THE DOUBLE-EDGED SWORD OF RECOMBINATION IN BREAKTHROUGH INNOVATION

Sarah Kaplan*
University of Toronto, Rotman School
105 St. George Street
Toronto, ON, M5S 3E6, Canada
416-978-7403
sarah.kaplan@rotman.utoronto.ca

Keyvan Vakili
London Business School
Sussex Place, Regent's Park
London, NW1 4SA, UK
079 6690 8742
kvakili@london.edu

**Corresponding author. We are grateful to Michael Lounsbury for the suggestion to study fullerenes and nanotubes within the broad domain of nanotechnology, Juan Alcácer for his help in developing the initial data set, and Hanna Wallach for her early test-driving of topic modeling on these data. Grace Gui, Illan Kramer, Sara Sojung Lee, Octavio Martinez, Shamsi Mohtashim, Neal Parikh, Navid Soheilnia, Aaron Sutton, and Xihua Wang provided much-appreciated research assistance. Earlier drafts of this manuscript benefited from the comments of Ajay Agrawal, Christian Catalini, Kristina Dahlin, Andrea Fosfuri, Alfonso Gambardella, Constance Helfat, Nico Lacetera, Anita McGahan, Andrea Mina, Jasjit Singh, Kevyn Yong, 2 anonymous reviewers and participants in the DRUID conference 2012 (where it received the Best Paper Prize), EGOS Colloquium 2011, the West Coast Research Symposium 2011, the slump_management research group, and in seminars at Rotman, the Yale School of Management, Chicago Booth, University of Illinois-Urbana Champaign and Wharton. This work is partially supported by the Mack Institute for Innovation Management at the Wharton School, University of Pennsylvania (where Kaplan is a Senior Fellow) and the Canadian Social Sciences and Humanities Research Council under grant #410-2010-0219. All errors and omissions remain our own.*

THE DOUBLE-EDGED SWORD OF RECOMBINATION IN BREAKTHROUGH INNOVATION

Abstract:

We explore the double-edged sword of recombination in generating breakthrough innovation: recombination of distant or diverse knowledge is needed because knowledge in a narrow domain might trigger myopia; but, recombination can be counterproductive when local search is needed to identify anomalies. We take into account how creativity shapes both the cognitive novelty of the idea and the subsequent realization of economic value. We develop a text-based measure of novel ideas in patents using topic modeling to identify those patents that originate new topics in a body of knowledge. We find that, counter to theories of recombination, patents that originate new topics are more likely to be associated with local search, while economic value is the product of broader recombinations as well as novelty.

Keywords: breakthrough innovation; recombination; patents; creativity; topic modeling; text analysis; nanotechnology; cognition

Research on innovation has long sought to determine the sources of innovative breakthroughs because they are the basis of change in scientific and technological ideas and of potential increases in social (Kuhn 1962/1996), firm (Hall *et* al 2005, Phene *et* al 2006) and individual economic value (Ahuja *et al* 2005). Most of this research has drawn on theories of recombination based in a 'tension' view of the relationship between knowledge and creativity (Weisberg, 1999): deep knowledge in one domain dampens creativity by entrenching researchers into one way of thinking. Breakthrough innovation therefore requires a broad search for information and the recombination of different kinds of knowledge to break those bonds and produce novel ideas that achieve high economic value (Ahuja & Lampert 2001; Guilford 1967). Bridging distant or diverse knowledge should therefore enhance creativity (e.g., Hargadon & Sutton 1997; Audia & Goncalo 2007).

A contrasting, and less tested, theory of creativity – the 'foundational' view – differs from the tension view in arguing that local search to identify anomalies is most likely to produce breakthrough innovations (Weisberg 1999; Taylor & Greve 2006). That is, in order to break out of existing constraints and advance a field beyond its current state, one must have a deep understanding of the foundations of a particular knowledge domain, its assumptions and its potential weaknesses. Wide recombination may be detrimental to innovation because only a deep dive can produce breakthroughs. In theory, all innovations are based on some sort of recombination. Here, we follow the lead of other innovation scholars in referring to long-jump search as recombination involving distant or diverse knowledge, where narrow recombination of similar knowledge can be seen as local search (Gavetti & Levinthal 2000; Fleming 2001; Ethiraj & Levinthal 2004).

Rather than seeing the tension and foundational views of recombination as alternatives, we might usefully conceptualize them jointly as a "double-edged sword" (Sternberg & O'Hara 1999: 256): recombination of distant or diverse knowledge is required to create breakthroughs because knowledge in a narrow domain might trigger intellectual lock-in and lead to incremental innovation; on the other hand, such recombination might be counterproductive because local search (narrow recombination) in a domain may be required in order to find the openings for breakthrough innovations. Yet, to date, the field of management has not examined how the two blades of the sword are interrelated. In our study, we develop a new method for examining these relationships in the context of patented innovations.

We can reconcile the tension and foundational views of recombination if we take into account how creative processes shape both the novelty of the idea (the cognitive dimension of breakthroughs) as well as the realization of value (the economic dimension of breakthroughs). Innovations potentially differ along these two dimensions (Amabile 1983; Audia & Goncalo 2007; Amabile & Pillemer 2012). An innovation might be novel in the sense that it introduces a potential new technological trajectory or incremental because it follows on an existing trajectory. Subsequently, an innovation may or may not turn out to be valuable in terms of generating economic returns for the owners of the invention. Art must appeal to collectors or museums, books must be sold to publishers and readers, new startups must attract venture capital funding, patents must garner licensing or sales revenues, etc. Inventors will first generate insights with different levels of novelty. Subsequently, if they understand which ideas will have the most economic value and 'sell' the idea to the appropriate audience, the novel ideas will be recognized as valuable (Sternberg & O'Hara 1999). By corollary, if they do not recognize which ideas are most valuable or are unable to sell the ideas, the economic value of the invention will not be realized.

Importantly, while innovation studies have suggested a positive link between novelty and the value (Trajtenberg *et al* 1997; Singh & Fleming 2010; Phene *et al* 1997), creativity scholars have suggested that the skills for generating cognitively novel ideas and those for selecting and promoting ideas that have economic value are weakly related at best (Sternberg & O'Hara 1999; Sternberg 1997). Thus, a comparison of the tension (wide recombination) and foundational (local search) processes of creativity would benefit from considering their separate impacts on novelty and value.

In studies of innovation, particularly of innovative patents, research has repeatedly found strong evidence for various forms of recombination as the main mechanism producing breakthroughs (e.g., Trajetenberg *et al* 1997; Fleming 2001; Hall *et al* 2001, Rosenkopf & Nerkar 2001; Hall 2002). Yet, the measure of breakthroughs to which they have been constrained is one of citation counts to patents, which have been shown to correlate (if noisily, Bessen 2008) with measures of economic value

(Griliches 1990), such as inventors' or other experts' estimates of future financial value (Albert *et al* 1991, Harhoff *et al* 1999), patent renewal fee payments (Harhoff *et al* 1999; Hegde & Sampat 2009), filing patents for the same invention in multiple jurisdictions (Lanjouw & Schankerman 2004), and firms' stock market values (Deng *et al* 1999; Hall *et al* 2005). As a result, citations may most appropriately measure the value or usefulness of patents but do not capture their novelty.

To develop a separate measure of cognitive novelty, we draw on Kuhn's (1962/1996) argument that scientific ideas are embedded in vocabularies and therefore shifts in ideas can be detected in shifts in language. If we are interested in understanding the emergence of breakthrough novel ideas, then, we need methods that pay attention to the language representing the innovations. In this paper, we introduce just such an approach. Borrowing a computer science technique called 'topic modeling' that discovers the latent topics in a collection of documents and identifies which composition of these topics best accounts for each document, we map the formation of new topics in patent data – which can be seen as the emergence of novel ideas – and locate the patents that introduce them. Those patents that originate new topics can be thought of as cognitive breakthroughs. In introducing this new measure of cognitive breakthroughs, we can unpack the relationship between contrasting creative processes – tension vs. foundational – and differing creative outcomes – cognitive novelty vs. economic value.

To develop and validate this approach, we examine the formation of novel ideas in a domain of nanotechnology, that of Buckminsterfullerenes (and the related area of carbon nanotubes). This is a useful setting because fullerenes can be seen as a 'general purpose technology' (Bresnahan & Trajtenberg, 1995; Helpman, 1998) with potential applications in many areas (Rothaermel & Thursby, 2007). Thus, a wide variety of cognitive and economic breakthroughs should be possible to identify.

In this field, we find that – consistent with tension theories of creativity – distant and diverse recombinations are positively associated with economic value as measured by patent citation rates. However, in support of foundational theories, we find that cognitive novelty (a patent that originates a new topic) is associated with local search. At the same time, we show that novel ideas tend to have higher economic value. This suggests an alternative model of innovation, where novelty is one source of economic value produced by innovators who ‘draw on a single domain in a practiced manner’ (Taylor & Greve 2006, p. 727), while the recombination of distant or diverse knowledge positively influences economic value but reduces the likelihood of cognitively novel breakthroughs. Few patents in our study were both cognitive and economic breakthroughs (less than one percent of our sample), but they appear to have a greater impact on future innovation than any other kind of invention.

A text-based approach to the analysis of patents gives the researcher new traction in understanding breakthroughs and the emergence and evolution of technologies. First, we use texts of patents to develop a measure of cognitive novelty and find that novelty contributes to the creation of economic value. At the same time, we highlight the conflicting creative processes that lead directly to novelty and value, the former requiring local search and the latter distant and diverse recombinations. This contrasts with the common view that local search should be associated with exploitation and not exploration. It also draws attention to the organizational design implications for managing the double-edged sword of recombination in innovation, where innovation strategies must deal with the trade-offs and interrelationships between allocation of resources towards the development of deep knowledge in particular domains and the creation of opportunities for long jump search and recombination.

HYPOTHESES: EXPLORING THE DOUBLE-EDGED SWORD

The creativity literature proposes that the creative process involves the generation of novelty and then the subsequent achievement of economic value through the recognition and promotion of those novel ideas that have the most promise (Sternberg & O’Hara 1999). This can be seen as a process of variation (either “blind” or intentional production of novelty) followed by selection and retention (realization of economic value) (Campbell 1960; Simonton 1999). Creativity research has proposed two alternative models – the tension and foundational views – of the role of knowledge in these creative processes. The tension view asserts that deep knowledge can lead to myopia such that recombination of distant or diverse knowledge is needed in order to see new ideas. The foundational view suggests that the only way to see potential anomalies that could lead to breakthroughs is through search in a narrower domain, i.e. local search. These are the two blades of the double-edged sword: wide recombination is either seen as promoting or detracting from innovation.

Figure 1 portrays the two blades of the sword as they relate to each innovative outcome – cognitive novelty and economic value. In the tension view, wide recombination should be positively associated with novelty (Path A) and economic value (Path C). In the foundational view, local search (narrow recombination) is more likely to produce novel (Path A') and economic value (Path C'). In both cases, novelty, once achieved, should also be associated with economic value (Path B).

--Insert Figure 1 about here --

Note that most innovation studies have an implicit model of innovation based in the tension theory of creativity: they argue that recombination generates novel ideas which, in turn, are more likely to be valuable (in these studies, the focus is patented inventions, so economic value is measured as citations as prior art by subsequent patents) (Fleming 2001; Hall *et al* 2001, Trajtenberg *et al* 1997; Gittelman & Kogut 2003; Singh & Fleming 2010). The dynamics are represented in Figure 1 in Paths A and B. To date, however, these studies have mainly looked at the effect of recombination in patents (using a variety of measures) on subsequent citations to those patents. In other words, rather than testing paths A and B separately, they have tested Path C in Figure 1. They find that recombination is positively associated with economic value but do not analyze directly the intervening step associated with the generation of novelty from those recombination processes.

In using topic modeling to analyze breakthrough patents, we can explore this relationship directly by measuring the presence of novel ideas (as indicated by shifts in vocabularies in the patent texts) and determining if this variable mediates the association between wide recombination and value (as indicated by forward citations) or if local search (narrower recombination) is more likely to lead to cognitive and economic breakthroughs. This approach will allow us to understand if there are any contradictions between the processes leading to novelty and those engendering economic value. We will accomplish this through a test of mediation (Baron & Kenny 1986; Iacobucci *et al* 2007; Zhao *et al* 2010) so that we can examine each of the paths and their joint effects. In testing paths C and C', we replicate prior innovation studies showing the association between wide recombination and economic value. In testing paths A (and A') and B, we explore the relationship between tension and foundational views in producing novel ideas that should subsequently be associated with economic value.

