A Dynamic Approach for Identifying Technological Breakthroughs with an Application in Solar Photovoltaics

Bixuan Sun[1*], Sergey Kolesnikov[2], Anna Goldstein[3], Gabriel Chan[1]

[1] Humphrey School of Public Affairs, University of Minnesota

[2] Centre for Environment, Energy and Natural Resource Governance, Department of Land Economy, University of Cambridge

[3] Department of Mechanical and Industrial Engineering, University of Massachusetts Amherst

Abstract

This paper presents a conceptual framework for understanding technological breakthroughs and a novel empirical approach for investigating potential breakthrough inventions in the patent record. We define technological breakthroughs as inventions that are initially novel to a technological field but become increasingly relevant for describing the development of the field over time. We operationalize these notions of novelty and relevance by applying topic modeling to a corpus of the full text of patents in a technological field. The identified topics define a "technological space," from which we develop continuous measures of a patent's novelty and relevance to the mainstream trajectory of technological development in this space over time. Our method allows us to identify potential breakthrough inventions that may have driven changes to the rate and direction of technological development in the focal field. We apply the method to silicon solar photovoltaic patents granted in the United States between 1977 and 1996, generating a list of 98 patents representing potential breakthroughs that can be subsequently validated with other approaches. This method can help researchers identify sources and patterns of technological breakthroughs to inform research and development policy.

Keywords: innovation; technological breakthrough; technological trajectory; technological space; topic modeling; solar photovoltaics

[*] Corresponding author. Email: sunxx731@umn.edu

1. Introduction

Understanding breakthroughs is central to theories of technological change. These theories suggest that technological progress often arises through the recombination of existing knowledge and components, as in theories of "recombinant innovation" (Fleming, 2001). In particular, breakthrough inventions are likely to come from novel combinations of more technologically "distant" prior inventions (Nemet, 2012), although not all novel exploration yields fruitful innovation (Kaplan and Vakili, 2015). Breakthrough inventions can lead to societal and economic change after significant follow-on incremental innovation (Mokyr, 1990). Empirical evidence on the sources of technological breakthroughs can help inform the development of policies and research-management practices to accelerate future innovation. This is especially important in domains critical to sustainable development, such as clean energy technologies, where there is underinvestment in private innovation (Jaffe, Newell and Stavins, 2005).

Many empirical studies of innovation assume high patent citation counts are indicative of breakthrough inventions, yet patent citations do not capture an invention's novelty relative to the technological field. Some scholars have developed more nuanced measures to distinguish "radical" or "destabilizing" inventions from "incremental" or "consolidating" inventions using patent citation networks (Dahlin and Behrens, 2005; Funk and Owen-Smith, 2017). But citation network analysis is also limited: not all relationships between inventions result in a citation, and not all citations represent influence (Jaffe, Trajtenberg and Fogarty, 2000; Kuhn, Younge and Marco, 2020).

In addition, activities that spur innovation are often idiosyncratic and highly contextual. Investigating breakthrough inventions in specific industries can generate a relevant and timely understanding of innovation dynamics for research and development policy decisions. Qualitative methods, such as interviews with industry experts, can reveal deeper insights into the origins of breakthroughs and the circumstances under which novel ideas become integrated with existing technologies. One challenge with qualitative studies, in addition to time and resource constraints, is that they may only identify prominent influential technologies that are familiar to industry experts, while overlooking less well-known cases, including those that involved inter- and trans-disciplinary research.

The goal of this paper is two-fold. First, we present a conceptual framework that defines technological breakthroughs in terms of their potential impact on technological change in a focal field. Specifically, the impact of a potential breakthrough invention is disaggregated into two components, "novelty" and "relevance". Conceptually, a technological breakthrough should be initially novel to the focal field and then become increasingly relevant to the mainstream of the field as it evolves over time—perhaps in response to the breakthrough. Second, to operationalize our theoretical definition, we develop continuous measures of novelty and relevance based on the full texts of patents and provide a set of analytical steps that can identify patents that show patterns consistent with technological breakthroughs. These measures are dynamic, taking into account changes in mainstream technologies over time and the gradual integration process of initially novel inventions into the mainstream of a technology field. A patent whose contents appear novel at first but become less novel and more relevant to the field over time may embody useful knowledge that has influenced the development of the focal field, or, in other words, a technological breakthrough.

