics. For computer science applications such as the development of highly refined predictive models for text searches, the best fit model often produces very large numbers of topics.[3] However, Chang et al (2009) show that these best fit models do not produce topics that represent clearly distinct meanings and that smaller numbers of topics make interpretation more feasible and reliable. In our data, the best fit according to machine learning criteria would include thousands of topics, where the distinctions between them are statistically reliable but hard to interpret even by experts in the field. Therefore, we specified in the model the maximum number of topics that would be still be interpretable by

[1] Latent Dirichlet allocation (LDA) is a generative probabilistic model for sets of discrete data (Blei, Ng & Jordan, 2003). This is the most recent approach to semantic or thematic analysis of texts. Using LDA corrects problems with earlier computer science approaches to text analysis. Latent semantic analysis (developed in the 1980s) based on linear algebra cannot capture multiple meanings of a word and can produce clusters of words that are hard to interpret. The correction to this was probabilistic latent semantic analysis (pLSA), which added a probabilistic model and therefore better statistical grounding. LDA adds the further assumption that topic distributions have a Dirichlet prior (a typical assumption in Bayesian statistics). The Dirichlet distribution is a "distribution over distributions" that gives the probability of choosing a group of items from a set given that there are multiple states to consider (it is a distribution over multinomials, just as beta is a distribution over binomials). Furthermore, unlike pLSA, LDA doesn't suffer from overfitting problems that arise from the increase in the number of parameters as the size of the training corpus increases. See Blei, Ng,and Jordan (2002) for more details on LDA and its comparison with other methods.

[2] See http://nlp.stanford.edu/software/tmt/tmt-0.3/ for further description details on the toolbox. As an alternative, one can use the module developed by Grun and Hornik (2011) for R Package.

[3] This often involves holding out part of the corpus, training the algorithm on one portion and then testing its accuracy in modeling the text in the held out portion, see Wallach et al, 2009.

nanotechnologists with specific knowledge of fullerenes and nanotubes, which for our analysis was 100 topics. This provided both statistically and semantically meaningful topics.

2.2. Sample of fullerene and related patents

To test this approach, we focused on a single technical domain, that of buckminsterfullerenes (and the chemically related carbon nanotubes). Prior studies of the emerging field of nanotechnology[4] have found fullerenes and nanotubes to be a useful site for analysis (Wry et al., 2010) because they can be applied in a very broad range of potential applications from medicine to automotive to electronics to sports and therefore can be conceptualized as general purpose technologies (Bresnahan & Trajtenberg, 1995; Helpman, 1998). They have the chemical formula of C_{60} or Carbon 60. Buckminsterfullerenes (also known as fullerenes or "buckyballs" because of their geodesic dome shape) were discovered in 1985 by Dr. Richard Smalley, Robert Curl and Harold Kroto (for which they won the Nobel Prize in Chemistry in 1996). Carbon nanotubes are in the fullerene family and their discovery is attributed to Sumio Iijima of NEC Corporation in 1991. Methodologically, the choice of fullerenes and nanotubes is appropriate for the application of topic modeling because they are subject to substantial patenting over time and are productive of a multiplicity of uses and interpretations.

To date, very little commercialization of anything technologically significant has taken place in this field, thus we must rely on patents as indicators of potential future commercial applications. These patents show that revolutionary new applications are being developed (e.g., implantable medical devices to control insulin levels for diabetics, more targeted treatments for cancer, structural materials for combat and sports gear, the material for super lightweight batteries and new computing processors that provide quantum leaps in speed and storage capability). Because of this range of potential applications, researchers and managers in organizations have broad purview to guide the research and development of the technology in many directions. As a result, their interpretations of what the technology is and how it might be used have consequences for the development and evolution of the technology. Research and development (and ultimately commercialization) resources will be placed in some areas and not others

[4] Nanotechnologies are "very small" (a nanometer = 1 billionth of a meter) technologies that often have different properties at the nanoscale than they do at larger sizes. Explorations in this field are occurring across a wide range of disciplines (e.g., chemistry, physics, biology, medicine, engineering and materials science). While nanotechnology is recognized as involving an increasing ability to manipulate matter, there is no consensus as to how this will happen or where such inventions could be applied.

depending on the interpretations and choices these researchers make.

To examine the evolution of fullerenes and nanotubes, we collected the 2,826 patents granted by the US patent office through December 2008. To assure our sample is not biased by previous methodologies used to categorize patents, we identified the population of patents using three separate search techniques. First, we selected all utility patents with the terms "fullerene" or "carbon nanotube" in the title, abstract or claims. Second, we used the Derwent technology classifications to select all patents they identify as pertaining to either of these technologies.[5] Finally, the US Patent Office established a nanotechnology "cross reference" patent class (#977) in 2004, which was applied retroactively to all previously-granted patents the USPTO deemed relevant as well as to all new nanotechnology patents. Several of the subclasses pertain to fullerenes and nanotubes (977/735-752). All the patents classified in these categories were selected. We did not use the International Patent Classification system because the most relevant classes (B82B 1/00 and 3/00) were not granular enough to separate out fullerenes and nanotubes from other nanotechnology patents.

Figure 1 demonstrates that no individual sampling technique provided a comprehensive picture of patents that could plausibly be associated with fullerenes and carbon nanotubes, and we believe our approach to developing the population of patents in this field compensates for biases created by any one method of classification. Note in particular that the text search method adds a proportionally large number of patents to our population (1,585 out of the 2,826 identified patents are unique to the text search method). This provides us with a more comprehensive portrait of the emerging field than the published classification systems. These data suffer from right truncation due to the average 34-month lag (with a range of 5 to 98 months) between patent application and patent grant, which we account for in the analyses presented below.

-- Insert Figure 1 about here --

2.3 Deriving fullerene and nanotube topics

For each patent, the abstracts from its US Patent Office document were used in the topic modeling

[5] We selected the following DWPI (Derwent World Patent Index) codes: B05-U; C05- U; E05-U; E31-U02; L02-H04B; U21-C01T; X12-D02C2D; X12-D07E2A; X12-E03D; X16-E06A1A (mainly related to "fullerene type cage structures").

analysis.[6] Analysis of the data shows that in several cases multiple patents with the same abstract have been granted to protect a single invention. To prevent multiple counting of such texts, we grouped patents with identical abstracts and assignees into patent families. This resulted in 2,384 patent families based on the 2,826 patents (there are 336 families, most of which include only 2 patents, with an average of 2.56 and a maximum of 15). Using the "Stanford Topic Modeling Toolbox", we subsequently identified 100 separate topics based on the abstracts. As noted above, the algorithm first created a pool of all the words appearing in all of the abstracts. Based on standard lists of stop words, we removed 788 different words from the abstracts. Using the latent Dirichlet allocation method (Blei, Ng & Jordan 2002), we then identified the topics and the frequency with which each of the words may appear in each topic.

There are two important parameters that are inputs to the LDA module: "topic smoothing" (alpha) and "term smoothing" (beta). The alpha parameter affects the mixing of topics in a document. The value of beta determines the granularity of the model. The Stanford Topic Modeling Toolbox uses 0.1 for both parameters as a default. A smaller beta results in more fine-grained topics (Griffiths & Steyvers, 2004), thus, because we are studying a narrow field of technology, we lowered the beta parameter to 0.01 to acquire more granular results. We kept the alpha parameter at its default value as recommended by the developers.

The algorithm does not provide a summary name for each topic; the topics are simply numbered (though some scholars have experimented with the automatic labeling of topics, this approach is not reliable enough to have been widely adopted, Mei, Shen & Zhai, 2007). Thus, a final step requires topical experts to evaluate the words in the topic and assign a name that most accurately represents the meaning of the topic. In our study, we provided a spreadsheet with each of 100 topics and the top 20 words associated with each topic to 2 field experts (researchers in nanotechnology). They each separately created a short name for each topic. We then met as a group to discuss differences and agreed on the most appropriate name based also on reference to the most important patents in each topic. This allowed us to substantiate the usefulness and the face validity of the topics generated using the automated algorithm.

[6] The abstracts available from Derwent were not used because Derwent often retrospectively modifies the texts of the abstracts they provide in their database, thus making those abstracts less useful when attempting to understand the contemporaneous interpretations of the technology.

Table 1 shows sample topics with the top 20 words associated with each. The algorithm also produces a vector of weights of each topic in each patent abstract. Patents may contain several topics, though of different weights. Figure 2 shows a sample abstract and the weight of its most important topics. For example, this particular abstract for patent number 7352617 "Nano tube cell and memory device using the same" contains topic 92 ("*application of nanotubes in memory devices*") with 92 percent weight and topics *24* ("*application of nanotubes in circuit switch and non-volatile memory*") and 62 ("*Application of nanostructures in energy conversion and energy-related devices*") each with 3 percent weight. Note that the sum of the weights of all the topics for any given patent is 100 percent.