Foundational vs. tension theories and economic value

Tension assumptions about recombination (Hargadon & Sutton 1997; Weisberg 1999) have dominated management scholarship on innovation. The view is that deep knowledge in a single or small number of domains may lock inventors into one way of thinking and therefore block their ability to generate breakthroughs. Local search enables only narrow recombinations that produce incremental innovations. Therefore, to generate breakthroughs, inventors must combine knowledge from distant and diverse sources. These studies use forward citation counts to measure breakthroughs, which capture the economic value of the patent. The relationship they test is represented by path C in Figure 1.

Specifically, research has suggested that recombination is most likely to lead to higher citation rates if the knowledge combined is technologically distant. The idea is that combining knowledge from 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 and ultimately become highly cited (Phene *et al* 2006, Rosenkopf & Nerkar 2001, Trajtenberg *et al* 1997). Similarly, scholars have argued that highly cited patents are more likely to be combinations of not just distant but also diverse knowledge domains (Hall *et al* 2001), where greater diversity (lower concentration) of knowledge avoids intellectual lock in. According to these theories, the positive impact of long-jump search on value occurs through the (untested) mechanism of novelty. It is noteworthy that these scholars have not claimed that novelty is the sole driver of economic value nor that recombination only serves to generate novel ideas and has no direct effects on the degree of economic value created. For example, combining diverse knowledge domains might enlarge audiences for the innovation, increase the likelihood it will be found by inventors or patent examiners in a search for prior art or broaden the network in which innovations would diffuse. Taking these insights together, we hypothesize that:

> *H1a: A patent based on distant and diverse recombinations is more likely to receive a higher number of citations than a patent produced based on local search (narrow recombination) (path C in Figure 1).*

While the existing empirical evidence for breakthrough patents has overwhelmingly supported the tension view of recombination, the foundational view would propose the reverse relationship between recombination and measures of economic value. Under this logic, inventors should need to explore a relatively narrow domain in-depth in order to know how to “defy the crowd” and “buy low and sell high” (Sternberg & O’Hara 1999; Sternberg & Lubart 1995). Inventors cannot see new sources of value without understanding what assumptions are behind the existing sources of value, and these insights can only come from focused, local search. This view of creativity is consistent with ecological theories that domain-spanning activities may suffer market penalties due to both deficiencies in production of innovations as well as problems of market reception. Recombinations of distant or diverse knowledge might get in the way of identifying value because they would disperse effort and distract from obtaining the incisive insight that comes from a deep appreciation of one domain (Hannan, Polos & Carroll, 2007). Thus, wide recombination could prevent the realization of economic value because it produces superficial or incremental work. One might also infer that wide recombination could compromise the realization of economic value because market audiences penalize offerings that span categories (Hsu, Kocak & Hannan 2009; Rao, Monin & Durand 2005; Ruef & Patterson 2009; Zuckerman 1999). That is, wide recombination of inputs could potentially lead to difficulties in classifying the innovation and thus to penalties in the form of fewer citations over time. We therefore offer a competing hypothesis (as represented in Path C’) to the recombination model of creativity in generating economic value:

> *H1b: A patent based on local search (narrower recombination) is more likely to receive a higher number of citations than a patent produced based on distant and diverse recombinations (path C’ in Figure 1).*

Foundational vs. tension theories and novelty

By calling out novelty as a separate creative output from the generation of economic value, we are able to interrogate existing research that has privileged recombination processes as the source of innovative breakthroughs. Implicit in the arguments made in studies of breakthroughs is the idea that wide recombination generates novel ideas (path A in Figure 1), which in turn are more likely to be cited as prior art by subsequent patents (path B).

With regard to the connection between novelty and economic value, while the creativity literature makes it clear that not every truly novel idea will become valuable (Amabile 1983; Sternberg & O'Hara 1999; Sternberg 1997), they also indicate that novelty will increase the probability that economic value can be obtained, all else equal. This logic is consistent with the arguments made in the innovation literature on patents discussed above. Of course, it is a requirement of the US Patent Office that every patent be novel to some extent, though some inventions may be 'improvements,' built upon existing technological trajectories, while others may be truly novel, introducing new technological trajectories. Our concern here is with those that meet this latter standard. Thus, we hypothesize:

> *H2: Patents that represent cognitive breakthroughs (truly novel ideas) are more likely to receive a higher number of citations than patents that do not (path B in Figure 1).*

Novelty has been portrayed in the innovation literature as an (unmeasured) output of recombination and an input to the creation of economic value (citations). For example, Trajtenberg *et al* (1997: 29) claim that 'synthesis of divergent ideas is characteristic of research that is highly original.' Similarly, Phene *et al* (2006: 370) suggest that, 'Knowledge that is technologically… distant provides the organization with an opportunity to make novel linkages.' These arguments are based in the tension view of recombination, which assumes that knowledge and creativity are opposing forces, such that 'knowledge may provide the basic elements…out of which are constructed new ideas, but in order for these building blocks to be available, the mortar holding the old ideas together must not be too strong' and too much knowledge of a domain can be habit-forming and inertial (Weisberg 1999, p. 226). Recombination of distant or diverse knowledge can break these habits. In introducing a measure of novel ideas, we can make the implicit model in innovation studies explicit: novel ideas – what we conceptualize as cognitive breakthroughs – are the products of distant and diverse recombinations:

> *H3a: Patents produced through distant and diverse recombinations are more likely to be cognitive breakthroughs (truly novel ideas) than those that are not (path A in Figure 1).*

The 'foundational' view makes the opposite claim (Weisberg 1999). Here, immersion in a particular domain is required in order to produce novelty (Csikszentmihalyi 1996). Local search and narrower recombinations based on deep knowledge in one area enables the identification of anomalies that lead to new insights by exposing the tensions or challenges in the current ways of thinking. This is consistent with the Kuhnian (1962/1996) model in which paradigm shifts are triggered by the accumulation of anomalies. Narrow but deep search leads to truly novel breakthroughs in knowledge because it enables researchers to identify 'what rules to break' (Taylor & Greve 2006, p. 726). These findings are also consistent with work at the inventor level of analysis suggesting that specialization is important to push the frontier of knowledge outward as the 'burden of knowledge' increases over time (Jones 2009, Agrawal *el al*, 2013, Conti *et al* forthcoming). We therefore offer a competing hypothesis (as represented in Path A') to the tension model of creativity:

> *H3b: Patents produced through local search (narrower recombination) are more likely to be cognitive breakthroughs (truly novel ideas) than those that are not (path A' in Figure 1).*

If developing truly novel inventions were the only mechanism through which recombination would lead to higher economic value, we should expect full mediation of Path C when introducing Paths A and B. However, there are reasons to expect that this may not be the case (as outlined in the argument for H1a). But, without a measure of novel ideas, prior scholars have not been able to tease apart the effect of novelty from other effects of recombination.

TOPIC MODELING OF PATENT TEXTS: A MEASURE OF NOVEL IDEAS

Crucial to our analysis is the introduction of topic modeling as a way to create a new measure of novelty to contrast with existing citation-based measures of economic value in patents. The intuition behind topic modeling as a method to identify novel ideas is the following: the algorithm uses the co-location of words in a collection of documents to infer the underlying (or latent) topics in those texts and the weight of each topic in each individual document. We can then identify the documents that are the originators of each topic by finding the earliest documents with a significant weight in the topic. These originating documents can be seen as cognitive breakthroughs. In our case, because we study patents, we call these topic-originating patents. Because topic modeling is a new method in strategic management, we introduce it first here before delving, in the next section, into the empirical methods and measures of other variables, which are more standard in the field. We explain how our method works and then show how we have implemented it in our sample of fullerene patents. For reasons of space, we have left certain technical details (useful to anyone looking to implement topic modeling in their own work) to an online appendix that is available as a companion to this article.

Our methodological move is to treat the texts of patents as representations of the inventive ideas embodied in them. Bibliometric techniques to understand the evolution of science and technology have a long tradition starting from the pioneering work of de Solla Price (1965a; 1965b). However, most of the work to date has used citation analyses (e.g., Leydesdorff *et al* 1994; Meyer *et al* 2004; Dahlin & Behrens 2005). Text analysis has been much less frequent, and, until recently, the main uses of the texts were counts, factor analyses and co-word analyses of keywords (typically in the titles of papers or patents) (Yoon & Park 2005; Azoulay *et al* 2007; Mogoutov & Kahane 2007, Upham *et al* 2010). With the increasing power of computation and availability of texts in electronic form, scholars are exploring the possibilities of more complete uses of texts, which would therefore require unsupervised approaches.

The study reported here follows these recent trends. It is premised on the idea that studying language in documents should provide a reading of their cognitive content (Duriau *et al* 2007, Whorf, 1956). In management studies, this idea has been adapted methodologically to use word counts to represent themes (Huff, 1990; Abrahamson & Hambrick, 1997; Kaplan *et al* 2003; Kaplan 2008). Where the concern is in identifying themes over large numbers of texts, topic modeling – a text analysis technique developed in computer science – offers exciting potential (see Blei 2012 for an overview). The advantage of topic modeling over word counts and keyword analyses is that it allows for polysemy – words can take on different meanings depending on their contexts – and it is inductive – the scholar does not have to specify categories *a priori* but can allow them to emerge from the data. This is particularly important given Kuhn's (1962/1996, p. 205) argument that 'proponents of different theories are like the members of different language-culture communities,' where vocabularies might share many of the same words but the actors attach different meanings to them.

Thus, we believe topic modeling should be a fruitful approach to measuring interpretations in the emergence of a new technological field (Hall, D. *et al* 2008). For our purposes, we use the texts in the abstracts of patents to understand how different actors describe what the technology is and could be, and then to identify shifts in language that represent the emergence of novel ideas.

A primer on topic modeling

The topic modeling approach we use is based in the Bayesian statistical technique of Latent Dirichlet Allocation (LDA)[1] (Blei *et al* 2003). Topic modeling allows the researcher to uncover automatically themes that are latent in a collection of documents and to identify which composition of themes best accounts for each document. The documents and the words in the documents are observed but the topics, the distribution of topics per document and distribution of words over topics are unobserved and represent a 'hidden structure' (Blei 2012). Topic modeling uses the co-occurrence of observed words in different documents to infer this structure. According to Blei (2012, p. 79), 'This can be thought of as 'reversing' the generative process – what is the hidden structure that likely generated the observed collection?' Computationally, the algorithm identifies the posterior distribution of the unobserved variables in a collection of documents.

This idea is represented schematically in Figure 2. The shaded circle denotes what can be observed (**w**, the words in the documents in the collection). The unshaded circles denote latent (unobservable) variables: **z**, the topic assignment in each document; **θ**, the per-document topic proportions; and **α** and **β**, the parameters of the Dirichlet priors for θ and the distribution of topics over words respectively. Each box is a plate where the N plate denotes the words within documents, the D plate denotes the documents within the collection and the T plate denotes the distribution of words over topics. Each word is assumed to be drawn from one of T topics. All topics are used in every document,

[1] 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). Blei *et al* (2003) provide more details on LDA and its comparison with other methods such as latent semantic 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). See http://nlp.stanford.edu/software/tmt/tmt-0.4/ for further details on the toolbox. Another good option is MALLET (http://mallet.cs.umass.edu/). Many new implementations are emerging as topic modeling becomes more prevalent, e.g., one can use the R package recently developed by Grun and Hornik (2011).

but exhibit them in different proportions (usually where only a few topics are important). The arrows indicate conditional dependencies between the variables such that assigning topics to words (z) depends on the per-document topic proportions (θ), and the appearance of a word in a document is inferred to depend on the distribution of topics over words (β) and the topics in each document (z). (See the online appendix for details on the formulas.)

-- Insert Figure 2 about here --

Given a collection of documents, the topic model algorithm provides two outputs, the first being a list of topics with a vector of words weighted by their importance to the topic, and the second being a list of documents (in our case, patent abstracts) with a vector of topics weighted by their importance to the document. This method allows the researcher to quantify meaning over large numbers of texts and to identify shifts in thought. The polysemy of a word is measured by the co-occurrence with other words in a document, where the number of topics in a word corresponds with the number of different meanings it has (Chang *et al* 2009).