We apply our empirical framework to silicon photovoltaics (PV) patents granted by the US Patent and Trademark Office (USPTO) between 1977 and 1996. We select silicon PV as the technology area to demonstrate this methodology because innovation in PV is seen as critical for sustainable development (Dincer, 2000) and because prior research provides a useful comparison for our approach (Nemet and Husmann, 2012). Further, the dynamics of PV innovation are known to have been shaped by several technological breakthroughs (Green, 2005; Husmann, 2011).

Our proposed method of screening for technological breakthroughs is unsupervised, hence reducing the information load for researchers when searching for potential breakthroughs. This method is applicable to a broad range of technological fields, especially in areas where there is a weaker systematic understanding of historical breakthrough channels. Instead of relying on patent citations or patent classification-based measures with their well-known limitations, we use text mining and topic modeling to capture the technical content of inventions, which allows us to create continuous measures of a patent's relationship to changes in the focal technological field.

The rest of this paper is organized as follows. Section 2 reviews prior literature on technological breakthroughs. Section 3 presents our conceptual framework and operationalizes the framework

by describing our novelty and relevance metrics and the analytical steps for identifying potential breakthrough patents in a focal field. Section 4 demonstrates the application of the framework in silicon PV patents. Section 5 discusses further applications of the methodological framework, its limitations, and validation approaches for assessing the candidate breakthrough patents identified by our method. Section 6 concludes.

2. Prior literature on technological breakthroughs

Building on concepts of genetic evolution, Mokyr (1990) frames technological breakthroughs as "punctuated equilibria", sudden outbursts of technological change that occur between long periods of stagnation. In order for a breakthrough invention to create a meaningful technological breakthrough, it must not only be novel; it must also be followed by a set of continuous smaller improvements that complement the initial invention and allow it to integrate into practice. This theoretical proposition has been examined empirically in energy technologies (Popp et al., 2013).

But how do technological breakthroughs occur? Theoretical literature has described the importance of recombination of existing knowledge as the main driver of breakthrough innovation (Usher, 1954; Fleming, 2001). In particular, many have claimed that combinations of distant knowledge or knowledge flows between distinct sectors drive breakthrough innovation (Rosenberg, 1994). However, some have raised the point that not all combinations of distant technologies result in mainstream innovations. For example, Ahuja and Lampert (2001) found an inverted-U relationship between a firm's breakthrough inventions and its exploration of novel technologies outside the organization. Exploring novel technologies can break "familiarity traps" and lay the foundation for breakthrough inventions, but excessive exploration of novel technologies can also become harmful, leading to "frenzies of experimentation" that create information overload and confusion (Levinthal and March, 1993). Similarly, Kaplan and Vakili (2015) described combinations of distant or diverse knowledge as a "double-edged sword," which may be necessary to create breakthroughs but can also be counterproductive when additional research is needed to integrate breakthroughs into the domain in a useful way. Arthur (2007) argued that radically novel technologies come from existing building-block technologies plus inventors' deep theoretical understandings of how to manipulate technologies to serve a

purpose. Therefore, breakthrough inventions should embody novel combinations of existing technologies in ways that are relevant to the focal technological domain.

Existing empirical measures at the invention level largely do not reflect this theoretical understanding of what constitutes technological breakthroughs. Empirical literature mostly relies on patent citation counts and other citation-based metrics to measure the influence of an invention. Breakthrough inventions are often assumed to be associated with highly-cited patents (Trajtenberg et al., 1997; Ahuja and Lampert, 2001; Phene et al., 2006; Schoenmakers and Duysters; 2010; Kelley, Ali and Zahra, 2013; Jaffe and Trajtenberg, 2002), but this metric fails to account both for an invention's novelty and its evolving relationship to the mainstream of a technological field.

There are many *ex-ante* measures of technological novelty derived from the knowledge base of a patent, such as the number of backward citations to scientific articles (Gittelman and Kogut, 2003), lack of citations to prior art (Ahuja and Lampert, 2001), novel combinations of patent subclasses in the patent (Fleming, 2007; Youn et al., 2015; Kim et al., 2016), newly introduced patent classes (Strumsky and Lobo, 2015), emergence of novel citation-linkages between patent classes or scientific disciplines (Verhoeven, Bakker and Veugelers, 2016) or the breadth of patent classes cited by the patent outside its own field (Rosenkopf and Nerkar, 2001; Shane, 2001; Briggs and Buehler, 2018). However, novel patents identified with such approaches do not necessarily represent breakthrough inventions unless they are adopted and used in the focal technological field; as Mokyr notes, "radical insight is not enough…just as a mutant who survives but cannot reproduce" (Mokyr 1990).