-- Insert Table 1 and Figure 2 about here --

The topics as generated from the text of patents do not give us the same information as that captured by patent office-assigned classifications. The correlation between categories developed using topic modeling and the USPTO assigned technological classes (using primary topics and primary 3-digit patent class) is 0.23 with a standard deviation of 0.10. The average correlation is the average over all the calculated maximum correlation values for each topic with all the possible patent classes. This shows that patents containing the same topic may be classified in several different USPTO patent classes, and similarly, patents in the same assigned patent class may belong to different topics. Topics are generated by the writings of the inventors (and others who help construct the patent) and patent classifications are assigned by patent examiners based on their understanding of what the patent application contains and using previously established classification categories.

2.4 Identifying cognitive breakthroughs

There are many potential analytical uses of the data produced in the topic modeling exercise (we return to this in the conclusion). For the purposes of this study, we focus on the identification of cognitive breakthroughs. To do so, we identify the entry of new topics into the sample. We then select all patents over a threshold weighting for that topic (in our case, 0.2) and appearing in the first year of the topic formation (based on application date). The average number of topic-generating patents using this method is 1.89 per topic for a total of 189 breakthroughs in the dataset (though one topic has 12 separate patents associated with it in the first year). The median is 1.

Note that the selection of breakthroughs is sensitive to the cutoff points we set. With regard to the time frame, we chose 1 year as reasonable estimate of the time for which the knowledge of that invention would not be widespread (given that the average lag between application and granting of a patent in our data is 34 months). Thus, any patents applied for in this 12-month window could be considered simultaneous inventions. The threshold for topic weight is also an important choice. Choosing a high cutoff threshold may result in filtering out the first patents that have actually triggered the formation of a topic. On the other hand, choosing a very low cutoff threshold may result in choosing patents that do not represent a topic in a meaningful manner. After analyzing different cutoff thresholds and comparing the selected patents against the primary topics assigned to them, we selected .20 as the cutoff threshold. As a robustness check, we also repeated the analyses with thresholds of .10, .15, .25, and .30. Using these alternative means of counting cognitive breakthroughs, we found qualitatively similar results in the analyses, although less satisfactory due to under- or over-sampling issues.

3. Sources of cognitive and economic breakthroughs

To understand and validate our cognitive measure of breakthroughs, we compared it to the economic measure based on forward citations. Using the typical criterion of patents in the top 5 percent of the distribution of forward citations in the first 5 years of the patent (since application date), we find that 119 patents in our dataset would qualify as economic breakthroughs, 21 of which are also cognitive breakthroughs. However, the fact that 168 of the cognitive breakthroughs are not economic breakthroughs and 98 of the economic breakthroughs are not cognitive breakthroughs suggests that different forces may be at play in shaping these phenomena (Figure 3).

-- Insert Figure 3 about here --

3.1 Sources of breakthroughs

We identified a number of factors that have previously been shown to be associated with economic breakthroughs as measured by citation rates and test their relationship to cognitive breakthroughs as measured by topic generation. We draw from a wide series of studies using patents to examine different aspects of breakthroughs (Singh & Fleming 2010; Phene et al 2006; Fleming 2001; Trajtenberg et al 1997; Hall et al 2001; Hall 2002; Fleming & Sorenson 2001; Podolny & Stuart 1995;

Gittleman & Kogut 2003; Harhoff et al 2003; Rosenkopf & Nerkar 2001). Because these studies vary as to whether they use a count of forward citations or a dummy variable indicating the "breakthroughs" in the top tier of cited patents as the dependent variable, we examine both outcomes in our analyses.

Research on breakthroughs has focused on the characteristics of the inventors and their social settings and the types of knowledge that have been incorporated in the making of the invention. We briefly describe each of the related studies and how we construct the focal measure in our dataset (as summarized in Table 2 for the characteristics of the inventors and Table 3 for the characteristics of the invention).

--Insert Tables 2 and 3 about here --

3.1.1 Inventor-level sources of breakthrough innovations

Inventors: teams and organizations. Singh and Fleming (2010) showed that economic breakthroughs (whether the patent is in the top 5% of the distribution of forward citations) were more likely to be associated with larger inventive teams due to greater diversity of viewpoints and higher capacity to iterate ideas and select better ones. We follow Singh and Fleming's (2010) lead and measure this as a dummy (*Team*) but in separate analyses (available from the authors) test the count of team members and find similar results. Further, Singh and Fleming (2010) argue that inventors embedded in organizations will be more likely to produce breakthroughs because they are able to draw on a rich amount of knowledge accumulated collectively in the organization. This is measured as dummy variable (*Assigned*) indicating the patent was assigned to (any kind of) organization.

Inventors: experience. Singh and Fleming (2010) also argue that teams with greater inventive experience on average will be better able to create valuable inventions. This is measured as the average number of previous patents by the inventors of the focal patent, using a log normal transformation to deal with the skewness of the data: *Ln(average experience)*. On the other hand, they argue that if teams have a good deal of experience working together, they might get locked into one way of thinking and produce more incremental innovations. This is measured as the number of previous patents by the same team of

inventors: *Ln(joint experience)*.[7]

3.1.2 Invention-level sources of breakthrough innovations

Invention: knowledge distance. Technological distance of the knowledge incorporated into a patent has been shown to correlate with citation rates. The idea is that exploratory or long jump search (Gavetti & Levinthal 2000; March 1991) is more likely to produce inventions that break from the existing technological and scientific models. Several authors have suggested that search in technologically distant arenas can lead to more highly cited patents (Phene et al 2006, Rosenkopf & Nerkar 2001, Trajtenberg, Henderson & Jaffe 1997). We use the technological distance measure proposed by Trajtenberg et al. (1997) as follows:

$$technological\ distance_i = \sum_{j}^{\#\ of\ backward\ citations\ of\ i} \frac{tech\ distance_{i,j}}{\#\ of\ backward\ citations\ of\ i}$$

where $tech\ distance_{i,j}$ is 0 if both patents i and j belong to the same 3-digit technological class, is 0.33 if they are in the same 2-digit class, is 0.66 if they are in the same 1-digit class, and 1 if they are in different 1-digit classes.

Invention: knowledge combination. Drawing on Schumpeterian logic (Schumpeter 1939) and theories of creativity (Hargadon & Sutton 1997), various scholars have suggested that it is not simply the distance of the knowledge that is incorporated but how it is combined that matters. They argue that breakthroughs are more likely to be the product of combinations of different knowledge domains (Trajtenberg et al 1997; Hall, Jaffe & Trajtenberg 2001) or, even more specifically, unique combinations of knowledge domains that have not been made in the past (Fleming 2001).

Hall, Jaffe and Trajtenberg (2001) examined recombination based on a Herfindahl index of citation concentration in a measure they called "*patent originality*." The greater the concentration of patent classes in the prior art cited by a focal patent, the less "original" is the patent. Further, according to Hall (2002), we adjusted the measure to correct for the downward bias for patents with few citations. Thus, we

[7] Note that Conti et al (2011) argue that greater experience of an individual inventor will increase the rate at which an inventor will produce breakthroughs but that this is driven by an increase in the productivity of the inventor (due to well-honed heuristics for inventing and patenting), while the likelihood that any one invention is a breakthrough will decrease with experience due to lock in to one way of thinking. Our results are not consistent with this finding, however, Conti et al have access to survey data that allows them to control for a whole host of additional inventor characteristics such as age, gender and mobility that allow for a more precise characterization of patenting experience.

measure originality as:

$$\widehat{Patent\ originality_i} = \frac{number\ of\ patents_i}{number\ of\ patents_i - 1}(Patent\ originality_i), where\ patent\ originality_{\ i} = 1 - \sum_j^{n_i} s_{ij}^2$$

where s_{ij} denotes the percentage of citations made by patent i to patents in class j, out of n_i patent classes. A high value of patent originality shows that inventors have combined different ideas from different topics in the field.[8] For patents that cite no prior art, patent originality cannot be calculated. Therefore, we also include a dummy (*No prior art*). Note that Ahuja and Lampert (2001), when examining patent references at the firm level, suggest that the lack of prior art is an indicator of a *de novo* technological solution, one that is "pioneering," and hypothesize that pioneering technologies are more likely to be breakthroughs.