A topic is a multinomial over a set of words, and therefore is not labeled by the algorithm (Blei & Lafferty 2007). Automatic labeling of topics is not reliable (Mei *et al* 2007). Thus, a further step in using topic models is the labeling of topics based on the words in them, which serves an important function in validating the topics produced by the model as well as generating a label to characterize each topic. For computer science applications such as the development of predictive models for text searches, the best-fit model often produces a very large number of topics. However, Chang *et al* (2009) show that these best-fit models do not produce topics that represent distinct meanings and that smaller numbers of topics make interpretation more feasible. Thus, scholars have found it most useful to constrain the number of topics (where the typical number selected is 100) (Blei & Lafferty 2007; Hall, D. *et al* 2008). Following their lead, we limited the model to 100 topics. This provided both statistically and semantically meaningful topics.

Sample of fullerene and related patents

To test the use of topic models to identify shifts in vocabularies, we focused on a single technical domain, that of buckminsterfullerenes (and the chemically related carbon nanotubes). This narrow focus is essential because it allowed us to identify domain experts to validate the topics generated using the topic-modeling algorithm (Grimmer & Stewart 2013). Prior studies of the emerging field of nanotechnology have found fullerenes and nanotubes to be a useful site for analysis (Kuusi & Meyer, 2007; Wry, *et al* 2010) because they can be applied in a broad range of applications from medicine to electronics to sports and therefore can be conceptualized as general purpose technologies (GPTs) (Bresnahan & Trajtenberg, 1995; Helpman, 1998).[2] They have the chemical formula of C_{60} or Carbon 60. Buckminsterfullerenes (also known as fullerenes) 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.

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 associated with a multiplicity of interpretations. These patents show that inventors envision technologies for revolutionary new applications (e.g., implantable medical devices to control insulin levels for diabetics, more targeted treatments for cancer, structural materials for combat and sports gear, 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 universities and firms 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 depending on the interpretations and choices these researchers make.

We collected the 2,826 fullerene and nanotube patents granted by the US Patent and Trademark

[2] Fullerenes compare quite well in the generality index with other GPT's studied (Hall *et al.* 2001). For example, from 1990 to 2000, the average generality index of fullerene patents is above 0.6 which is nearly 50 percent higher than that reported for patents in computers and communications (the technologies Hall *et al.* (2001) identify as having the highest generality).

Office (USPTO) through 2008. We identified the population of patents using three separate search techniques. First, we used Derwent's technology classifications to select all patents they identify as pertaining to either of these technologies: B05-U; C05- U; E05-U; E31-U02; L02-H04B; U21-C01T; X12-D02C2D; X12-D07E2A; X12-E03D; X16-E06A1A. Second, the USPTO established a nanotechnology 'cross reference' class (#977) in 2004, which was applied retroactively to all previously-granted patents deemed relevant as well as to new nanotechnology patents. We selected all the patents in subclasses pertaining to fullerenes and nanotubes (977/735-752). To complement the use of these formal classification systems, we also selected all utility patents with the terms 'fullerene' or 'carbon nanotube' in the title, abstract or claims. Figure 2 demonstrates that no individual sampling technique provided a complete picture of patents that could plausibly be associated with fullerenes, and we believe our approach to developing the population of patents in this field compensates for biases created by any one method of classification.

-- Insert Figure 2 about here --

Deriving fullerene and nanotube topics

For each patent, we used the abstracts from its USPTO document. Patents' abstracts are appropriate for an analysis of shifts in language because they are meant to represent a summary of the novel aspects of the invention. Specifically, the USPTO instructs applicants that, 'The purpose of the abstract is to enable the [USPTO] and the public generally to determine quickly from a cursory inspection the nature and gist of the technical disclosure,' where, 'the form and legal phraseology often used in patent claims… should be avoided'[3] (see also Emma 2006). Further, the USPTO requires that abstracts be 150 words or less, thus assuring that the documents compared in our analysis are of approximately equal size.

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

[3] First quote from, http://www.uspto.gov/web/offices/pac/mpep/documents/0600_608_01_b.htm (accessed November 15, 2012). Second quote from the Manual of Patent Examining Procedure (MPEP), Eighth Edition, August 2001, Latest Revision July 2010, Chapter 600 Parts, Form, and Content of Application, 608.01(b) 'Guidelines for the Preparation of Patent Abstracts' Section C: 'Language and Format'

assignees into patent families. This resulted in 2,384 patent families based on the 2,826 patents (there are 336 families with more than one patent, most of which include only 2 patents, with an average of 2.56 and a maximum of 15). We use the data associated with the chronologically first patent in the family. As is typical practice, we removed 'stop words' such as 'the,' 'and,' 'that,' or 'were' that do not contribute to the identification of topics. We identified 100 separate topics, the probability that each of the words appeared in each topic, and the weight of each topic in each abstract.

To label the topics and validate their usefulness in identifying separate ideas, three nanotechnology experts[4] separately reviewed each of the 100 topics. Based on the list of the top 20 words and their weights as produced by the topic-modeling algorithm, we asked each coder to provide a short name to label the topic. We obtained a Krippendorff's α, a common measure of inter-rater reliability, of 0.78. Disagreements for 22 topics were all resolved in joint discussions. A series of topics focused on production processes such as chemical functionalization of nanotubes, metal catalysts for production, or using a reaction vessel for producing nanotubes. Other topics covered applications into such areas as neural networks, reinforced golf balls, optical devices, batteries, transistors, magnetic memory, recording devices, temperature sensing devices, x-ray devices, DNA detectors or plasma display panels. A third category included topics related to the equipment – primarily scanning probe microscopes – used for visualizing and manipulating nanoscale matter.

The topics as generated from the abstracts of patents do not give us the same information as that captured by patent office classifications. The correlation between categories developed using topic modeling and the USPTO classes (using primary topics and primary 3-digit patent class) is 0.22 with a standard deviation of 0.10 (where the average correlation is the average over all the calculated maximum correlation values for each topic with all the patent classes). This is perhaps not surprising: topics are generated from the writings of inventors (and others who help construct the patent) to describe the nature of the invention, while classifications are assigned by patent examiners using previously

[4] Two post docs and one graduating PhD student, each with experience in research on fullerenes and nanotubes.

established classification systems to facilitate their search for prior art (USPTO, 2005).

Identifying topic-originating patents that represent truly novel ideas

There are many potential analytical uses of the data produced using topic modeling. For the purposes of this study, we focus on the identification of patents that originate novel ideas. To do so, we detect 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 12 months of the topic formation (based on application date). The average number of topic-originating patents using this method is 1.89 per topic for a total of 189. The median is 1. Two topics had more than 10 patents associated with them in the first year; results are robust to their omission. Thus, *topic originating patents* is an indicator variable where a 1 identifies those patents that are over the threshold (though results are robust to the use of a continuous variable where topic originating patents are measured according to the weight of the topic).

The selection of topic-originating patents is sensitive to the cutoff points we set. For the time frame, we chose 12 months as reasonable estimate of the time for which the knowledge of that invention would not be widespread (where 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. To identify the appropriate threshold, for each of the 100 topics, we provided our three expert coders with a chronological list of patents (and their abstracts) that had a greater than ten percent weight in the topic. They were asked to identify the first patent chronologically that represented the essence of the topic. The weight that best matched the expert assessments was a 0.20 threshold. Nevertheless, to check the robustness of this threshold, we conducted our statistical analyses with topic-originating patents as identified using thresholds of 0.15 and 0.25. These results are qualitatively similar to those using the best-fit threshold, with the effect of topic originating patents in our regressions slightly lower (but still significant) for patents using the tighter 0.25 threshold and slightly higher (but also still significant) for patents using the less-strict 0.15 threshold.

One might be worried that the identification of topic-originating patents is merely mechanical.

Since topic modeling calculates posterior probabilities, the patents we identify as topic-originating could be supposed to exist only because of the many citations to those patents that follow or because certain topic originating patents are associated with more heavily populated topics. We do not believe this to be the case for several reasons. First, the number of patents per topic differs from topic to topic, ranging from 9 to 44 (mean 23.2, standard deviation 8.7). Second, in examining the 5-year forward citations for each of the topic-originating patents, we see that there is substantial variation around the mean of 22.17, where the minimum is 0 and the maximum is 187. Indeed, only 20 of the 189 topic-originating patents are also in the top 5 percent of cited patents (those that are often considered 'breakthroughs').

Moreover, the correlation between the topic size and the number of citations to its originating patents is negative and insignificant. In statistical analyses, there is no direct positive relationship between the size of the topic and the number of citations that topic's originating patents receive from follow-on patents. Many of the forward citations are made by non-fullerene patents (and therefore not in our dataset and not used in the text analysis that identified the topics), and for those that are fullerene patents, the citation pattern does not indicate any skewed dependence on topics. Finally, because prior art is assigned based on patent classifications, the low correlation between such classes and the topics would also mitigate any argument of reverse causality. The advantage of topic modeling is precisely that it allows us to identify both heavily and sparsely populated topics. This technique enables the researcher to identify shifts in vocabulary, whether or not many other patents then continue the conversation.

MEASURING AND TESTING THE DOUBLE-EDGED SWORD OF RECOMBINATION

We will use this measure of truly novel ideas (topic-originating patents) as the mediator variable in an analysis of the creative processes producing these novel ideas and subsequent economic value (where citations capture the value of the patent). We describe the dependent and independent variables below and explain how their relationships will be tested using structural equation modeling.

Dependent and independent variables

Dependent variables

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 indicate that a patent represents a breakthrough (Trajtenberg 1990). Because studies of innovative breakthroughs vary as to whether they use a count of forward citations or a dummy variable indicating the 'breakthroughs' (for the top tier of cited patents) as the outcome of interest, we examine both dependent variables in our analyses. We examine the 5-year count of forward citations as a dependent variable in negative binomial count models and a dummy variable indicating whether a patent is in the top 5 percent of cited patents in a probit model. The two dependent variables are measured from the grant date in our reported regressions. Our results are robust to the use of a 5-year window since application date.

Independent variables

To test for the competing effects of the tension and foundational views of recombination, we can draw on measures of distant and diverse recombination from prior studies, where higher levels of either measure would capture recombination (the tension view) and lower levels would capture local search (the foundational view).

Technological distance. To measure the distance of the knowledge recombined, we use the technological distance measure proposed by Trajtenberg *et al* (1997), which uses the USPTO classification system. The USPTO categorizes patents into classifications based on a hierarchical organization of knowledge domains. Each patent is given one or more 3-digit classifications where the first digit indicates a broad domain and each successive digit indicates a more specific location in technological space. Each patent also includes a listing of "prior art" patents which are identified by the inventor, her agents and the patent examiner as related to the invention. To examine the distance of recombination, Trajtenberg *et al* (1997) look at all of the 3-digit classes of the prior art patents cited by the focal patent. They measure the distance of the focal patent's prior art based on USPTO patent classifications 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 prior art patents i and j belong to the same 3-digit technological class, 0.33 if they are in the same 2-digit class, 0.66 if they are in the same 1-digit class, and 1 if they are in different 1-digit classes or have no prior art. The higher the value of this measure, the more distant the knowledge combined. (Note that we control for cases where patents have no prior art.)

Technological diversity. To capture the breadth of recombination, we draw on Hall *et al* (2001) who take advantage of the USPTO classification system as well. They use a Herfindahl index of the concentration of USPTO patent classes in the prior art cited by a focal patent in a measure that represents *technological diversity* (which they termed 'patent originality'). The higher value of diversity, the more diverse sources of knowledge a patent combines. As suggested by Hall (2002), we adjusted the measure to correct for downward bias in patents with few citations to prior art.

$$Technolo\widehat{gical}\ diversity_i = \frac{number\ of\ patents_i}{number\ of\ patents_i - 1}(Technological\ diversity_i),$$

$$where\ technological\ diversity\ _i = 1 - \sum_{j}^{n_i} s_{ij}^2$$

and s_{ij} denotes the percentage of citations made by patent i to patents in class j, out of n_i patent classes. For patents that cite no prior art, this measure cannot be calculated. Therefore, we set the value to zero and include, as mentioned above, a dummy (*No prior art*) as a control.

Controls

We control for a variety of measures that will allow us to isolate the knowledge effect of distant and diverse combinations on novelty and economic value.