Dahlin and Behrens (2005) identify breakthrough inventions that were both novel and useful by integrating both *ex-post* and *ex-ante* metrics based on forward and backward citations. Subsequently, Funk and Owen-Smith (2017) create a dynamic, citation-network-based measure of technological change that distinguishes inventions that consolidate or destabilize existing technology fields. Destabilizing inventions break away from an existing technological trajectory, while consolidating inventions reinforce the status quo trajectory (Funk and Owen-Smith, 2017). However, citation relationships as a measure of knowledge flow or of an invention's impact do not necessarily indicate the meaningful integration of a patent into a focal field.

At best, citations are a noisy proxy for knowledge flows (Jaffe and Trajtenberg, 2002). When measuring knowledge flows, scholars sometimes exclude citations added by examiners without the knowledge of the inventors, as examiner-added citations comprise nearly two-thirds of all citations (Alcácer and Gittelman, 2006). Jaffe, Trajtenberg and Fogarty (2000) found that half of all patent citations did not correspond to any apparent knowledge flows according to the inventors they surveyed. Kuhn, Younge and Marco (2020) show that widespread violation of core assumptions about the patent citation process, along with dramatic changes in patent citation practices over time, threaten the validity and reliability of citation-based metrics.

Other approaches to identifying potential breakthrough inventions are based on the emergence of new terms in the patent text, where the first appearance of a term indicates the emergence of a novel idea, and growing usage of the term indicates an invention's impact on the development of the field. These approaches typically apply modern machine learning and text mining techniques that reduce noise by eliminating unspecific terms and capturing multi-word phrases, synonyms, acronyms and abbreviations (Tseng, Lin and Lin, 2007; Zhang et al., 2014; Suominen and Newman, 2017). Building on an established method of co-word analysis (Callon et al., 1983), scholars have traced the emergence and evolution of novel patterns in keyword vectors (Geum, Jeon and Seol, 2013; Lee, Kang and Shin, 2015; Wang and Chen, 2019), keyword clusters and networks (Yoon and Park, 2004; Lee, Yoon and Park, 2009; Joung and Kim, 2017), semantic structures (Yoon and Kim, 2012; Gerken and Moehrle, 2012) or topics (Kaplan and Vakili, 2015; Ranaei and Suominen, 2017). These methods allow researchers to study how novel ideas within a technological domain emerge and change over time.

However, it is not always the case that more frequent usage or greater network centrality of a particular term is associated directly with an invention's impact. There may be temporary "hype" that does not result in innovation (Suominen and Newman, 2017) or parallel developments in related technologies within the same domain. Domain-level text analyses are well-suited for studies of technology emergence (Rotolo, Hicks and Martin, 2015; Suominen and Newman, 2017) or the construction of technology evolutionary pathways (Momeni and Rost, 2016; Zhang et al., 2017; Huang et al., 2017). But such approaches are not sufficiently granular to track how an individual invention may have impacted the trajectory of a technological domain and if it could be on a pathway toward becoming a technological breakthrough. A more detailed, patent-

level, text-based method is needed to trace the relationship of radically novel ideas to subsequent inventions in the field.

3. The novelty and relevance framework

Building upon prior theoretical and empirical studies, we present a conceptual framework of technological breakthroughs that embodies two important components of an invention: "novelty" and "relevance." Novelty indicates the level of unfamiliarity of an invention to a focal technological field at a certain moment in time, while relevance indicates how closely an invention aligns with the subsequent innovation trajectory in a field.