Fleming (2001) has extended these ideas to examine components and types of combinations, inferring implications about the knowledge of the inventors. He suggests that if inventors are familiar with the components of an invention, they will be more able to select and recombine those components into useful inventions (as measured by forward citation counts). Familiarity is inferred from how frequently and recently patent subclasses have been used previously by any researcher. This variable – *Ln(component familiarity)* – is measured as the average time-discounted count of previous usage of the focal patent's subclasses across all patents listed by the USPTO. Following Fleming's (2001) formulation:

$$Average\ component\ familiarity\ of\ patent\ i = \frac{\sum_{all\ subclasses\ j\ of\ patent\ i} I_{ij}}{\sum_{all\ subclasses\ j\ of\ patent\ i} 1}$$

$$where \quad I_{ij} = \sum 1\{patent\ k\ uses\ subclass\ j\} \times e^{-\left(\frac{app.date\ of\ patent\ i - app.date\ of\ patent\ k}{time\ constant\ of\ knowledge\ loss\ (5\ years)}\right)}$$

Similarly, Fleming (2001) argues that the more recently and frequently specific combinations of subclasses have been used, the more likely inventors will be to refine these combinations into useful inventions. This variable – *Ln(combination familiarity)* – is measured as the time-discounted count of the

[8] We also calculated this measure using topics rather than patent classes. To do so, we needed to assign topics to the patents that were cited by the patents in our original sample. Since our main topic modeling analysis was done only for the 2,826 core fullerene and nanotube patents, we first needed to develop topics for all of the cited patents (the backwards citations). We used an extended sample that included the fullerene and nanotube patents plus all of their backward citations for a total of 17,735 patents. Then, using the same topic modeling method, we identified a new set of 100 topics and assigned a primary topic to each of the patents in the extended sample. With topics assigned to all the cited patents, we then were able to calculate the "topic originality" for each of the patents in the core sample using the same methodology as Hall (2002) did for patent classes. These two measures are correlated at 0.70, suggesting that while specific topics are not highly correlated with specific classes (as mentioned above), the degree of recombination of knowledge is. In regression results not reported here but available from the authors, we find very similar effects for the two measures of originality.

previous use of the focal patent's particular subclass combination across all patents listed by the USPTO:

$$cumulative\ comb.use\ of\ patent\ i = \sum_{\substack{all\ patents\ k\ granted \\ before\ patent\ i}} \left[1\{patent\ k\ used\ same\ comb.of\ subclasses\ as\ patent\ i\} \times e^{-\left(\frac{app.date\ of\ patent\ i - app.date\ of\ patent\ k}{time\ constant\ of\ knowledge\ loss\ (5\ years)}\right)}\right]$$

On the other hand, Fleming (2001) suggests that too much cumulative use of a combination may mean that it has been exhausted of its usefulness, and therefore such combinations would be less likely to be useful. This variable – *Ln(cumulative combination)* – is the same as combination familiarity but without the time discount.

Invention: knowledge intensity. Scholars have also suggested that the degree to which a patent draws on basic scientific knowledge is also associated with its future impact. This has typically been measured as a count of the "non-patent references" listed by a focal patent (Gittelman and Kogut 2003; Deng, Lev & Narin 1999), and we follow this approach in calculating *# non-patent references*. This is admittedly a noisy measure, because non-patent references can include manuals, presentations, and other documents in addition to publications in scientific journals.

Controls. We include three other measures as controls because they have been shown to be associated with the forward citations garnered by patents. We control for the total number of patents cited as prior art (*# domestic references*) because it is assumed that patents that cite more will also be cited more (Fleming & Sorenson 2001; Podolny & Stuart 1995). We also control for the number of claims in the focal patent (*# claims*), because it has been argued that the greater the scope of the patent, the more likely the invention will receive future citations (Singh & Fleming 2010). Finally, we control for *family size*, where the family is the set of patents that contain identical abstracts and assignees and therefore are assumed to represent a cluster of patents around a single invention. We assume that patents in large families will be more likely to receive higher numbers of future citations (this is related to arguments by Cockburn & Henderson 1998; Gittleman & Kogut 2003; Harhoff et al 2003, who measure patent families as patents that are patented in multiple jurisdictions). Annual time dummies are included as a simple control for possible time trends.

3.2 Descriptive statistics

Table 4 shows the descriptive statistics. The majority of the variables have the same means and standard deviations as reported in other papers. The mean and standard deviation of component and combination familiarity measures are above what reported in Fleming (2001) mainly due to the fact that we are looking at a much more recent sample of patents which means we observe a substantially longer history for the observed subclasses and combinations in the sample. The high correlation between combination familiarity and cumulative combination is due to their very similar definitions. Fleming (2001) reports the same high correlation and suggests that because of the large sample properties of maximum likelihood estimators, this should not be an issue except for potentially inflated standard errors. Nonetheless, our results for these variables are generally similar to what Fleming (2001) reports.

-- Insert Table 4 about here --

3.3 Results examining the sources of economic vs. cognitive breakthroughs

We test the effects of these variables on the standard measure of economic breakthroughs – a dummy variable indicating whether a patent is in the top 5 percent of cited patents (based on a count of forward citations in the first 5 years after application date) – in a simple logit model (Models 1 and 2 in Table 5). Because scholars have sometimes used citation counts to measure the impact or usefulness of patents, we examine the 5-year count of forward citations as a dependent variable separately in a negative binomial count model (Models 3 and 4). We constrain our dataset to the time period 1991-2005 in order to account for the potential effects of right truncation of citation counts for more recently granted patents. We report the odds ratios and incident rate ratios associated with each of the selected explanatory variables for each of the two dependent variables, economic breakthroughs and citation counts respectively. Ratios greater than 1 show a positive effect on the dependent variables, and less than 1 indicate a negative effect.

-- Insert Table 5 about here --

Though the samples of the studies whose measures we use are vastly different (in terms of numbers, time periods, technological arenas, etc.), we by-and-large are able to replicate their results in our fullerene patent dataset for both measures of economic impact (Models 1 and 3 in Table 5). To control for

the variance in the grant delay of patents, we assured that these results are robust to the use of a 4-year window since grant date instead of a 5-year window since application date.

Looking at the inventor-level factors, the effects of *team* and *assigned* are unambiguously positive and significant. The results suggest that teams of inventors have 3 times more chance to come up with economic breakthroughs compared to lone inventors. Similarly, a patent assigned to an institution (a firm, university, government, etc.) rather than an individual is almost 3 times more likely to be an economic breakthrough. Average experience is also positive but only significant when the citation count is used as the dependent variable. The result for *joint experience* suggests, if anything, that a patent produced by a team that has substantial experience patenting together will likely garner more citations. While this is counter to the effect proposed by Singh & Fleming (2010), it is consistent with some of their actual regression results.

With regard to the knowledge used in each invention, we find that *technological distance* has a positive, but only marginally significant, effect on the citation count of patents suggesting that exploration of distant areas is indeed associated with more citations. We also confirm the findings that recombinations are positively associated with future citation rates, as the ratios for *patent originality*, *Ln(component familiarity)*, and *Ln(combination familiarity)* are larger than 1 and mainly significant. Further, *Ln(cumulative combination)* – the construct capturing the degree to which a particular combination has been exhausted of potential – has a negative and marginally significant effect on the forward citation rate of patents. We also find that Ln(*# non-patent references)* – used as a proxy for the science intensity of a patent – positively and significantly affects the citation count and the likelihood of economic breakthroughs.

Importantly, as shown in models 2 and 4, when we enter our measure of topic-generating patents into these regressions, we find that they are strongly positively associated with future citation rates, even when controlling for a whole host of other explanations. This suggests that the creation of a new topic is a separate and distinct mechanism from other inventor- or invention-related dynamics studied in the literature to date. If a patent is topic generating, it increases the odds of generating an economic breakthrough almost by a factor of 2. In other words, holding everything else at their means, the

probability of being an economic breakthrough is 0.048 for a topic-generating patent compared to 0.024 for other patents. Also, a topic-generating patent is likely to receive 1.5 times more citations than the average patent.

When we then turn in Model 5 to what factors might be associated with the production of these topic-generating patents (cognitive breakthroughs), we find very different results from those reported for economic breakthroughs and citation counts. At the inventor level, we find little to suggest that collaboration in teams (*team)* or within organizations (*assigned*) leads to cognitive breakthroughs (the coefficients are not significant and the ratios are much closer to 1). On the other hand, we do find some evidence, much as for economic breakthroughs, that the joint experience of the team matters, as the calculated ratios for *Ln(joint experience)* are all greater than 1 and significant.