Previous use of components and combinations. According to Fleming (2001), previous use of the components of an invention and their prior combinations will increase the probability that an invention is useful (highly cited). Controlling for previous use allows us to separate out the effect of distance and diversity of recombination on citation rates from the effect of the field's familiarity with those combinations. These measures take advantage of the fact that in addition to the 3-digit classes assigned by the USPTO, each class is accompanied by a subclass that further narrows the technological domain in question. Familiarity of components is inferred from how frequently and recently a focal patent's subclasses have been used previously by other researchers. This variable – *Ln(component* *familiarity)* – is measured as the average time-discounted count of all previous usage of the focal patent's subclasses across all patents listed by the USPTO. Following Fleming's (2001) formulation:

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

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

Similarly, prior use of combinations – *Ln(combination familiarity)* – is measured as the time-discounted count of the previous use of the focal patent's subclass combination across all patents:

$$cumulative\ comb.use\ of\ patent\ i =$$

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

On the other hand, Fleming (2001) suggests that too much use of a combination may mean that it has been exhausted of its potential. We control for this possibility using the variable *Ln(cumulative combination)*, which is the same as combination familiarity but without the time discount.

Inventor context. It has been argued that inventors with more experience or those working in teams and organizations may be more likely to create valuable inventions. The logic is often based on the idea that such factors will increase recombination. For example, some scholars claim that more experienced inventor teams will be better able to generate recombinations that are valuable (Singh & Fleming 2010; Conti *et al* forthcoming). Collaborations (relative to solo inventors) have been argued to produce more valuable inventions because of the greater diversity of views and backgrounds that enable wider recombinations (Singh & Fleming 2010). Relatedly, it has been shown that if an inventor is embedded in an organization (relative to independent inventors), she could be drawing on a greater variety of accumulated knowledge to make recombinations that are highly cited (Audia & Goncalo 2007; Singh & Fleming 2010).

While the mechanism identified by these scholars is that of broader recombination, it is also equally likely that experience, teams and organizations produce more highly cited patents because of other social processes such as networks or prominence that allow for greater diffusion of ideas separate from the characteristic of the knowledge. On the flip side, we might expect that various forms of experience could be positively correlated with narrow recombination (local search) in that experience could make inventors more prone to recombine familiar and local knowledge rather than exploring new paths. These narrower recombinations might still become more highly cited because experienced inventors would be more prone to avoid combinations that have failed in the past. Given these potential competing effects of inventor context, we introduce controls for each. Inventor experience is measured as the average number of previous patents by the inventors of the focal patent, using a log normal transformation to correct for skewness: *Ln(average experience)*. Following the lead of Singh and Fleming (2010), we measure invention in a team as a dummy variable (*Team*) (but find similar results with the count of team members) and invention in an organization as dummy variable (*Assigned*).

Additional controls. We include four 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 (Podolny & Stuart 1995). Similarly, we control for the number of non-patent references (*# non-patent references*) because they have also been shown to be associated with higher forward citation rates (Gittelman & Kogut 2003; Deng *et al* 1999. We also control for the number of claims (*# 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 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; Gittelman & Kogut 2003, who measure patent families as patents that are patented in multiple jurisdictions). We include annual time dummies as a simple control for possible time trends.

Table 1 shows the means and standard deviations for the whole sample – the majority of which have similar values to those reported in the studies we cite above – as well as for subsamples of topic-originating patents, highly cited patents vs. all others. Note that, relative to other patents, topic-originating patents have higher numbers of citations but statistically significantly lower values for measures of distant and diverse recombinations, an initial indication of support for the foundational view of creativity in producing novelty. In contrast, highly cited patents have higher values for the recombination measures, in line with prior results reported in the innovation literature based on the tension view. Note also that the measure of novelty is positively correlated with economic value. The correlation table (not reported here for reasons of space) confirms these results.

-- Insert Table 1 about here --

Structural equation modeling to test for mediation

We use structural equation modeling (SEM) to examine the double-edge sword of recombination. This is the recommended approach for testing mediated relationships where there are more than one independent variable (the case in our analysis) (Iacobucci *et al* 2007; Zhao *et al* 2010; Cho & Pucik 2005). Positive signs for our recombination variables as tested in paths A, A*B (the indirect effect of recombination on citations as mediated by novelty) and C in Figure 1 would support the tension view of creativity; negative signs (Paths A', A'*B, and C') would be evidence for the foundational view. This approach will also allow us to verify prior studies showing a direct, positive association between distant and diverse recombination and citations (Path C). The advantage of SEM relative to running three separate regressions (the traditional approach to testing mediation, according to Baron & Kenny 1986) is that the simultaneous equations control for measurement errors that might lead to under- or over-estimation of mediation effects (Shaver 2005). The indirect effect of one of the independent variables (IV) on the dependent variable (DV) through the mediator can be calculated by multiplying the estimated direct effect of the IV on the mediator (path A) and the estimated direct effect of the mediator on the DV (path B). As a robustness check, we also conducted a mediation analysis with separate regressions for each path in Figure 1 and found highly consistent results in both the effect size and significance.

The nature of our dependent variables (one is a count and the other binary) and mediator variable (also binary) places additional constraints on the SEM approach. Using a linear model with count and categorical dependent and mediator variables can lead to biased results. We therefore use the Generalized SEM (GSEM) model introduced in Stata 13 that allows generalized linear response functions with count and binary outcomes. We use a negative binomial function for regressions with count outcomes and a probit function for regressions with binary outcomes. Employing a maximum likelihood estimator, GSEM provides consistent, efficient and asymptotically normal estimates for paths A, B and C. We further use nonparametric bootstrapping (with 1,000 replications) to adjust estimates for bias and to estimate the indirect effects (A*B), total effects ([A*B]+C), their standard errors and their confidence intervals. All the significance levels are determined by the bias-adjusted bootstrap confidence intervals (Mooney & Duval 1993, Efron & Tibshirani 1993).

RESULTS

Tables 2 (for citation counts) and 3 (for breakthroughs in the top 5 percent of citations) report the results of the structural equation models. Column 1 shows the direct effect of the independent variables measuring recombination processes and mediator measuring novel ideas on the dependent variable (paths C and B). Column 2 shows the direct effect of recombination on novel ideas (path A). Column 3 identifies mediation effects in the analysis and shows the indirect effect of recombination as mediated by novel ideas (A*B). Column 4 represents the total effect of the independent variables and mediator on the dependent variables, taking into account the direct and indirect effects ([A*B]+C).

-- Insert Tables 2 and 3 about here --

Testing path C

In testing H1a and H1b for path C, we are replicating the prior studies on breakthrough innovations linking distant and diverse recombination with subsequent citations (a measure of economic value). Though the samples of the prior studies are vastly different (in terms of numbers of observations, time periods, technological arenas), we find support for H1a in our fullerene patent dataset in terms of direction and, in one case, significance of effect for each of the dependent variables (Model 1 in Tables 2 and 3). More distant (*technological distance*) and diverse (*technological diversity*) combinations of knowledge are positively associated with subsequent citations, though the effect is only significant for *technological distance* in the patent count model (Table 2). Note also that, as previous studies have found, other measures sometimes linked with recombination – previous use of components and combinations, greater experience of inventors, and invention in organizations and teams – are positively associated with citation rates. The significance of the effects are attenuated when using the dummy variable for citation-based breakthroughs (in Table 3), but this is likely due to the reduction in variance in the dependent variable.

Testing mediation (paths B and A)

Confirming H2, Model 1 in both Tables 2 and 3 shows that the measure of truly novel ideas (topic-originating patents) is strongly positively associated with subsequent citation rates. This relationship (for path B) is statistically and economically significant. Looking at Model 1 in Table 2, a topic-originating patent is likely to receive 1.4 times more citations than the average patent. Similarly, looking at Model 1 in Table 3, if a patent is topic-originating, the odds of it becoming an economic breakthrough as measured by citation rates increase by a factor of 1.7. In other words, holding all other variables at their means, the probability of gaining a breakthrough level of citations is 0.072 for topic-originating patents (those representing novel ideas) compared to 0.024 for other patents. The marginal effect of topic originating patents on the likelihood of becoming a top cited patent is 0.031, which is substantially greater than the marginal effects of the recombination variables.

On the other hand, we do not see the positive effect of distant and diverse recombination on novelty anticipated by tension theories (H3a). First, looking at the direct effect of recombination variables on novel ideas in Model 2, we find that *technological distance* and *technological diversity* are negatively and significantly associated with topic-originating patents. That is, topic-originating patents are not the result of the combination of distant or diverse knowledge but are instead produced through local search (confirming the foundational view of creativity as represented in H3b). Of the recombination-related controls, only inventor experience appears to have a significant association with novelty. The positive and significant signs for experience on both value and novelty suggest that inventors' previous experience in patenting may increase capabilities in recombination (as suggested by the tension view) or deep knowledge in the field (as suggested by the foundation view), or both. Future research might explore the effect of experience on these two different creative processes.

Turning to Model 3, which is the test of mediation, we do not find the complementary mediation suggested by H3a, but instead a competing mediation relationship. Breakthrough novel ideas are associated with higher citation rates, but distant and diverse recombinations do not appear to produce that novelty, and indeed pull in the opposite direction. Because distant and diverse recombinations have a positive direct effect on citations but are negatively associated with cognitive breakthroughs, their total effect (Model 4) is not statistically significantly different from zero. This finding is the essence of the double-edged sword of recombination. These competing mediation effects are statistically significant based on multiple tests (Kenny 2013, MacKinnon & Dwyer 1993, Iacobucci 2012).[5]

Relationship between cognitive novelty and economic value

We find that patents that are especially novel are also especially valuable. Following the methodology introduced by Rysman and Simcoe (2008) to evaluate patent citation patterns and rates adjusted for confounding factors such as cohort effects, we found that topic-originating patents are more likely to have higher citations than other patents both in their first generation and in their second generation (that is, in patents citing patents that cite the focal patent) (results available from the authors).

On the other hand, this relationship is not perfect. As mentioned above, only 20 of the 189 cognitive breakthroughs (as measured by topic modeling) are also economic breakthroughs (patents in the top 5 percent of 5-year forward citations – there are 109 of these in our dataset). Our method thus highlights the separate but interrelated nature of cognitive and economic breakthroughs. Those patents that represent breakthrough levels of both novelty and value are rare (less than one percent of our

[5] In the first set of mediation estimations reported in Table 2, the main dependent variable is a count variable (citation counts), whereas the mediator is a binary variable (a dummy for whether or not the patent is topic originating). Hence estimates associated with paths A and B in Figure 1 are calculated using different response functions (probit versus negative binomial). As a result, one might be concerned that the estimates for the total indirect effects (A*B) and their standard errors might not be accurate. While the properties of GSEM estimates and the bootstrapping method should take care of this concern, we nevertheless performed a robustness test suggested by Iacobucci (2012) to examine the significance of the indirect effects. Here we find the z-statistic of each indirect effect is significant, consistent with the results from the GSEM method. Furthermore, in the estimations where both the dependent variable and the mediator are dummies (as reported in Table 3), we used the method proposed by Kenny (2013, see also, MacKinnon & Dwyer 1993) as a robustness check of our results. In this method, the estimates for paths A and B are first scaled to similar levels and then their product is estimated using the delta method. Here, again, we find the indirect effects are significant. Both robustness tests confirm the findings estimated by the GSEM method.

sample) but appear to have a greater impact on future innovation than any other kind of invention. As a result, our approach may offer empirical handholds for addressing questions of cumulative research, especially where, as Scotchmer (1991, p. 39) suggests, a 'first technology has very little value on its own but is a foundation for second generation technologies' (see also Furman & Stern 2011).

There are several reasons to believe that generating novel ideas may not automatically lead to breakthrough levels of citations (which represent economic value). A novel idea embodied in a patent still depends on other factors to become known and used, including the reputation and status of the inventors (Merton, 1968; Azoulay, Stuart & Wang, 2014), the distribution of the idea in the relevant network (Singh, 2005), the match of the invention with the environmental demand (Sørensen & Stuart 2000) and the presence of complementary technologies (Rosenberg 1996). In the absence of such factors, a patent that represents a novel idea may not gain traction. Similarly, not all highly cited patents represent novel breaks in knowledge. Patents with a broad scope and general claims, patents inside patent thickets (dense networks of patents with overlapping claims), patents that make an original idea more understandable and usable or patents that distribute an idea strategically in a network, all may lead to higher citations whether or not the patent introduces a truly novel idea. By adding a direct measure of novelty, our analysis is a first step in isolating the effects of novelty from these social dynamics (often associated with recombination processes) that should increase value.