Specifically, our definition of technological breakthroughs contains the following four features. First, there should be a clearly defined focal technological field that a breakthrough can influence (e.g. energy storage technologies or lithium-ion batteries). Patent classifications and keyword combinations can be a useful guide to determine the scope of the focal field, although caution has to be taken when considering multiple patent classes and, especially, keywords with particular technological products or industrial sectors (Pilkington et al., 2002; Schmoch, 2008; Nemet, 2012; Ranaei and Suominen, 2017). Second, the novelty and relevance of an invention are measured relative to a trajectory of technological change in the focal field (Dosi, 1982) and thus change over time. Third, a breakthrough invention should be measurably distant (i.e., novel) from the mainstream technological trajectory of its focal field at the time of its invention. An important technological breakthrough disrupts the status quo of the dominant technological paradigm, bringing novel ideas and changing the existing pathway of innovation. The fourth and final feature of our definition is that a breakthrough invention integrates into the mainstream of the focal field over time through gradual, follow-on, and complementary inventions. Hence, a breakthrough should become increasingly relevant and decreasingly novel relative to the subsequent mainstream technological trajectory.

We can look to the history of the technological field of solar photovoltaics for a classic example of an important technological breakthrough to illustrate our definition. In 1876, William Grylls Adams and Richard Evans Day first discovered that selenium can produce electrical current when exposed to sunlight, and the first solar cells were invented by Charles Fritts in 1883 (U.S.

Department of Energy, 2004). Improvements in selenium-based solar cells continued along a photovoltaic technological trajectory afterwards. But in 1953, Gerald Pearson at Bell Labs inadvertently made silicon solar cells that were more efficient than selenium solar cells when he was searching for potential silicon applications in electronics (U.S. Department of Energy, 2004). This breakthrough PV invention was inspired by recombinant innovations within the electronics sector and was initially considered novel relative to the selenium-based technological trajectory the field of solar cells had been on. Later innovations in the solar PV field built upon silicon-based technology for applications such as solar-powered satellites and spacecraft, as well as solar modules with improved efficiency and reduced costs. Thus, over time, the Pearson breakthrough became highly relevant for the subsequent development of the solar PV field as silicon solar cells became the mainstream solar PV technology.

Based on this conceptual framework, we present a dynamic and granular approach to identifying potential breakthrough inventions and tracking how breakthrough inventions integrate into a focal field over time based on the novelty and relevance of patents. First, we use topic modeling of the full text content of all patents in the focal field to construct a quantitative representation of the technological space of the field, a "topic space." Then we calculate the average position of patents in the field in a given year, called the "field centroid." Next, we calculate each patent's novelty and relevance metrics based on its position in the topic space and its distance to the evolving technology trajectory traced by the field centroid over time. Finally, we propose a set of analytical steps to identify potential breakthrough inventions based on these metrics.

3.1. Topic model representation of technological space and trajectory

The concept of "technological space" is commonly used in the literature on technological distance and diversity metrics (Jaffe, 1986; vom Stein, Sick and Leker, 2015, Kaplan and Vakili, 2015; Aharonson and Schilling, 2016). Typically, in a technological space, each entity, such as a patent or a firm, occupies a unique position defined by its knowledge base. Position vectors can be based on patent classification (Jaffe, 1986; McNamee, 2013; Aharonson and Schilling, 2016), patent citation network relationships (Kay et al. 2014), or a representation of the technical content of patents by natural language processing methods in a semantic network of keywords (Yoon and Park, 2004; Kim, Suh and Park et al., 2008; Lee, Yoon and Park, 2009), semantic structures (Cascini, Fantechi and Spinicci, 2004; Yoon and Kim, 2011; Gerken and Moehrle,

2012) or latent topics that cover the technological space (Blei and Lafferty, 2007; Kaplan and Vakili, 2015; Suominen et al., 2017).

In our work, we construct a technological space of latent topics that represent the technical content of a focal technology domain. Specifically, we apply the Latent Dirichlet Allocation (LDA) algorithm (Blei et al., 2003), a topic modeling method, to analyze the full text of patents in a corpus of patents related to a focal field of interest and to identify patents' positions in the resulting topic space. The LDA method assigns a patent to a set of latent topics based on word co-occurrence in the patent's full text, providing a reduced-dimension representation of the patent by a vector of topic probabilities. It has been previously shown that topic modeling can be successfully used to characterize scientific (Blei and Lafferty, 2007; Yau et al., 2014) and technological (Suominen et al., 2017; Kaplan and Vakili, 2015) domains defined by sets of scientific publications or patents, respectively. A particularly useful feature of topic modeling for our framework is that it is possible to calculate the Euclidean distance between any two points in the topic space, including vectors corresponding to any two patents in the focal field[1].