Surprisingly, at the invention level, we find that topic-generating patents do not appear to evolve as the combination of distant knowledge or of patents in different technological classes. Indeed, *technological distance* is negatively associated with cognitive breakthroughs. And, counter to Hall et al (2001), we find that *patent originality* as measured by the dispersion of classes in the prior art, is negatively associated with topic-generating patents. All three coefficients for Fleming's (2001) measures of *component* and *combination familiarity* and *cumulative combination* are not significant. Similarly, *Ln(# non patent references)* (science intensity) is not significant (different from findings by Gittelman & Kogut, 2003, and Deng, Lev & Narin, 1999). Interestingly, the *Ln(# claims)* and the *Ln(family size)* controls also lose their significance in the regressions on topic-generating patents.

Taken together, these results suggest that cognitive and economic breakthroughs are distinctive phenomena that are produced through different processes. Further, introducing new technological ideas (topic generation) is strongly associated with higher future economic impact.

4. The impacts of cognitive and economic breakthroughs on future innovation

Recall from Figure 3 that among the 119 highly cited patents (economic breakthroughs) in our sample, only 21 are identified as cognitive breakthroughs (generating a new topic). There are also 168 patents that are topic generating but not among the top 5 percent cited. Thus, there are also 2,097 patents

that are neither topic-generating nor highly cited. To explore further what this new dimension adds to our understanding of breakthrough technologies, we examine the difference in pacts of the patents in each of these groups on future innovations.

To address this question, we analyze and compare the forward citation patterns of these four types of patents for the first generation (forward citations to the focal patents) and the second generation (forward citations to patents citing the focal patents), following the method proposed by Mehta, Rysman, and Simcoe (2010) and Rysman and Simcoe (2008). The goal is to evaluate the citation patterns and rates of each patent adjusted for confounding factors such as cohort effects. Their approach benefits from variance in the processing time of patents from the same application-year cohort (which arises due to variation in the USPTO review process) to identify the age effects. The model hence uses the age since grant and a full set of application and citing year dummies. Following their method, we estimate the first and second-generation citation age profiles for each of the four types of patents using the following model:

$$C_{it} = f(\alpha_y, \alpha_t, \alpha_a, \varepsilon_{it})$$

where C_{it} stands for the number of citations (1st generation or 2nd generation depending on the model) received by patent i in year t (as measured by the application year of the citing patent), α_y and α_t are the fixed effects for the application year and citing year respectively, α_a is the set of citing age effects, and ε_{it} is a patent-year error term uncorrelated with the three sets of fixed effects. $f()$ is defined as a Poisson process.

Since patents may receive citations before they are officially granted, age effects include negative values as well. In practice, we drop any citation age less than 2 years before grant year due to scarcity of such observations. We also cut the citing age at 8 years since there are very few patents more than 8 years old particularly in the top two cells of Figure 3 (highly cited patents of either type). Further, as we move toward the end of the sample, we observe some truncation because we cannot observe the pending patents (not granted yet) with long wait times to grant. Therefore, similar to analyses in the previous section, we limit our sample to those patents applied between 1991 and 2005.

The equation is then estimated separately for each of the four types of patents and for first

generation and second-generation citations. Using the estimation results, we then predicted the expected citation count for each of the four cells for ages between -2 and 8 from application. Figure 4 presents the 1st and 2nd generation citation age profiles for the top cited patents (in the top 5 percent of citations), comparing those that are topic generating vs. those that are not. The results show a significant difference between the two types of patents. For first-generation citations (Figure 4a), starting 1 year after their grant year, top cited patents that are also topic generating consistently receive between 1.3 to 2.7 times more citations compared to other top cited patents. A similar pattern can be seen for the second-generation citations (Figure 4b). Figure 5 shows the comparison results between the two cells in the bottom of Figure 3 (that is, the two types of patents that do not attract high citations, those that are cognitive breakthroughs and those that are not). Again the results are striking. Among the patents in the bottom 95 percent citation rate, the topic-generating patents receive more than twice as many citations as do other patents. These results are robust to the use of 4-year window since grant date rather than a 5-year window since application to capture a longer citation window and control for different pendency lags.

-- Insert Figures 4 and 5 about here --

The average age of the citations is 1.91 years for top cited topic-generating patents, while it is only 1.44 for other top cited patents. The average age of citations for patents the bottom 95 percent is 3.24 for topic generating and 3.12 for non-topic generating. Thus, topic-generating patents (cognitive breakthroughs) not only receive relatively more citations after their grant year, they also receive a larger share of their forward citations in later years, suggesting a slower ramp up but a higher and more enduring effect. This result is all the more notable given the results from Hall et al. (2005) suggesting that unexpected future citations (which would produce a flatter age profile) are more valuable than an average citation (Rysman & Simcoe 2008).

5. Discussion and conclusion

5.1 Implications for the understanding of breakthroughs

The topic modeling approach treats patents not as proxies of innovation but rather as historical documents produced by inventors, lawyers, patent examiners and others. By paying attention to how these

actors interpreted the technology, we usefully complement existing research on technology evolution, in particular that which draws on patent data to understand the sources and impact of innovation. The imperfect relationship between topic-generating patents and those that receive high citations also indicates that there are different kinds of "breakthroughs," those that are cognitive and those that are associated with later market value. As a result, our analyses find results that contrast with conclusions drawn previously by scholars looking to determine the sources and trends in inventive activity. In this study, we make two main contributions.

First, we introduce a methodology for identifying cognitive breakthroughs as distinct from economic breakthroughs. This allows us to separate out the contemporaneous emergence of new technological ideas from the *ex post* identification of economically valuable patents. Identifying cognitive breakthroughs and understanding their sources are quite important for several reasons. Cognitive breakthroughs mark the origins of new technological paths. They can disrupt previous technological paths and they can be the source of competitive shifts in industries. Furthermore, identifying cognitive breakthroughs helps us understand different mechanisms through which new ideas spread over time and space and explain why some new ideas become the wheels of economic fortune and some simply grind to a halt after a few years. While most of the empirical evidence has been massed to link forward citations to economic value (Harhoff et al 1999, Griliches 1990, Trajtenberg 1990), the use of citation rates in innovation studies has become so common that the meaning of a highly cited patent has bled over into general attributions of its importance or technological significance (e.g., Albert et al 1991, Fleming 2001). Our results suggest that such attributions can introduce both type I (counting certain patents as breakthroughs when they are not) and type II errors (missing out on some breakthroughs that do not directly attract high citation rates), resulting in biases in the results.

Our method thus highlights the separate but interrelated nature of cognitive and economic breakthroughs. While we find that patents that are cognitive breakthroughs are likely to get higher levels of forward citations (even when controlling for factors previously identified as contributing to economic breakthroughs), not all cognitive breakthroughs are also economic successes and not all economic breakthroughs represent introductions of new technological ideas to the field. Those patents that are both

cognitive and economic breakthroughs are rare (less than one percent of our sample) but appear to have a greater impact on future innovation than any other kind of invention.

Second, we show that the dynamics that have been shown to predict economic breakthroughs do not, in the main, predict cognitive breakthroughs. This is somewhat surprising because many of the variables have been ascribed to cognitive mechanisms. Not only does this suggest that there are unique factors contributing to cognitive breakthroughs but it also pushes us to refine our understanding of what constructs many of the previously identified variables – such as teams, organizations, recombination, or theuse of non-patent references – truly represent.

For example, because *team* and *assigned* are associated with citation rates and not cognitive breakthroughs, it might be assumed that patents that are produced by teams inside organizations may get more citations because of social effects rather than due to the quality of the ideas produced from diverse inputs (Podolny & Stuart 1996; Dahlin & Behrens 2005). The inventors are more likely to be known simply because there are more of them to know and teams of inventors on average have a more extended network compared to an individual inventor. Also, organizations (assignees) often have more presence in the intellectual landscape than individuals. By increasing the chance that others will know about the patent, it is more likely to be used, and therefore cited, over time. As another example, the fact that *# non-patent references* to be positively correlated with citations but not with topic generation suggest a potential alternative explanation of the effect of citations to the scientific literature: such citations may be a means that inventors attempt to signal the novelty or legitimacy of the invention rather than a representation of the underlying science upon which the inventors might have drawn. This is especially true because these types of citations are voluntary and carry no legal implications, unlike citations to prior patents.

The results for the measures of combination are particularly striking. They suggest that patents with a very high concentration of patent classifications in the prior art (low *patent originality*) are more likely to be topic generating, while patents that combine a wide variety of elements (high *patent originality*) are more likely to be highly cited. Perhaps the patents that draw on prior art from a wide range of patent classes (and that have more claims and are in larger patent families) are simply more applicable in a variety of domains and therefore more likely to be used (and cited) in the future. On the other hand,

generating new topics may require deep immersion in one particular domain rather than linking to more distant knowledge or building on familiar components and combinations.