DISCUSSION AND CONCLUSION

Our primary objective for this study was to examine the double-edged sword of recombination in creating innovative breakthroughs. To do this, we look at the effect of distant and diverse recombinations on two innovative outputs: novelty and economic value. We introduce a new method – topic modeling – for measuring the novelty of ideas embedded in patent texts by identifying those patents that originate new topics in a body of knowledge. This measure allows us to distinguish inventions that are cognitively novel in the Kuhnian sense – they introduce new language and therefore new ways of thinking – from inventions that are economically valuable (as measured by the subsequent citations they receive). It also enables us to examine contrasting creative processes – those based in either ‘tension’ or ‘foundational’ assumptions – that contribute to novelty and value.

Implications for understanding breakthroughs

Our approach takes seriously the idea put forth by Griliches (1990) and pursued in recent studies (Jaffe *et al* 2000; Alcácer & Gittelman 2006; Alcácer *et al* 2009; Benner & Waldfogel, 2008; Hegde & Sampat, 2009; Tan & Roberts, 2010) that patents should be assessed as historical documents produced by inventors, prosecuted by patent attorneys and evaluated by patent examiners. An implication is that it should be useful to analyze the texts in these patents, which is also consistent with the cognitive turn being made in studies of technology emergence and evolution (Kaplan & Tripsas, 2008). In doing so, we complement existing research on technology evolution, in particular that which draws on patent data to understand the sources of innovation.

The imperfect relationship between topic-originating patents and those that receive high citations may indicate that there are different kinds of ‘breakthroughs,’ those that introduce truly novel knowledge and those that are associated with economic value. Distinguishing between the novel and the valuable (and understanding the sources of each) is quite important for several reasons. Breakthroughs in knowledge mark the potential origins of new technological paradigms. Furthermore, identifying the patents that mark shifts in knowledge may help 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 (Podolny & Stuart 1995).

By operationalizing the concept of novel ideas implicit in many studies of the sources of innovation, we are able to distinguish processes that produce novelty from those that produce economic value. The contrasting results for distant and diverse recombination are particularly striking. They suggest that generating new topics require deep immersion in a narrower domain rather than linking to more distant or diverse knowledge. On the other hand, patents that cite prior art from a wide range of patent classes are more applicable in a variety of domains and therefore more likely to be cited in the future. These findings highlight the double-edged sword of recombination based on the tension and foundational models of the role of knowledge in creativity (Weisberg 1999; Taylor & Greve 2006).

Theories of recombination are based in the former, while our results on the sources of cognitive breakthroughs are better explained by the latter: local search to uncover anomalies is more likely to produce breaks in the existing knowledge and language. Identifying breakthroughs in knowledge using topic modeling may help us develop further insights into Kuhn's (1962/1996, p. 62) model of technological change based on, 'the previous awareness of anomaly, the gradual and simultaneous emergence of observational and conceptual recognition, and the consequent change of paradigm categories and procedures.'

As an early foray into the use of a new method, this study is, not surprisingly, constrained by some limitations. Most importantly, topic modeling is sensitive to the corpus of documents selected for the analysis. Because the technique is based in the generation of posterior probabilities, the identification of topics and topic-originating patents will be affected by which documents are included in the analysis. This is in turn affected by which inventions are patented and by which documents the researcher selects to include in the corpus. Patents are an imperfect source of information on new scientific and technological ideas. Not all inventions are patented (Scherer 1983; Griliches 1990). We are therefore surely missing ideas and topics that withered on the vine. This constraint is balanced by the rich bibliometric data that patents provide, which allow the scholar to examine the effects of citations, inventors, assignees, patent classes and the like. We have also addressed potential bias in our sample that pre-established classification systems create by using three different search methodologies to identify patents related to fullerenes.

Implications for organizations

Our results are shaped by the possibility that the processes we observe are endogenous to each other. We measure the impact of recombination on two different aspects of innovation: novelty and value. Assuming that individuals and organizations are strategic in setting their goals, they simultaneously decide about how much novelty and value they should pursue in their innovative activities. In other words, the decision to achieve a certain amount of economic value is simultaneously determined with the decision to achieve a certain amount of novelty. As a result, what we observe in our data is the resulting outcome of such simultaneous decisions by individuals and organizations about how much effort to place on long jump or local search. In that sense, while pursuing novelty can lead to high citation rates (as we see in our results), pursuing economic value at the same time can influence the level of effort put by individuals and organizations to achieve novel innovations.

This endogeneity has an empirical implication. The simultaneity and mutual dependency between the two decisions directly influence the number of valuable vs. novel innovations we observe in our sample and also the size and significance of the regression coefficients. One can think of another equilibrium in which organizations would have put much more effort in finding novel innovations, which could change the number of topic originating patents in our sample and consequently the size and significance of the effects we find. In that sense, what we measure in paths A and B is influenced by what we measure in path C and vice versa (and this is precisely why we use the GSEM methodology). While this simultaneity and correlation of outcomes can influence our estimated coefficients, our results nevertheless highlight important contrasting effects from distant and diverse recombination on novel and valuable innovative outcomes.

This endogeneity also has an organizational implication. The results suggest that an effective innovation strategy needs to bridge between wide recombination and local search in order to facilitate the transformation of novel ideas into economically valuable ones. By understanding the double-edged sword of recombination in driving novelty and value, organizations can think about how to manage these conflicts. The literature has not yet studied the ways in which tension and foundational views of creativity interact. These views have been positioned as alternative theories of creativity rather than as two processes operating simultaneously in organizations. Our model might be consistent with a variation-selection-retention view of innovation where variation (novelty) is produced by one set of processes while selection and retention of the most potentially valuable ideas is produced by another. Future research could explore the organizational design implications of the presence of these conflicting effects, potentially across different stages of innovation.

Extensions of topic modeling as a tool in studies of innovation

In addition to identifying the sources and impacts of breakthroughs, topic modeling may usefully contribute to other areas of research on science, technology and knowledge. For example, topic modeling can allow us to analyze at a more fine-grained level technological distance, ties and spillovers between firms and other entities. To date, this research has primarily been conducted through cross-citation analyses of the overlaps in USPTO patent classifications amongst the patents of different entities (e.g., Jaffe 1986; Ahuja 2000; Song *et al* 2003) or citations between entities (e.g., Jaffe *et al* 1993; Mowery *et al* 1996; Henderson *et al* 1998). Scholars are increasingly raising concerns about the degree to which patent classifications are proxies of location in technological space (Benner & Waldfogel 2008) and about the noisiness of patent citations as measures of knowledge flows (Duguet & MacGarvie 2005; Alcácer & Gittelman 2006; Roach & Cohen, 2013). In the case of nanotechnology, for example, ethnographic research has found that knowledge flows are not fully captured by co-authoring and citations, where exchanging students and experimental materials, commenting on each other's work, or participating in problem-solving workshops were more powerful and frequent mechanisms (Mody 2011). Yet, we have lacked reasonable alternative quantitative measures for knowledge flows (Roach & Cohen 2013).

Topic modeling of patents may provide one solution to augment existing approaches. Because topic models produce a vector of weights of each topic for each patent, one can evaluate the content of ties using topics and the strength of ties using weights. This approach may be a useful complement to patent classes because it tracks the language of the actors rather than the classifications assigned by others. It also adds greater nuance than available in current cross-citation approaches by, first, examining the ideas directly rather than inferring them from citation ties and, second, allowing for the possibility that connections amongst ideas occur even if the patents are not cited.

Topic modeling, thus, offers a new means of generating inductively classifications of ideas from texts, which may be advantageous as we look beyond patents to other collections of documents. With the burgeoning interest in classification and categorization (e.g., Lounsbury & Rao 2004; Kennedy

2005; Navis & Glynn 2010; Pontikes 2012), topic models can identify themes or frames as they emerge and evolve over time (e.g., Ruef & Nag, 2014; DiMaggio *et al*, 2013). This approach has the distinct appeal of dispensing with the requirement to use pre-established categories or to come up with ex-post classification systems. Instead, the data can speak for themselves, thus allowing the researcher to observe paths that fall away as well as paths that become consolidated over time. As such, topic modeling could be vital in understanding the emergence and institutionalization of new fields. We hope that our early foray into the application of topic modeling to social science questions can instigate further explorations in these directions.

References

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

Agrawal A, Goldfarb A, Teodoridis F. 2013. Does Knowledge Accumulation Increase the Returns to Collaboration? NBER Working Paper 19694, National Bureau of Economic Research, Cambridge, MA. Available at: http://www.nber.org/papers/w19694.

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

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

Ahuja G, Coff R, Lee P. 2005. Managerial Foresight and Attempted Rent Appropriation: Insider Trading on Knowledge of Imminent Breakthroughs. Strategic Management Journal 26(8): 791-808.

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

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.

Amabile TM. 1983. The Social-Psychology of Creativity – A Componential Conceptualization. Journal of Personality and Social Psychology 45(2): 357-376.

Amabile TM, Pillemer J. 2012. Perspectives on the Social Psychology of Creativity. Journal of Creative Behaviour 46(1): 3-15.

Audia PG, Goncalo JA. 2007. Past Success and Creativity Over Time: A Study of Inventors in the Hard Disk Drive Industry. Management Science 53(1): 1-15.

Azoulay P, Ding W, Stuart T. 2007.The Determinants of Faculty Patenting Behavior: Demographics or Opportunities? Journal of Economic Behavior & Organization 63(4): 599-623.

Azoulay P, Stuart T, Wang Y. 2014. Matthew: Effect or Fable? Management Science 60(1): 92-109.

Baron RM, Kenny DA.1986. The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations. Journal of Personality and Social Psychology 51(6): 1173-1182.

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

Bessen J. 2008. The Value of US Patents by Owner and Patent Characteristics. Research Policy 37(5): 932-945.

Blei DM. 2012. Probabilistic topic models. Communications of the ACM 55(4): 77-84.

Blei DM, Lafferty JD. 2007. A Correlated Topic Model of Science. Annals of Applied Statistics 1(2): 634-634.

Blei DM, Ng AY, Jordan M. 2003. Latent Dirichlet allocation. Journal of Machine Learning Research 3: 993–1022.

Bresnahan TF, Trajtenberg M. 1995. General purpose technologies: Engines of growth? Journal of Econometrics 65(1): 83-108.

Campbell D. 1960. Blind variation and selective retention in creative thought as in other knowledge processes. Psychological Review 67(6): 380-400.

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.Available at: http://goo.gl/QzeEv9.

Cho HJ, Pucik V. 2005. Relationship between innovativeness, quality, growth, profitability, and market value. Strategic Management. Journal 26(6): 555-575.

Cockburn IM, Henderson RM. 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. Forthcoming. Learning to Be Edison? Individual Inventive Experience and Breakthrough Inventions. Organization Science. Available at: http://dx.doi.org/10.1287/orsc.2013.0875
Csikszentmihalyi M. 1996. Creativity: Flow and the Psychology of Discovery and Invention. HarperCollins Publishers: New York, NY..
Dahlin KB, Behrens DM. 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.
de Solla Price D. 1965a. Is technology historically independent of science? A study in statistical historiography. Technology and Culture 6: 553–568.
de Solla Price D. 1965b. Little Science, Big Science. Columbia University Press: New York, NY..
DiMaggio P, Nag M, Blei D. 2013. Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of Government Arts Funding in the U.S. Poetics 41(6): 570-606.
Duguet E, MacGarvie M. 2005. How Well do Patent Citations Measure Flows of Technology? Evidence from French Innovation Surveys. Economics of Innovation and New Technology 14(5), 375-393.
Duriau VJ, Reger RK, Pfarrer MD. 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.
Efron B, Tibshirani R. 1993. An Introduction to the Bootstrap. Chapman and Hall: New York, NY..
Emma P. 2006. How to Write a Patent. IEEE Micro 26(1) 143-144.
Ethiraj SK, Levinthal D. 2004. Modularity and innovation in complex systems. Management Science 50(2): 159-173.
Fleming L. 2001. Recombinant Uncertainty in Technological Search. Management Science 47(1): 117-132.
Furman JL, Stern S. 2011. Climbing Atop the Shoulders of Giants: The Impact of Institutions on Cumulative Research. American Economic Review 101(5) 1933-1963.
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.
Griliches Z. 1990. Patent Statistics as Economic Indicators - A Survey. Journal of Economic Literature 28(4): 1661-1707.
Grimmer J, Stewart BM. 2013. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. Political Analysis 21(3): 267-297.
Grun B, Hornik K. 2011. Topic models: An R Package for Fitting Topic Models. Journal of Statistical Software 40(13): 1-30.
Guilford JP. 1967. The nature of human intelligence. McGraw-Hill: New York, NY..
Hall BH, Jaffe A, Trajtenberg M. 2005. Market Value and Patent Citations. The RAND Journal of Economics 36(1): 16-38.
Hall BH, Jaffe A, Trajtenberg M. 2001. The NBER Patent Citation Data File: Lessons, Insights, and Methodological Tools. NBER Working paper 8498, National Bureau of Economic Research, Cambridge, MA. Available at: http://www.nber.org/papers/w8498.
Hall BH. 2002. A Note on the Bias of Herfindahl-Type Measures Based on Count Data. In Patents, Citations, and Innovations, Jaffe A, Trajtenberg M (eds). MIT Press: Cambridge, MA: 454-459.