5.2 Extensions of topic modeling as a tool in studies of innovation

In addition to identifying the sources and impacts of cognitive breakthroughs, the topic modeling approach may usefully contribute to other areas of research on science and technology. For example, it enables us to look at the origin of technological paths, something that we cannot normally do with technological classes particularly in nascent fields. The low correlation between topics and patent classes indicates that the classification system is only one means for representing the knowledge embodied in a patent. Indeed, in the case of emerging technologies such as nanotechnology, the classification system cannot capture the point of emergence. As Benner and Waldfogel (2008) note, the use of patent classifications to proxy location in technological space can be problematic. The classifications, as the USPTO points out, "reflect the uneven growth derived from the first general scheme created in 1790" and revised many times since then (USPTO 2005, p. 1), they are "primarily designed to assist patent examiners performing patentability searches" (p. 1) and the classification of patents into these categories are inherently the "subjective" assessments of examiners based on their interpretations of the claims and the rules for making classification decisions (p. 9). And, for nanotechnology class 977, its use by researchers like us is made even more problematic by its status as a cross-reference class[9] and therefore, by the rules of the Patent Office, cannot be included as a primary patent class. Thus, an emerging field such as nanotechnology – for which the USPTO has only created a cross-reference class to facilitate their own search – would not be captured in primary three-digit classification measures that scholars in technology management tend to use.

Topic modeling can also allow us to analyze at a more fine-grained level technological distance, ties and spillovers between entities. To date, this research has primarily been conducted through an examination of the overlaps in USPTO patent classification amongst the patents of the different entities

[9] The USPTO, in response to the launching of the National Nanotechnology Initiative by President Clinton in 2000, created the 977 Cross-Reference Art Collection in an "effort to improve the ability to search and examine nanotechnology-related patent documents" and as a means to document the "collection of issued U.S. patents and published pre-grant patent applications relating to nanotechnology" http://www.uspto.gov/web/patents/biochempharm/crossref.htm, accessed June 8, 2011. Cross-reference classifications cannot be used as a primary patent classification.

(e.g., Jaffe 1986, Ahuja 2000, Song, Almeida & Wu 2003) or citations between entities (e.g., Stuart & Podolny 1996; Mowery, Oxley & Silverman 1996). These vectors of overlaps have been constrained to indicator variables (either there is an overlap or not). However, because topic modeling provides a weight of each topic for each patent, there is an opportunity to evaluate the content of the ties using topics and the strength of the ties using weights. This approach may be a useful complement to patent classes because it tracks the actual language of the actors rather than the classifications assigned by others. It also supplements the cross-citation approach by, first, examining the ideas directly rather than inferring them from citation ties and, second, allowing for the possibility that such connections occur even if specific patents are not cited. It also provides a new tool to track the diffusion of ideas and therefore can contribute to the literature on knowledge and technology spillovers.

These examples are just a few of the possibilities that are created by the application of topic modeling to studies in the management of technology. It is worth noting that the topic modeling methodologies are quite new, even to the field of computer science, and their application to the social sciences is even newer (Ramage et al 2009). The approach used in this paper is the state-of-the-art for which a toolkit is available. New techniques such as correlated topic modeling (which allows for the correlation of topics) (Blei & Lafferty, 2007), dynamic topic modeling (which takes into account the passage of time in computing topics) (Ahmed & Xing, 2009; Wang et al, 2008; Zhang & Wang, 2010) and author-matched topic modeling (Rosen-Zvi et al. 2010), may offer even better representations of topics as they emerge and evolve. Future analyses should take advantage of these approaches as they are refined and as toolkits become available.

References

Abrahamson, E. & Hambrick, D. C. 1997. Attentional homogeneity in industries: The effect of discretion. Journal of Organizational Behavior, 18: 513-532.

Ahmed, A., & Xing, E. P. 2010. Timeline: A Dynamic Hierarchical Dirichlet Process Model for Recovering Birth/Death and Evolution of Topics in Text Stream. In Proceedings of the 26th International Conference on Conference on Uncertainty in Artificial Intelligence (UAI 2010).

Ahuja, G. 2000. Collaboration Networks, Structural Holes, and Innovation: A Longitudinal Study. Administrative Science Quarterly, 45(3): 425-455.

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: 521-544.

Alcácer, J., Gittelman, M., & Sampat, B. 2009. Applicant and Examiner Citations in U.S. Patents: An Overview and Analysis. Research Policy, 38(2): 415-427.

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

Benner, M., & Waldfogel, J. 2008. Close To You? Bias and Precision in Patent-Based Measures of Technological Proximity. Research Policy, 37(9): 1556-1567.

Blei, D. M., Griffiths, T. L., Jordan, M. I., & Tenenbaum, J. B. 2004. Hierarchical topic models and the nested chinese restaurant process. Advances in Neural Information Processing Systems 16, 16: 17-24.

Blei, D. M. & Lafferty, J. D. 2007.A Correlated Topic Model of Science (Vol 1, Pg 17, 2007). Annals of Applied Statistics, 1(2): 634-634.

Blei, D. M.; Ng, A. Y.; Jordan, M. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research 3: pp. 993–1022.

Bresnahan, T. F. & Trajtenberg, M. 1995.General purpose technologies: Engines of growth'? Journal of Econometrics, 65(1): 83-108.

Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S., & Blei, D. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. In Proceedings of *Neutral Information Processing Systems 2009*.

Cockburn, I. M., & Henderson, R. M. 1998. Absorptive Capacity, Coauthoring Behavior, and the Organization of Research in Drug Discovery. The Journal of Industrial Economics, 46(2): 157-182.

Conti, R., Gambardella, A., & Mariani, M. 2011. Learning to Be Edison? Individual Inventive Experience and Breakthrough Inventions. Working Paper. Bocconi University.

Dahlin, K. B., & Behrens, D. M. 2005. When is an Invention Really Radical? Defining and Measuring Technological Radicalness. Research Policy, 34(5): 717-737.

Deng, Z., Lev, B. & Narin F. 1999. Science and technology as predictor of stock performance. Financial Analysts Journal, 53(3) 20–32.

Duriau, V. J., Reger, R. K., &Pfarrer, M. D. 2007. A Content Analysis of the Content Analysis Literature in Organization Studies: Research Themes, Data Sources, and Methodological Refinements. Organizational Research Methods, 10(1): 5-34.

Fleming, L. 2001. Recombinant Uncertainty in Technological Search. Management Science, 47(1): 117-132.

Fleming, L. & Sorenson, O. 2001. Technology as a Complex Adaptive System: Evidence from Patent Data. Research Policy, 30(7): 1019-1039.

Gavetti, G., & Levinthal, D. 2000. Looking Forward and Looking Backward: Cognitive and Experiential Search. Administrative Science Quarterly, 45(1): 113-137.

Gittelman M. & Kogut B. 2003. Does Good Science Lead to Valuable Knowledge? Biotechnology Firms and the Evolutionary Logic of Citation Patterns. Management Science, 49(4): 366-382.

Griffiths, T. L., & Steyvers, M. 2004. Finding Scientific Topics. PNAS, 101: 5228-5235.

Griliches, Z. 1990. Patent Statistics as Economic Indicators - a Survey. Journal of Economic Literature,

28(4): 1661-1707.

Grun B, Hornik K (2011). Topic models: An R Package for Fitting Topic Models." Journal of Statistical Software, 40(13): 1-30. URL http://www.jstatsoft.org/v40/i13/.

Hall, B. H., Jaffe, A., & Trajtenberg, M. 2005. Market Value and Patent Citations. The RAND Journal of Economics, 36(1): 16-38.

Hall, B. H., Jaffe, A., & Trajtenberg, M. 2001. The NBER Patent Citation Data File: Lessons, Insights, and Methodological Tools. NBER Working paper 8498.

Hall, B. H. 2002. A Note on the Bias of Herfindahl-Type Measures Based on Count Data. In A. Jaffe, M. Trajtenberg, eds. Patents, Citations, and Innovations. MIT Press, Cambridge, MA, 454-459

Hall, D., Jurafsky, D., & Manning, C. D. 2008, Studying the Histories of Ideas Using Topic Models. In Proceedings of *The Conference on Empirical Methods in Natural Language Processing 2008.*

Hargadon A., & Sutton, R. I. 1997. Technology Brokering and Innovation in a Product Development Firm. Administrative Science Quarterly, 42(4): 716-749.

Harhoff, D., Narin, F., Scherer, F. M., & Vopel, K. 1999. Citation Frequency and the Value of Patented Inventions. The 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

Hedge, D., & Sampat, B. 2009. Examiner Citations, Applicant Citations, and the Private Value of Patents. Research Policy, 105(3): 287-289.

Helpman, E. 1998. Diffusion of General Purpose Technologies In E. Helpman (Ed.), General Purpose Technologies and Economic Growth: 85-117. Cambridge, Mass.: MIT Press.

Huff, A. S. 1990. Mapping strategic thought. Chichester, New York: John Wiley and Sons.

Schumpeter, J. 1939. Business Cycles. New York: McGraw-Hill.

Jaffe, A. B. 1986. Technological Opportunity and Spillovers of Research-and-Development - Evidence from Firms Patents, Profits, and Market Value. American Economic Review, 76(5): 984-1001.

Jaffe, A. B., Trajtenberg, M., & Fogarty, M. 2000. Knowledge Spillovers and Patent Citations: Evidence from a Survey of Inventors. American Economic Review Papers and Proceedings, 90(2): 215-218.

Kaplan, S., Murray, F., & Henderson, R. 2003. Discontinuities and Senior Management: Assessing the Role of Recognition in Pharmaceutical Firm Response to Biotechnology. Industrial and Corporate Change, 12(2): 203-233.

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

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

Mei, Q., Shen, X. & Zhai, C. 2007. Automatic Labeling of Multinomial Topic Models. In Proceedings of *The 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

Mimno, D., Wallach, H. M., & McCallum, A. 2008. Gibbs Sampling for Logistic Normal Topic Models with Graph-Based Priors. Paper presented at the Proceedings of the 2008 Workshop on Analyzing Graphs: Theory and Applications.

Mowery, D. C., Oxley, J. E., & Silverman, B. S. 1996. Strategic Alliances and Interfirm Knowledge Transfer. Strategic Management Journal, 17: 77-91.

Phene, A., Fladmoe-Lindquist, K., Marsh, L. 2006. Breakthrough Innovations in the U.S. Biotechnology Industry: The Effects of Technological Space and Geographic Origin. Strategic Management Journal, 27: 369-388.

Podolny, J. M., & Stuart, T. E. 1995. A Role-Based Ecology of Technological Change. The American Journal of Sociology, 100(5): 1224-1260.

Ramage, D., Rosen, E., Chuang, J., Manning, C., & McFarland, D. A. 2009. Topic Modeling for the Social Sciences. NIPS 2009 Workshop on Applications for Topic Models: Text and Beyond

Rosenkopf, L., Nerkar, A. 2001. Beyond Local Search: Boundary-Spanning, Exploration and Impact in the Optical Disc Industry. Strategic Management Journal, 22: 287-306.

Rosen-Zvi, M., Chemudugunta, C., Griffiths, T., Smyth, P., & Steyvers, M. 2010. Learning Author-Topic Models from Text Corpora. Journal of ACM Transactions on Information Systems, 28(1).

Rysman, M., & Simcoe, T. 2008. Patents and the Performance of Voluntary Standard-Setting Organizations. Management Science, 54(11): 1920-1934

Sapir, E. 1944. Grading: A study in semantics. Philosophy of Science, 11: 93-116.

Singh J. & Fleming, L. 2010. Lone inventors as sources of breakthroughs: Myth or reality? Management Science, 56(1): 41-56.

Song, J., Almeida, P., & Wu, G. 2003. Learning-by-Hiring: When Is Mobility More Likely to Facilitate Interfirm Knowledge Transfer? Management Science, 49(4): 351-365.

Tan, D., Roberts, P.W. 2010. Categorical coherence, classification volatility and examiner-added citations. Research Policy. 39(1) 89-102.

Trajtenberg, M. 1990. A Penny for 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.

US Patent and Trademark Office 2005. Handbook of Classification. Washington, DC: United States Government Printing Office.

Wallach, H. M., Murray, I., Ruslan, S., & Mimno, D. 2009. Evaluation Methods for Topic Models. Proceedings of the 26th Annual International Conference on Machine Learning, NY, USA.

Wang, C., Christopher Meek, B. T., & Blei, D. 2009. Markov Topic Models. In Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS).

Whorf, B. L. 1956.Science and linguistics. In J. B. Carroll (Ed.), Language, thought, and reality: Selected writings of Benjamin Lee Whorf: 207-219. Cambridge, MA: MIT Press.

Wry, T., Greenwood, R., Jennings, P. D., & Lounsbury, M. 2010. Institutional Sources of Technological Knowledge: A Community Perspective on Nanotechnology Emergence. Research in the Sociology of Organizations, 29: 149-176.

Zhang, X., & Wang, T. 2010. Topic Tracking with Dynamic Topic Model and Topic-Based Weighting Method. Journal of Software, 5(5): 482-489.

Figure 1: Sample of fullerene and nanotube patents

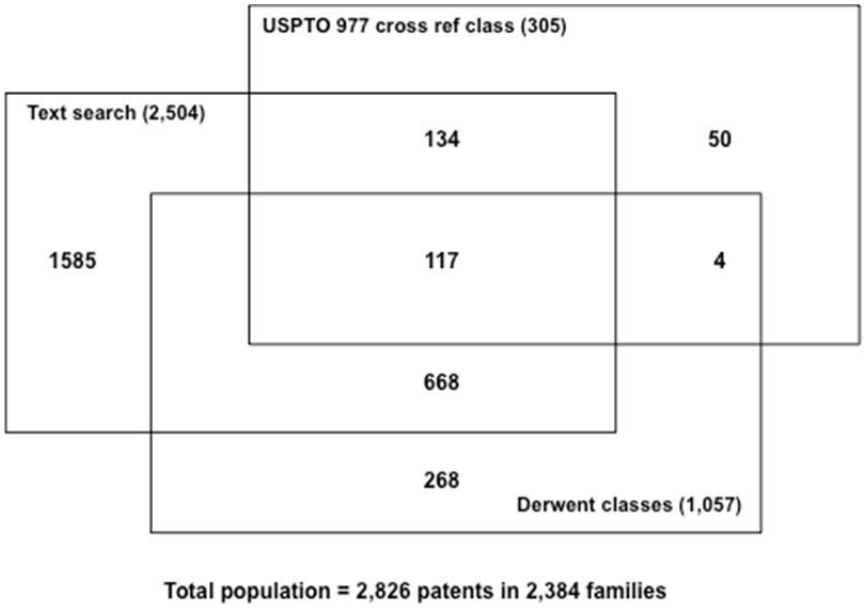

Figure 2: Example coded abstract

Patent Number: 7352617
Title: Nano tube cell and memory device using the same
Inventors: Kang; Hee Bok (Daejeongwangyeok-si, KR)
Assignee: Hynix Semiconductor Inc. (KR)
Application date: February 2005, **Issue date**: April 2008

Abstract
A nano[27] tube[75] cell[92] and a memory[92] device[93] using the same features[35] a cross[92] point[26] cell[92] using a capacitor[60] and a PNPN nano[92] tube[75] switch[24] to reduce[94] the whole memory[92] size[67]. In the memory[92] device[93], the unit[83] nano[92] tube[75] cell[92] comprising a capacitor[60] and a PNPN nano[92] tube[75] switch which does not an additional[76] gate[48] control[24] signal[24] is located[11] where a word[92] line[100] and a bit[92] line[100] are crossed[92], so that a cross[92] point[26] cell[92] array[79] is embodied. As a result[51], the whole chip[28] size[67] is reduced[62], and read[92] and write[92] operations[92] are effectively[6] improved[62].