Hall D, Jurafsky D, Manning CD. 2008, Studying the Histories of Ideas Using Topic Models. In Proceedings of The Conference on Empirical Methods in Natural Language Processing. Avilable at: http://dl.acm.org/citation.cfm?id=1613763.
Hannan MT, Pólos L, Carroll G. 2007. Logics of organization theory: audiences, codes, and ecologies. Princeton University Press: Princeton, NJ.
Hargadon A, Sutton RI. 1997. Technology Brokering and Innovation in a Product Development Firm. Administrative Science Quarterly 42(4): 716-749.
Harhoff D, Narin F, Scherer FM, Vopel K. 1999. Citation Frequency and the Value of Patented Inventions. The Review of Economics and Statistics 81(3): 511-515.
Hegde 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 General Purpose Technologies and Economic Growth, Helpman E (eds). MIT Press: Cambridge, MA.: 85-117.
Henderson R, Jaffe AB, Trajtenberg M. 1998. Universities as a Source of Commercial Technology: A Detailed Analysis of University Patenting, 1965-1988. Review of Economics and Statistics 80(1) 119-127.
Hsu G, Kocak O, Hannan MT. 2009. Multiple Category Memberships in Markets: An Integrative Theory and Two Empirical Tests. American Sociological Review 74(1): 150-169.
Huff AS. 1990. Mapping strategic thought. John Wiley and Sons: Chichester, NY.
Iacobucci D. 2012. Mediation analysis and categorical variables: the final frontier. Journal of Consumer Psychology 22(4): 582-594.
Iacobucci D, Saldanha N, Deng X. 2007. A Meditation on Mediation: Evidence that Structural Equations Models Perform Better than Regressions. Journal of Consumer Psychology 17(2): 140-154.
Jaffe AB. 1986. Technological Opportunity and Spillovers of Research-and-Development - Evidence from Firms Patents, Profits, and Market Value. American Econonomic Review 76(5): 984-1001.
Jaffe AB, 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.
Jaffe AB, Trajtenberg M, Henderson R. 1993. Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations. The Quarterly Journal of Economics 108(3): 577-598.
Jones BF. 2009. The Burden of Knowledge and the Death of the Renaissance Man: Is Innovation Getting Harder? Review of Economic Studies 76(1): 283-317.
Kaplan, Sarah. 2008. Cognition, Capabilities and Incentives: Assessing Firm Response to the Fiber-Optic Revolution. Academy of Management Journal. 51(4), 672-695.
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.
Kaplan S, Tripsas M. 2008. Thinking about Technology: Applying a Cognitive Lens to Technical Change. Research Policy 37(5): 790-805.
Kennedy MT. 2005. Behind the One-Way Mirror: Refraction in the Construction of Product Market Categories. Poetics 33(3-4): 201-226.
Kenny D. 2013. Mediation with dichotomous outcomes. Working paper. available at: http://davidakenny.net/doc/dichmed.pdf.
Kuhn TS. 1962/1996. The Structure of Scientific Revolutions. University of Chicago Press: Chicago, I.L.
Kuusi O, Meyer M. 2007. Anticipating technological breakthroughs: Using bibliometric coupling to explore the nanotubes paradigm. Scientometrics 70(3): 759-777.
Lanjouw JO, Schankerman M. 2004. Patent Quality and Research Productivity: Measuring Innovation with Multiple Indicators. The Economic Journal 114(495): 441-465.
Leydesdorff L, Cozzens S, Van den Besselaar P. 1994.Tracking areas of strategic importance using scientometric journal mappings. Research Policy 23(2): 217-229.

Lounsbury M, Rao H. 2004. Sources of durability and change in market classifications: A study of the reconstitution of product categories in the American mutual fund industry, 1944-1985. Social Forces 82(3): 969-999.
MacKinnon DP, Dwyer JH. 1993. Estimating mediated effects in prevention studies. Evaluation Review 17: 144-158.
March JG. 1991. Exploration and Exploitation in Organizational Learning. Organization Science 2(1): 71-87.
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.
Merton RK. 1968. The Matthew Effect in Science. Science 159(3810): 56-63.
Meyer M, Pereira TS, Persson O, Granstrand O. 2004. The scientometric world of Keith Pavitt, Research Policy 33(9):1405-1417.
Mody CC. 2011. Instrumental Community: Probe Microscopy and the Path to Nanotechnology. MIT Press: Cambridge, MA.
Mogoutov A, Kahane B. 2007. Data search strategy for science and technology emergence: A scalable and evolutionary query for nanotechnology tracking, Research Policy 36(6): 893-903.
Mooney CZ, Duval RD. 1993. Bootstrapping: A nonparametric approach to statistical inference. No. 94-95. Sage: Newbury Park, CA.
Mowery DC, Oxley JE, Silverman BS. 1996. Strategic Alliances and Interfirm Knowledge Transfer. Strategic Management Journal 17: 77-91.
Navis C, Glynn MA. 2010. How New Market Categories Emerge: Temporal Dynamics of Legitimacy, Identity, and Entrepreneurship in Satellite Radio, 1990-2005. Administrative Science Quarterly 55(3): 439-471.
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(4): 369-388.
Podolny JM, Stuart TE. 1995. A Role-Based Ecology of Technological Change. The American Journal of Sociology 100(5): 1224-1260.
Pontikes EG. 2012. Two Sides of the Same Coin: How Ambiguous Classification Affects Multiple Audiences' Evaluations. Administrative Science Quarterly 57(1): 81-118.
Ramage D, Rosen E, Chuang J, Manning C, McFarland DA. 2009. Topic Modeling for the Social Sciences. In NIPS 2009 Workshop on Applications for Topic Models: Text and Beyond .
Rao H, Monin P, Durand R. 2005. Border crossing: Bricolage and the erosion of categorical boundaries in French gastronomy. American Sociological Review 70(6): 968-991.
Roach M, Cohen WM. 2013. Lens or Prism? Patent Citations as a Measure of Knowledge Flows from Public Research. Management Science 59(2): 504-525.
Rosenberg N. 1996. Uncertainty and Technological Change. In The Mosaic of Economic Growth, Landau R, Taylor T, Wright G (eds). Stanford University Press: Stanford, CA: 334-353.
Rosenkopf L, Nerkar A. 2001. Beyond Local Search: Boundary-Spanning, Exploration and Impact in the Optical Disc Industry. Strategic Management Journal 22(4): 287-306.
Rothaermel FT, Thursby M. 2007. The nanotech versus biotech revolution: Sources of Productivity in Incumbent Firm Research. Research Policy 36(6): 832-849.
Ruef M, Nag M. 2014. The Classification of Organizational Forms: Theory and Application to the Field of Higher Education. In Remaking College: The Changing Ecology of Higher Education, Stevens ML, Kirst MW (eds) Stanford University Press: Stanford, CA: 147-160.
Ruef M, Patterson K. 2009. Credit and Classification: The Impact of Industry Boundaries in Nineteenth-century America. Administrative Science Quarterly 54(3): 486-520.
Rysman M, Simcoe T. 2008. Patents and the Performance of Voluntary Standard-Setting Organizations. Management Science 54(11): 1920-1934
Scotchmer S. 1991. Standing on the Shoulders of Giants - Cumulative Research and the Patent-Law. Journal of Economic Perspectives 5(1): 29-41.

Scherer FM. 1983. The propensity to patent. International Journal of Industrial Organization 1(1), 107–128.
Shaver JM. 2005. Testing for Mediating Variables in Management Research: Concerns, Implications, and Alternative Strategies. Journal of Management 31(3): 330-353.
Simonton DK. 1999. Creativity as blind variation and selective retention: Is the creative process Darwinian? Psychological Inquiry 10(4): 309-328.
Singh J. 2005. Collaborative Networks as Determinants of Knowledge Diffusion Patterns. Management Science 51(5): 756-770
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.
Sørensen JB, Stuart TE. 2000. Aging, Obsolescence, and Organizational Innovation. Administrative Science Quarterly 45(1): 81-112.
Sternberg RJ. 1997. Thinking styles. Cambridge University Press: Cambridge, U.K.
Sternberg RJ, Lubart TI. 1995. Defying the crowd: cultivating creativity in a culture of conformity. Free Press: New York, NY.
Sternberg RJ, O'Hara LA. 1999. Creativity and Intelligence. In Handbook of Creativity, Sternberg RJ (eds). Cambridge University Press: Cambridge, U.K.: 251-272.
Tan D, Roberts PW. 2010. Categorical coherence, classification volatility and examiner-added citations. Research Policy 39(1): 89-102.
Taylor A, Greve HR. 2006. Superman or the Fantastic Four? Knowledge Combination and Experience in Innovative Teams. Academy of Management Journal 49(4): 723-740.
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.
Upham SP, Rosenkopf L, Ungar LH. 2010. Innovating Knowledge Communities an Analysis of Group Collaboration and Competition in Science and Technology. Scientometrics 83(2): 525-554.
US Patent and Trademark Office. 2005. Handbook of Classification. United States Government Printing Office: Washington, DC.
Weisberg RW. 1999. Creativity and Knowledge: A Challenge to Theories. In Handbook of Creativity, Sternberg RJ (eds). Cambridge University Press: Cambridge, U.K.: 226-250.
Whorf BL. 1956. Science and linguistics. In Language, thought, and reality: Selected writings of Benjamin Lee Whorf, Carroll JB (eds). MIT Press: Cambridge, MA: 207-219.
Wry T, Greenwood R, Jennings PD, Lounsbury M. 2010. Institutional Sources of Technological Knowledge: A Community Perspective on Nanotechnology Emergence. Research in the Sociology of Organizations 29: 149-176.
Yoon B, Park Y. 2005. A systematic approach for identifying technology opportunities: Keyword-based morphology analysis. Technological Forecasting and Social Change 72(2): 145-160.
Zhao X, Lynch Jr JG, Chen Q. 2010. Reconsidering Baron and Kenny: Myths and Truths about Mediation Analysis. Journal of Consumer Research 37(2): 197-206.
Zuckerman EW. 1999. The categorical imperative: Securities analysts and the illegitimacy discount. America*n Journal of Sociology* **104**(5): 1398-1438.