Vector of top topics:
Topic 92 (application of nanotubes in memory devices): 89%,
Topic 24(application of nanotubes in circuit switch and non-volatile memory): 3%
Topic 62 (Application of nanostructures in energy conversion and energy-related devices): 3%
Other topics: less than 1%

Figure 3: Top cited patents (economic breakthroughs) and topic generating patents (cognitive breakthroughs

	Topic-generating patent	Non-topic generating patent
Top cited patent	21	98
Not top cited patent	168	2,097

Figure 4: Estimated number of first second generation forward citations at each year since grant date for top cited patents (topic-generating and not topic generating)
5 year window for top cited patents

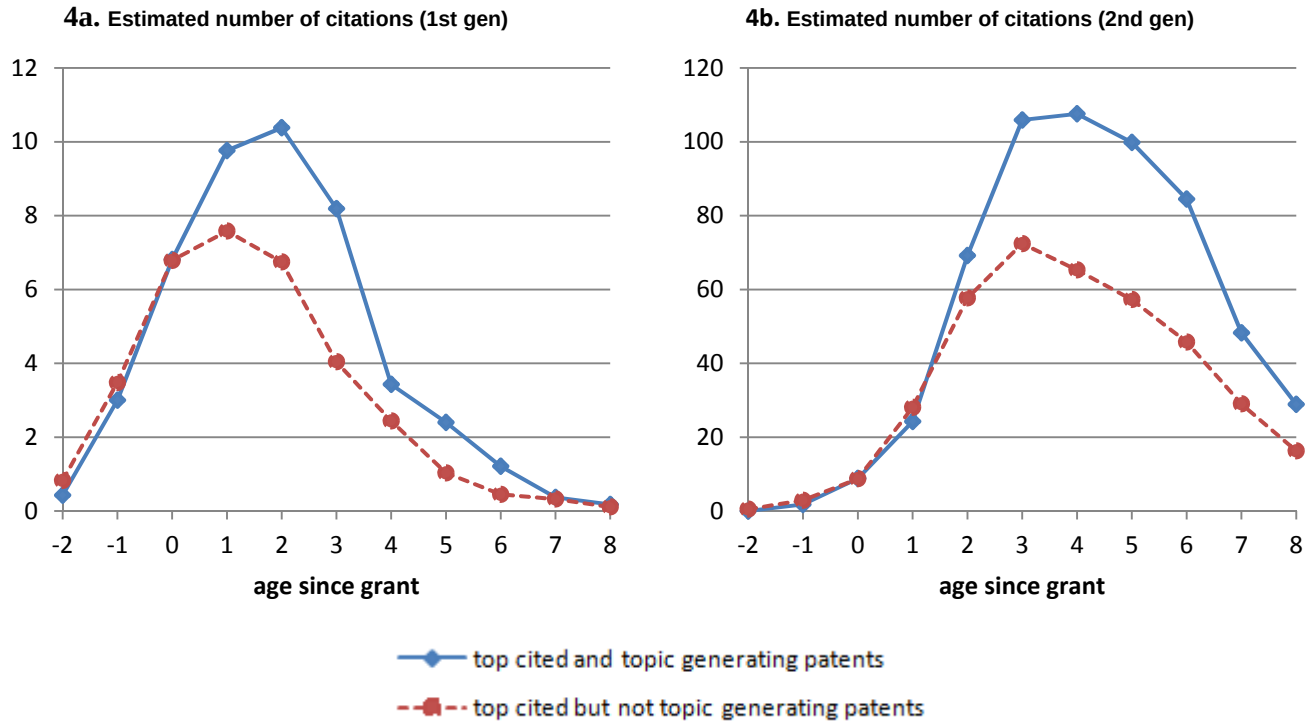

Figure 5: Estimated number of first second generation forward citations at each year since grant date for patents that are <u>not</u> top cited (topic generating and not topic generating)
5 year window for top cited patents

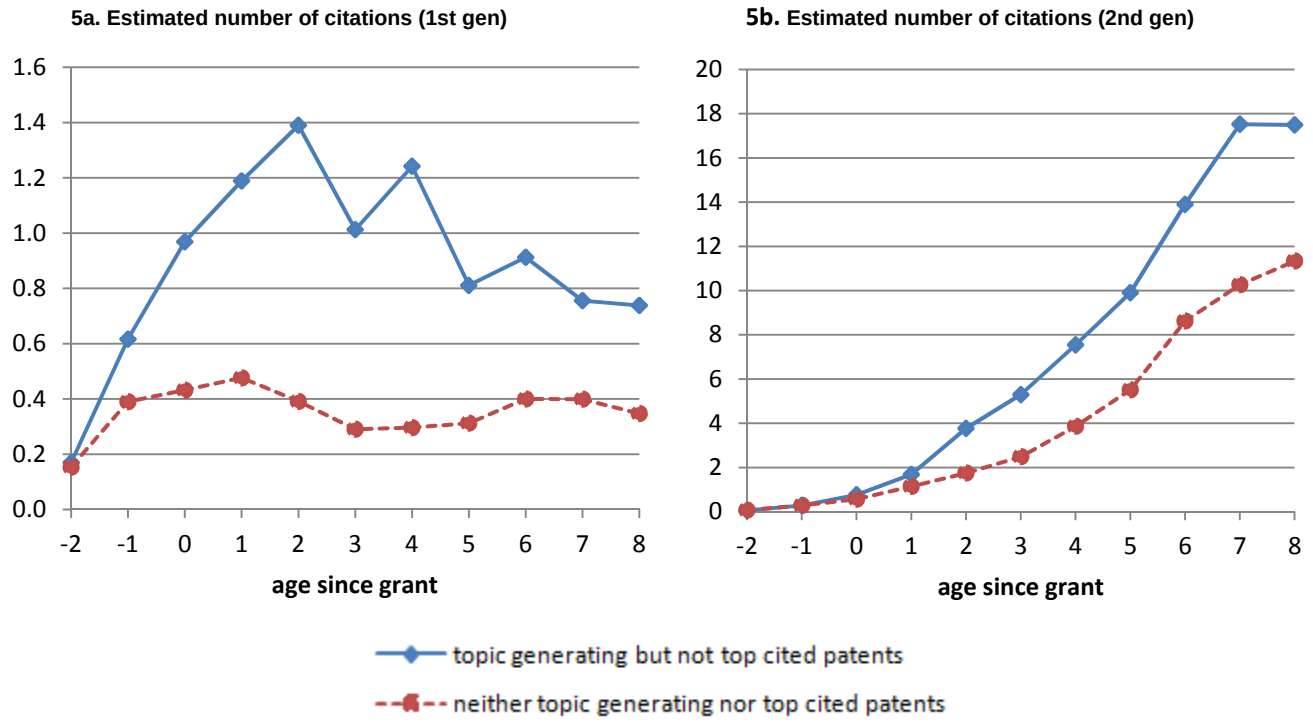

Table 1: Sample topics including the top 20 words associated with each (core fullerene dataset)

Topic 92 *application of nanotubes in memory devices*	**Topic 24** *application of nanotubes in circuit switch and non-volatile memory*	**Topic 62** *Application of nanostructures in energy conversion and energy-related devices*	**Topic 26** *Nanostructures in electro-optical modeling*	**Topic 67** *Nano-motors*	**Topic 56** *application of fullerenes and nanotubes in field emission devices*
memory	control	active	structure	nm	gate
cell	switching	improved	form	diameter	formed
cells	element	efficiency	plural	metallic	insulating
nano	output	conversion	mesh	hollow	cathode
bit	signal	energy	method	size	layer
word	input	reduced	fed	outer	electrode
device	node	achieved	point	average	emission
read	channel	performance	forming	diameters	substrate
stored	switch	addition	applying	pore	insulation
plurality	signals	improve	nano-sized	bearing	hole
array	disposed	density	manufacturing	rotor	field
ferroelectric	circuits	transfer	supporting	supported	display
charge	release	reduction	lateral	mixed	device
photovoltaic	terminal	power	barrier	shaft	cavity
data	elements	increased	decomposing	packed	focusing
solar	non-volatile	increase	defined	lubricant	covering
access	logic	varying	extension	uniform	focus
nonvolatile	relation	life	feedstock	characterized	electron
tube	circuit	accomplished	constituting	grease	triode
capacitor	communication	relates	resultant	rotatable	stack

Table 2: Inventor-level variables associated with breakthrough innovations

Variable	**Description**	**Mechanism**	**Source**	**Expected sign**
Team	A dummy variable that indicates whether patent invented by more than one person.	Inventors working in teams iterate more and make more combinations than do lone inventors due to greater diversity of viewpoints, thus increasing the likelihood the focal patent is a breakthrough	Singh & Fleming (2010)	+
Assigned	A dummy variable that indicates whether patent is assigned to any organization.	Inventors working in organizations are able to use and leverage knowledge accumulated in the organization more the lone inventors, thus increasing the likelihood the focal patent is a breakthrough	Singh & Fleming (2010)	+
Ln(average experience)	Average number of previous patents by the inventors of the focal patent.	Teams of inventors with more collective experience are more likely to have an advantage in inventive skills, thus increasing the likelihood that the focal patent is a breakthrough	Singh & Fleming (2010)	+
Ln(joint experience)	Number of patents produced by the same team of inventors as the focal patent.	The more a team works together, the more likely it will get locked into a particular way of thinking, thus decreasing the likelihood the focal patent is a breakthrough.	Singh & Fleming (2010)	-

Table 3: Invention-level variables associated with breakthrough innovations

Variable	Description	Mechanism	Sources	Expected sign
Backward technological distance	Measures the distance between the technological class of the focal patent and that of its backward citations as a sum over the share of backward citations in the same 3-digit technological class, in the same 2-digit technological class, and in the same 1-digit technological class as the focal patent.	Exploration of distant technological areas as opposed to exploiting the related technological classes should increase the likelihood a focal patent is a breakthrough	Trajtenberg et al. (1997), also Phene et al (2006) and Rosenkopf & Nerkar (2001)	+
# prior art subclasses	Number of subclasses assigned to the focal patent	Patents that combine knowledge from many different technological classes are more likely to be novel and therefore a breakthrough.	Singh & Fleming (2010), Fleming & Sorenson (2001)	+
Patent originality	Measured as 1 minus the Herfindahl index of technological concentration of prior art: 0 if cited patents belong to one 3-digit class, close to 1 if the cited patents belong to many, adjusted for # of prior art citations	The more the focal patent combines ideas from different technological fields, the more original the patent, and the more likely to be a breakthrough.	Trajtenberg et al. (1997) Hall et al. (2001), Hall (2002)	+
No prior art	Dummy = 1 if the patent cites no patents as prior art.	Patents with no prior art are seen as "pioneering" and therefore more likely to be breakthroughs.	Ahuja & Lampert (2001)	+
Ln(component familiarity)	The average time-discounted count of previous usage of focal patent's subclasses across all patents	The more recently and frequently components are used previously, the more likely the inventors will be familiar with them and therefore able to select and recombine them into useful inventions	Fleming (2001)	+
Ln(combination familiarity)	The time-discounted count of previous usage of the focal patent's particular subclass combination across all patents.	The more recently and frequently the combination of components in the focal patent has been used previously, the more likely the inventors will be able to refine it into a more useful invention	Fleming (2001)	+
Ln(cumulative combination)	The total count of previous usage of the focal patent's particular subclass combination across all patents	The greater the number of previous inventions with the same combination of components, the less likely the invention is to be useful due to exhaustion of possibilities for improvement	Fleming (2001)	-
# non-patent references	The total number of non-patent references cited by the focal patents	The greater the "science intensity," the more likely the invention is to be a breakthrough.	Gittelman & Kogut (2003); Harhoff et al. (2003); Deng et al. (1999)	+
Controls:				
# prior art patents	The total number of patents granted by the USPTO cited by the focal patent	Citing more prior art indicates inventions are more likely to garner future citations (more likely to be built on by others).	Fleming & Sorenson (2001), Podolny & Stuart (1995)	+
# claims	Number of claims in the focal patent	The greater the scope of the patent, the more likely the invention is to be a breakthrough.	Singh & Fleming (2010)	+
Family size	Number of patents with identical abstracts and assignees as the focal patent, indicating a cluster of patents around the same invention	The larger the cluster of patents around a single invention, the more likely any one patent in the family will receive future citations.	Related to: Gittleman & Kogut (2003)	+

Table 4: Descriptive statistics

	mean	std	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
(1) Economic breakthroughs	0.052	0.223	1.000															
(2) # forward citations (5-yr)	6.314	11.580	0.799**	1.000														
(3) Topic-generating patent	0.079	0.270	0.085 **	0.145**	1.000													
(4) Team	0.791	0.406	0.057**	0.054*	-0.050*	1.000												
(5) Assigned	0.889	0.314	0.064**	0.097**	-0.000	0.159**	1.000											
(6) Ln(average experience)	2.101	1.171	0.017	-0.002	-0.035+	0.074**	0.048*	1.000										
(7) Ln(joint experience)	0.866	1.235	0.030	0.067**	0.060**	-0.551**	-0.040+	0.398**	1.000									
(8) Technological distance	0.520	0.340	0.021	0.033	-0.035+	-0.029	-0.015	-0.098**	-0.022	1.000								
(9) Patent originality	0.648	0.371	0.024	0.039+	-0.070**	-0.022	-0.002	0.055**	0.039+	-0.033	1.000							
(10) No prior art	0.048	0.213	-0.025	-0.021	0.064**	0.009	0.007	-0.101**	-0.050*	0.311**	-0.391**	1.000						
(11) Ln(component familiarity)	4.661	1.001	0.019	0.008	-0.095**	0.098**	0.078**	0.157**	-0.007	-0.042*	0.057**	-0.101**	1.000					
(12) Ln(combination familiarity)	0.306	0.773	-0.022	-0.010	-0.016	0.027	0.002	0.093**	0.036+	-0.065**	-0.045*	-0.011	0.247**	1.000				
(13) Ln(cumulative combo)	0.408	0.965	-0.027	-0.016	-0.012	0.025	0.005	0.091**	0.033	-0.076**	-0.044*	-0.013	0.237**	0.984**	1.000			
(14) Ln(# non-patent references)	1.498	1.301	0.099**	0.153**	0.032	0.061**	-0.004	-0.020	0.024	0.050*	0.069**	0.040+	-0.090**	-0.107**	-0.113**	1.000		
(15) Ln(# prior art patents)	2.123	1.066	0.041*	0.062**	-0.062**	-0.008	-0.032	0.182**	0.105**	-0.140**	0.503**	-0.447**	0.117**	-0.021	-0.023	0.286**	1.000	
(16) Ln(# claims)	2.835	0.764	0.054*	0.097**	-0.042*	0.007	-0.012	0.009	-0.015	-0.008	0.110**	-0.079**	-0.002	-0.030	-0.034	0.141**	0.198**	1.000
(17) Ln(Family size)	0.126	0.331	0.129**	0.203**	0.066**	0.005	0.022	0.046*	0.086**	0.005	0.009	-0.040+	0.008	-0.039+	-0.039+	0.149**	0.129**	0.034

+<=.1, *<=.05, **<=.01

Table 5: Understanding sources and impact of cognitive breakthroughs (1991-2005)
5-year window since application date for citation counts and economic breakthroughs

	(1)	(2)	(3)	(4)	(5)
	Economic breakthroughs (5-year window after app date)		Citation counts (5-year window after app date)		Cognitive breakthroughs
	Logistic	Logistic	Negative Binomial Incidence-rate ratios	Negative Binomial Incidence-rate ratios	Logistic
	Odds ratios	Odds ratios			Odds ratios
Topic-generating patent		2.048*		1.520**	
		(0.655)		(0.148)	
Team	3.095**	2.899**	1.343**	1.332**	1.118
	(1.235)	(1.143)	(0.125)	(0.123)	(0.313)
Assigned	3.243*	3.146*	1.654**	1.654**	0.910
	(1.831)	(1.786)	(0.185)	(0.187)	(0.260)
Ln(average experience)	1.164	1.152	1.071*	1.069*	1.094
	(0.131)	(0.133)	(0.035)	(0.035)	(0.098)
Ln(joint experience)	1.278*	1.260*	1.131**	1.120**	1.211*
	(0.135)	(0.135)	(0.040)	(0.039)	(0.117)
Technological distance	1.223	1.298	1.164	1.180+	0.495*
	(0.399)	(0.427)	(0.114)	(0.114)	(0.175)
Patent originality	1.132	1.226	1.215+	1.227*	0.539*
	(0.386)	(0.430)	(0.125)	(0.123)	(0.155)
No prior art	0.585	0.601	1.190	1.178	2.025
	(0.439)	(0.445)	(0.221)	(0.217)	(1.013)
Ln(component familiarity)	1.241*	1.223+	1.109**	1.107**	1.076
	(0.133)	(0.132)	(0.039)	(0.039)	(0.106)
Ln(combination familiarity)	3.512	3.463	1.524+	1.573+	0.656
	(3.494)	(3.435)	(0.354)	(0.366)	(0.367)
Ln(cumulative combination)	0.317	0.316	0.729+	0.711*	1.519
	(0.259)	(0.256)	(0.124)	(0.120)	(0.642)
Ln(# non-patent references)	1.350**	1.366**	1.127**	1.122**	0.929
	(0.116)	(0.120)	(0.032)	(0.032)	(0.073)
Ln(# prior art patents)	0.981	0.958	1.062	1.055	1.451**
	(0.116)	(0.116)	(0.047)	(0.047)	(0.172)
Ln(# claims)	1.175	1.176	1.223**	1.227**	0.944
	(0.150)	(0.152)	(0.049)	(0.049)	(0.105)
Ln(Family size)	1.684*	1.702*	1.623**	1.635**	1.097
	(0.384)	(0.390)	(0.160)	(0.160)	(0.251)
Constant	0.000**	0.000**	0.111**	0.113**	0.011**
	(0.000)	(0.000)	(0.028)	(0.029)	(0.008)
Year fixed effects	Yes	Yes	Yes	Yes	Yes
Observations	2,067	2,067	2,276	2,276	2,276
Ll	-356.9	-354.5	-5976	-5967	-451.6
chi2	175.3	178.4	850.3	865.1	328.6
df_m	23	24	28	29	28

Robust standard errors in parentheses, ** p<0.01, * p<0.05, + p<0.1