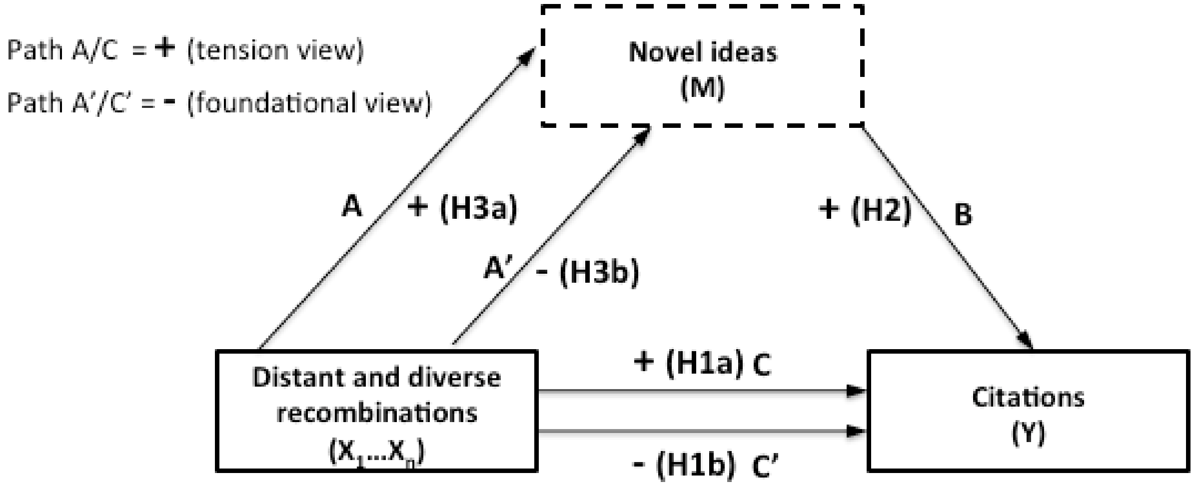


Figure 1: The double-edged sword of recombination

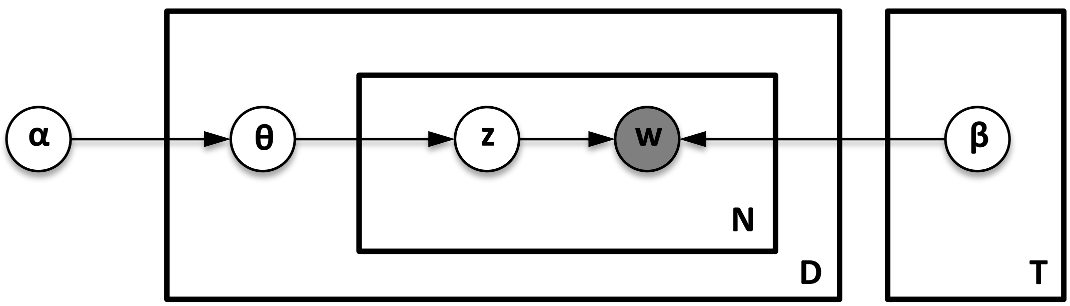


Figure 2: Latent Dirichlet allocation in topic modeling

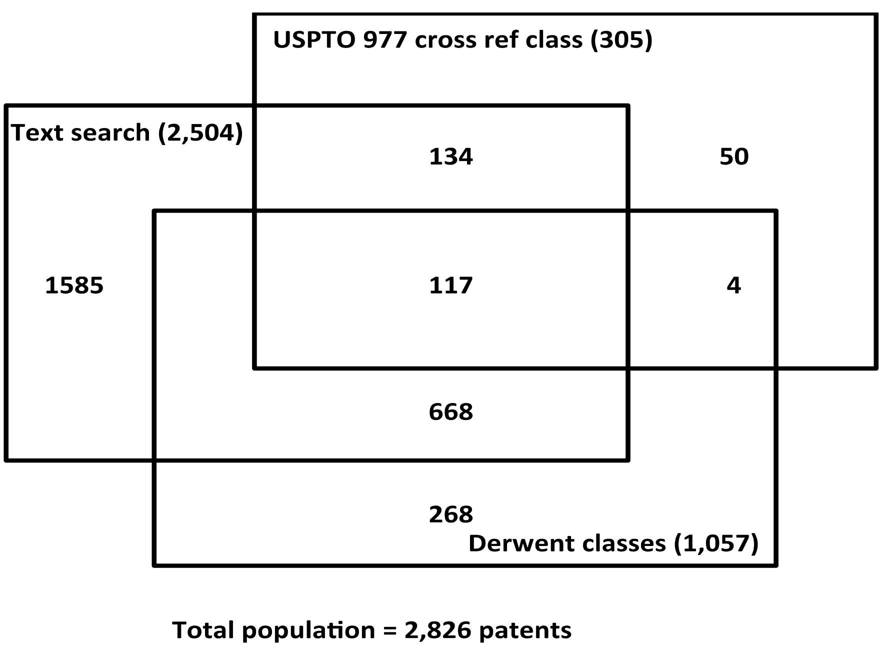


Figure 3: Sample of fullerene and nanotube patents

Table 1: Descriptive statistics, total sample, topic originating patents, top cited patents, and all others

Means and standard deviations, t-test

	All	Topic originating patent=1	Topic originating patent=0	Difference	Top 5% cited patent=1	Top 5% cited patent=0	Difference
Breakthroughs (top 5% cited)	0.048	0.111	0.042	0.069**	1.000	0.000	1.000**
	(0.213)	(0.315)	(0.202)	(p=0.000)	(0.000)	(0.000)	(p=0.000)
# forward citations (5-yr)	13.507	22.173	12.767	9.406**	91.982	9.559	82.422**
	(21.678)	(30.958)	(20.536)	(p=0.000)	(29.963)	(11.106)	(p=0.000)
Topic-originating patent	0.079	1.000	0.000	1.000**	0.183	0.074	0.110**
	(0.270)	(0.000)	(0.000)	(p=0.000)	(0.389)	(0.261)	(p=0.000)
Technological distance	0.405	0.366	0.409	-0.043+	0.474	0.402	0.072*
	(0.328)	(0.331)	(0.327)	(p=0.094)	(0.325)	(0.328)	(p=0.026)
Technological diversity	0.736	0.681	0.741	-0.059*	0.758	0.735	0.023
	(0.323)	(0.353)	(0.320)	(p=0.018)	(0.317)	(0.323)	(p=0.466)
Ln(component familiarity)	4.661	4.337	4.689	-0.353**	4.735	4.658	0.078
	(1.001)	(1.090)	(0.989)	(p=0.000)	(0.834)	(1.009)	(p=0.430)
Ln(combination familiarity)	0.306	0.264	0.309	-0.045	0.225	0.310	-0.084
	(0.773)	(0.709)	(0.778)	(p=0.448)	(0.687)	(0.776)	(p=0.265)
Ln(cumulative combination)	0.409	0.370	0.412	-0.041	0.287	0.415	-0.127
	(0.965)	(0.871)	(0.973)	(p=0.581)	(0.819)	(0.971)	(p=0.179)
Ln(average experience)	2.101	1.961	2.113	-0.152+	2.158	2.098	0.059
	(1.171)	(1.272)	(1.162)	(p=0.096)	(1.197)	(1.170)	(p=0.605)
Team	0.791	0.722	0.797	-0.075*	0.899	0.786	0.113**
	(0.406)	(0.449)	(0.402)	(p=0.017)	(0.303)	(0.410)	(p=0.005)
Assigned	0.889	0.889	0.889	0.000	1.000	0.883	0.117**
	(0.314)	(0.315)	(0.314)	(p=1.000)	(0.000)	(0.321)	(p=0.000)
No prior art	0.048	0.094	0.044	0.051**	0.027	0.049	-0.021
	(0.214)	(0.293)	(0.205)	(p=0.002)	(0.164)	(0.216)	(p=0.308)
Ln(# prior art patents)	2.123	1.896	2.142	-0.246**	2.413	2.108	0.304**
	(1.066)	(1.198)	(1.052)	(p=0.003)	(1.238)	(1.055)	(p=0.004)
Ln(# non-patent references)	1.498	1.641	1.486	0.156	2.360	1.455	0.905**
	(1.301)	(1.341)	(1.297)	(p=0.124)	(1.419)	(1.280)	(p=0.000)
Ln(# claims)	2.835	2.725	2.844	-0.119*	3.133	2.820	0.313**
	(0.764)	(0.816)	(0.759)	(p=0.045)	(0.749)	(0.762)	(p=0.000)
Ln(Family size)	0.126	0.201	0.120	0.081**	0.386	0.113	0.273**
	(0.331)	(0.390)	(0.325)	(p=0.002)	(0.547)	(0.311)	(p=0.000)

Table 2: Tests of mediation (dv=citation counts, 5-year window since grant date) (1991-2005)
Generalized structural equation model using bootstrapping (1000 repetitions), bias-corrected coefficients and robust standard errors

	Direct effect on citation counts (Paths C and B) (1)	Direct effect on topic-originating patents (Path A) (2)	Indirect effect on citation counts (Paths A *B) (3)	Total effect on citation counts ([A * B] + C) (4)
Topic-originating patent	0.330**			0.330**
	(0.094)			(0.094)
Measures of recombination:				
Technological distance	0.211*	-0.391*	-0.129*	0.082
	(0.093)	(0.184)	(0.073)	(0.119)
Technological diversity	0.065	-0.486**	-0.158**	-0.093
	(0.101)	(0.167)	(0.067)	(0.118)
Controls:				
Ln(component familiarity)	0.077**	0.028	0.009	0.086**
	(0.029)	(0.054)	(0.018)	(0.034)
Ln(combination familiarity)	0.286	-0.198	-0.068	0.217
	(0.213)	(0.328)	(0.115)	(0.241)
Ln(cumulative combination)	-0.275+	0.192	0.065	-0.209
	(0.162)	(0.252)	(0.090)	(0.186)
Ln(average experience)	0.070*	0.122**	0.040**	0.110**
	(0.028)	(0.042)	(0.018)	(0.033)
Team	0.226**	-0.134	-0.044	0.183*
	(0.073)	(0.116)	(0.041)	(0.084)
Assigned	0.385**	0.020	0.008	0.393**
	(0.089)	(0.155)	(0.054)	(0.103)
No prior art	0.255	0.491+	0.159+	0.413+
	(0.202)	(0.279)	(0.103)	(0.224)
Ln(# prior art patents)	0.090**	0.249**	0.082**	0.172**
	(0.032)	(0.067)	(0.031)	(0.044)
Ln(# non-patent references)	0.163**	-0.024	-0.007	0.156**
	(0.023)	(0.041)	(0.014)	(0.026)
Ln(# claims)	0.211**	-0.035	-0.012	0.199**
	(0.036)	(0.064)	(0.022)	(0.042)
Ln(Family size)	0.414**	0.113	0.037	0.451**
	(0.088)	(0.121)	(0.042)	(0.096)
Constant	0.106	0.272	0.089	0.195
Year fixed effects	Yes	Yes	Yes	Yes
Observations	2276	2276	2276	2276

** p<0.01, * p<0.05, + p<0.10

Table 3: Tests of mediation (dv=dummy for citation-based breakthroughs, top 5%, 5-year window since grant date) (1991-2005)
Generalized structural equation model using bootstrapping (1000 repetitions), bias-corrected coefficients and robust standard errors

	Direct effect on on citation-based breakthroughs (Paths C and B) (1)	Direct effect on topic-originating patents (Path A) (2)	Indirect effect on citation-based breakthroughs (Paths A *B) (3)	Total effect on citation-based breakthroughs ([A * B] + C) (4)
Topic-originating patent	0.538*			0.538*
	(0.188)			(0.188)
Measures of recombination:				
Technological distance	0.044	-0.391*	-0.209*	-0.165
	(0.189)	(0.184)	(0.126)	(0.227)
Technological diversity	0.118	-0.486**	-0.257*	-0.138
	(0.211)	(0.167)	(0.125)	(0.232)
Controls:				
Ln(component familiarity)	0.103	0.028	0.015	0.118
	(0.058)	(0.054)	(0.031)	(0.065)
Ln(combination familiarity)	0.564	-0.198	-0.111	0.452
	(0.508)	(0.328)	(0.196)	(0.535)
Ln(cumulative combination)	-0.519	0.192	0.105	-0.413
	(0.413)	(0.252)	(0.151)	(0.444)
Ln(average experience)	0.079	0.122**	0.065*	0.145*
	(0.054)	(0.042)	(0.032)	(0.062)
Team	0.320*	-0.134	-0.072	0.248
	(0.175)	(0.116)	(0.070)	(0.190)
Assigned	4.522**	0.020	0.011	4.533**
	(0.265)	(0.155)	(0.091)	(0.278)
No prior art	-0.462	0.491+	0.261+	-0.202
	(1.093)	(0.279)	(0.180)	(1.119)
Ln(# prior art patents)	-0.062	0.249**	0.132*	0.070
	(0.066)	(0.067)	(0.057)	(0.077)
Ln(# non-patent references)	0.247**	-0.024	-0.011	0.235**
	(0.047)	(0.041)	(0.023)	(0.052)
Ln(# claims)	0.200**	-0.035	-0.019	0.180*
	(0.072)	(0.064)	(0.037)	(0.078)
Ln(Family size)	0.550**	0.113	0.061	0.611**
	(0.128)	(0.121)	(0.072)	(0.144)
Constant	-12.737**	0.272	0.142	-12.594**
	(0.650)	(0.404)	(0.229)	(0.667)
Year fixed effects	Yes	Yes	Yes	Yes
Observations	2276	2276	2276	2276

** p<0.01, * p<0.05, + p<0.10

Appendix For: The Double-Edged Sword of Recombination in Breakthrough Innovation. Sarah Kaplan & Keyvan Vakili. *Strategic Management Journal*.

In this online appendix, we present some additional details on the topic modeling methodology used in "The Double-Edged Sword of Recombination." A good overview of the approach can be found in Blei (2012) and more technical details in Ramage *et al* (2009) and McFarland *et al* (2013).

Latent Dirichlet Allocation algorithm

The topic modeling approach used in our paper is based on a Bayesian statistical technique called Latent Dirichlet Allocation (LDA). The core idea behind the LDA algorithm is described in the section 'A primer on topic modeling' in the accompanying article. Also, Figure 2 in the paper illustrates a schematic representation of the relationship between the various elements in the algorithm, including topics, words, and documents. This section describes the fundamental mathematical model behind the LDA algorithm.

Let $\emptyset_1$,…, $\emptyset_T$ be T topics. Each topic is a distribution over a fixed set of terms. The Latent Dirichlet Allocation (LDA) method assumes each document has its own set of topic proportions θ, but the whole corpus is governed by the same set of latent topics $\emptyset_{1|T}$. In other words, for each document in the corpus, LDA assumes the following generative probabilistic process:

1. Choose θ, topic proportions $\sim$ Dirichlet(α)
2. For each term:
   a. Choose a topic assignment $z_n \sim$ Multinomial(θ)
   b. Choose term w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic assignment from the previous step.

θ is a T-dimensional Dirichlet random variable with values in the $(T-1)$-simplex and the following probability density:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{T} \alpha_i)}{\prod_{i=1}^{T} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \ldots \theta_T^{\alpha_T - 1}$$

where $\Gamma(x)$ is the Gamma function and α is a vector with k components $\alpha_i > 0$. The Dirichlet distribution over the simplex is in the exponential family and has convenient properties that facilitate the process of inference and parameter estimation.

With parameters α and β as given, the joint distribution of the set of N words w, the set of topic assignments z, and the topic proportions θ is as follows:

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta)$$

where $p(z_n|\theta)$ is equal to θ_i for the unique i at which $z_n^i = 1$. The marginal distribution of a document can then be calculated by integrating over θ and summing over z:

$$p(w|\alpha, \beta) = \int p(\theta|\alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta) p(w_n|z_n, \beta) \right) d\theta$$

Now we can calculate the probability distribution of a corpus by taking the product of the marginal probabilities of all the single documents in the corpus:

$$p(D|\alpha,\beta) = \prod_{d=1}^{M} \int p(\theta|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_n|\theta_d) p(w_n|z_n,\beta) \right) d\theta_d$$

The central problem in LDA is to calculate the conditional distribution of the unobserved variables: topics ($\emptyset_T$), topic proportions θ, and topic assignments z_n. The set of these hidden variables is called posterior. There are several algorithms proposed to estimate the posterior. The standard approach uses a Gibbs sampling, in which a Markov chain – a sequence of random variables each dependent on the previous one – with a limiting distribution converging to the posterior is constructed. The algorithm runs a series of Markov chains defined based on the hidden variables, collects samples from the limiting distributions, and calculates an approximation for the distribution based on the collected samples. See Steyvers and Griffiths (2006) for an elaborated description of the Gibbs sampling and the estimation method. Also see Ramage *et al.* (2009) and McFarland *et al.* (2013) for more details on the LDA algorithm.

Input parameters for topic modeling analysis

The goal of topic modeling techniques as developed in the computer sciences is unsupervised analysis of text designed both to generate a predictive model to aid search and to provide a representation of the topics in an existing corpus (Hall *et al.* 2008; Chang *et al.* 2009). In our paper, we focus on this second goal, and we use our data to track the emergence of new ideas (and hence new language to describe these ideas) over time and identify the patents that lead to these shifts in language. In the five sections under the header 'Topic modeling of patent texts: a measure of novel ideas' in the paper, we delineate the topic modeling technique and how we use it to identify the patents that originate new topics. In this section, we describe the specific set of parameters that are required for topic modeling analysis and our particular choices for the analyses reported in the paper.

All topic-modeling techniques require a text corpus as input. Our input corpus comprises the abstracts of 2,384 patent families based on the 2,826 fullerene and nanotube patents. The process of identifying the fullerene patents and patent families are explained in the 'Deriving fullerene and nanotube topics' section in the paper.

Furthermore, the LDA algorithm requires inputs for two specific parameters: α (sometimes called 'topic smoothing') and β (sometimes called 'term smoothing'). Values above one lead to more even distributions. Values below one favor more concentrated distributions across fewer topics or words. In order to produce semantically meaningful topics, most topic modeling toolkits (including the 'Stanford Topic Modeling Toolbox' that we used) recommend 0.1 for both parameters as a default. As a smaller β results in more fine-grained topics (Griffiths & Steyvers, 2004), we lowered this parameter to 0.01 because we are studying a narrow field of technology. The algorithm thus allocates documents to the fewest topics possible while at the same time assigning a high probability to as few words as possible for each topic.

The algorithm also requires input for the number of topics that it would produce at the end. There are a few methods to optimize the number of topics to achieve a best-fit outcome. However, such best-fit models often produce a large number of topics that do not represent distinct meanings (Chang *et al.* 2009). Hence, following the approach by computer science scholars (Blei & Lafferty 2007; Hall *et al* 2008), we capped our model to 100 topics. The section 'Deriving fullerene and nanotube topics' in the paper describes the process of extracting, labeling, and validating the topics.

Figure A1 shows a sample abstract and the weight of its most important topics. Patent number 7288970 'Integrated nanotube and field effect switching device' is dominated by topic 24 ('*Nanotube switching devices and applications*') with 63 percent weight and topics 54 ('*Electronic implementations of look up tables*') and 49 ('*Field emissions display devices*') each with 5 percent weight. No other topic is greater than 5 percent weight. The sum of the weights of all the topics for any given patent is 100 percent. Figure A2 provides graphical representations of sample topics with the top 20 words associated with each sized in proportion to their importance.

Figure A1: Example coded abstract

Patent Number: US7288970
Title: Integrated nanotube and field effect switching device
Inventors: Bertin, Claude L.
Assignee: Nantero, Inc.
Application Date: January 2005, **Issue Date**: October 2007
USPTO classifications: 326/120, 326/112, 977/940

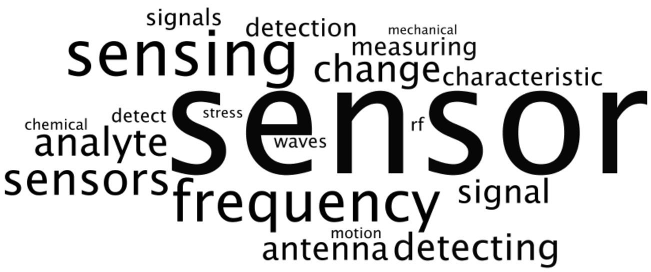
> **Abstract**
> Hybrid[57] switching[24] devices[32] integrate[54] nanotube[56] switching[24] elements[25] with field[49] effect[52] devices[32], such as NFETs and PFETs. A switching[24] device[93] forms[46] and unforms[24] a conductive[59] channel[52] from the signal[24] input[24] to the output[24] subject[22] to the relative[86] state of the control[24] input[24]. In embodiments[94] of the invention, the conductive[59] channel[52] includes a nanotube[56] channel[52] element[25] and a field[49] modulatable[58] semiconductor[46] channel[52] element[25]. The switching[24] device[93] may include a nanotube[56] switching[24] element[25] and a field[49] effect[52] device[93] electrically[59] disposed[42] in series[24]. According to one aspect[33] of the invention, an integrated[54] switching[24] device[93] is a four-terminal[29] device[93] with a signal[24] input[24] terminal[42], a control[24] input[24] terminal[29], a second input[24] terminal[29], and an output[24] terminal[29]. The devices[32] may be non-volatile[24]. The devices[32] can form[24] the basis for a hybrid[57] NT-FET[52] logic[54] family and can be used to implement[82] any Boolean logic[54] circuit[54].

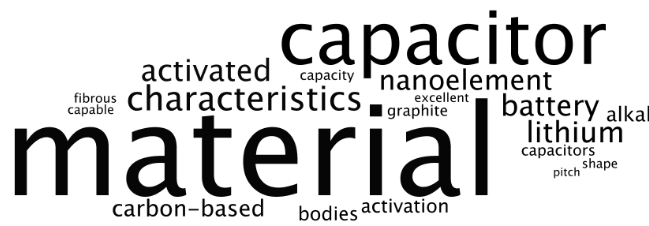
Topic 24 (Nanotube switching devices and applications): 63%
Topic 54 (Electronic implementations of look-up-tables): 5%
Topic 49 (Field emissions display devices): 5%
The rest: less than 5% each for a total of 27%

Figure A2: Graphical representation of words in example topics

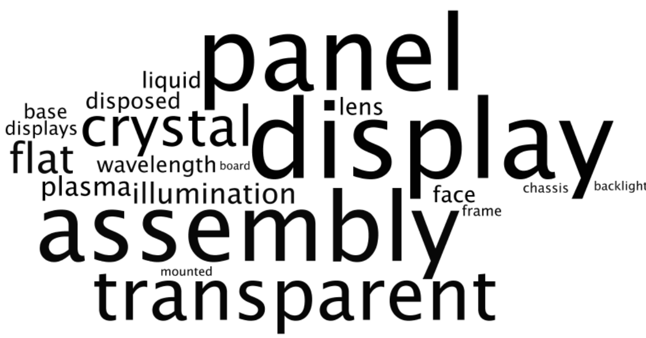
Topic 14: Sensors and detectors (1993)

Topic 60: Application to batteries and charge storage devices (1998)

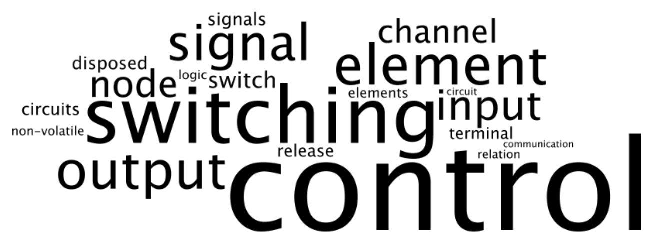




Topic 58: Application to plasma display panels (1997)



Topic 24: Nanotube switching devices (2004)



Note: date of topic-originating patent in parentheses. Size of words based on importance.

Checking the validity of the topic selection process

We used three expert coders to validate and label the topics. The coders found 25 topics to be very general and therefore difficult to label with distinctive codes. This is not surprising as topic models tend to place noisy data into broad or uninterpretable topics (which serves to bolster the coherence of the other topics). We performed the statistical analyses reported in the paper omitting these general topics and found fully consistent results. However, since our interest is in introducing a replicable approach using unsupervised analysis of texts, in the paper we only report the results based on all 100 topics rather than the results that depend on human intervention by coders.

As explained in the paper, we use a threshold for topic weight to identify the first patents that represent each topic. In the paper, we explain the details of how we chose the threshold in the section ‘Deriving fullerene and nanotube topics.’ For robustness tests, we also explored other methodologies for identifying topic-originating patents, such as selecting the first patent to represent a substantial jump in weight relative to prior patents. For example, if we look at patents that represent a 3 standard deviation jump from prior patents and all subsequent patents over that weight in the first year, we obtain a list of 141 patents, of which 73% are the same as those topic-originating patents identified using our 0.2 threshold. The results using this alternative measure are substantially the same as those we report in the paper. Note, however, that in every case, the selection of a topic-originating patent requires an assumption about a threshold, whether it is a weight of 0.2 or a number of standard deviations. Therefore, we prefer to use the threshold that has been validated by coders.

References

Blei DM. 2012. Probabilistic topic models. Communications of the ACM 55(4): 77-84.

Blei DM, Lafferty JD. 2007. A Correlated Topic Model of Science. Annals of Applied Statistics 1(2): 634-634.

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. Available at: http://goo.gl/QzeEv9.

Griffiths TL, Steyvers M. 2004. Finding Scientific Topics. Proceedings of the National Academy of Sciences 101: 5228-5235.

Hall D, Jurafsky D, Manning CD. 2008, Studying the Histories of Ideas Using Topic Models. In Proceedings of The Conference on Empirical Methods in Natural Language Processing. Avilable at: http://dl.acm.org/citation.cfm?id=1613763.

McFarland DA, Ramage D, Chuang J, Heer J, Manning CD, Jurafsky D. 2013. Differentiating language usage through topic models. Poetics 41(6): 607-625.

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

Steyvers, M, Griffiths T. 2006. Probabilistic topic models. In Handbook of latent semantic analysis, Landauer T, McNamara D, Dennis S, Kintsch W (eds). Lawrence Erlbaum: Mahwah, NJ: 427-448.