We use the full text of patents rather than patent abstracts or patent claims. Although topic modeling of scientific publications is routinely done using abstracts as a corpus (Yau et al., 2014), abstracts are poor representations of the complexity of technical information contained in patents. Often, patent abstracts are purposefully written to obfuscate, rather than reveal, important features of the inventions (Suominen et al., 2017). Patent claims, on the other hand, are not written in natural language; they are highly formalized and often hierarchically built from repeated sentence fragments, and as such are not optimal for text mining methods. Using the full texts of patents, which contain detailed descriptions of inventions and exclude descriptions of prior art, results in a more robust textual corpus.

A final methodological note on topic modeling is that the LDA algorithm requires the number of topics to be set exogenously. Cross-validation with training and test samples can be used to

[1] Specifically, if each entity in the topic space is represented by a vector of topic probabilities and M is the total number of topics in a topic model, then each m^{th} topic vector ($m = 1 \ldots M$) in this space is defined as (s_1,…, s_i,…,s_M), where $i = 1 \ldots M$, $s_i = 1$ if $i = m$ and $s_i = 0$ if $i \neq m$. These vectors are orthogonal (i.e., scalar products of each pair of vectors are equal to 0), and each has a unit length of 1. Therefore, the set of M topic vectors represents an orthonormal basis in the topic space, which in this case can be considered an M-dimensional metric Euclidean space. Euclidean distance is defined for any two points in this metric space, including the points corresponding to patents with coordinates in the orthonormal basis of M topic vectors that represent topic probabilities with values varying between 0 and 1.

determine the optimal number of topics based on various metrics, such as the perplexity value (Blei et al., 2003; Appendix A).

Having calculated the topic distribution for each patent in a technological field, we can characterize the technological mainstream of the field in the topic space in a given year. The vector of the average topic distributions of all patents granted in the focal field in that year is called the "field centroid". This centroid represents the average topical content of that year's inventions in the technological space, while the centroid's movement in the technological space over time represents the mainstream technological trajectory of the field.

Let $p_{i,t}$ denote an *i*th patent granted in year t with a topic distribution denoted by the vector $\boldsymbol{p_{i,t}} = (p_{1,i,t}, p_{2,i,t}, \dots, p_{m,i,t}, \dots, p_{M,i,t})$, where *m* denotes the topic number, $p_{m,i,t}$ denotes the probability that patent $p_{i,t}$ contains content from topic *m*, *M* is the total number of topics, *m* = 1 … *M*. There are N_{t_1} and N_{t_2} patents granted in years t_1 and t_2, and C_{t_1} and C_{t_2} are the centroids of these years with topic distributions calculated as:

$$\boldsymbol{C_{t_1}} = (\bar{p}_{1,t_1}, \dots, \bar{p}_{m,t_1}, \dots, \bar{p}_{M,t_1}), \text{ where } \bar{p}_{m,t_1} = \frac{1}{N_{t_1}} \sum_j p_{m,j,t_1}, \quad (1)$$

$$\boldsymbol{C_{t_2}} = (\bar{p}_{1,t_2}, \dots, \bar{p}_{m,t_2}, \dots, \bar{p}_{M,t_2}), \text{ where } \bar{p}_{m,t_2} = \frac{1}{N_{t_2}} \sum_j p_{m,j,t_2}. \quad (2)$$

We can then characterize the technology trajectory through the topic space between years t_1 and t_2 as the change in the position of the field centroid from year to year represented by the vector $\overrightarrow{C_{t_1} C_{t_2}}$.

3.2. The novelty and relevance metrics

Next, we study the position of a single patent relative to the field centroids before and after it is granted and develop its "novelty" and "relevance" measures.

The novelty of a patent $p_{i,t}$ ($t_1 < t < t_2$) is its shortest Euclidean distance to the vector connecting prior and future field centroids $\overrightarrow{C_{t_1} C_{t_2}}$. The relevance of the patent is the length of projection from the vector $\overrightarrow{C_{t_1} p_{i,t}}$ onto the field centroid vector $\overrightarrow{C_{t_1} C_{t_2}}$. Figure 1 depicts novelty and relevance in a schematic representation of a three-topic space.

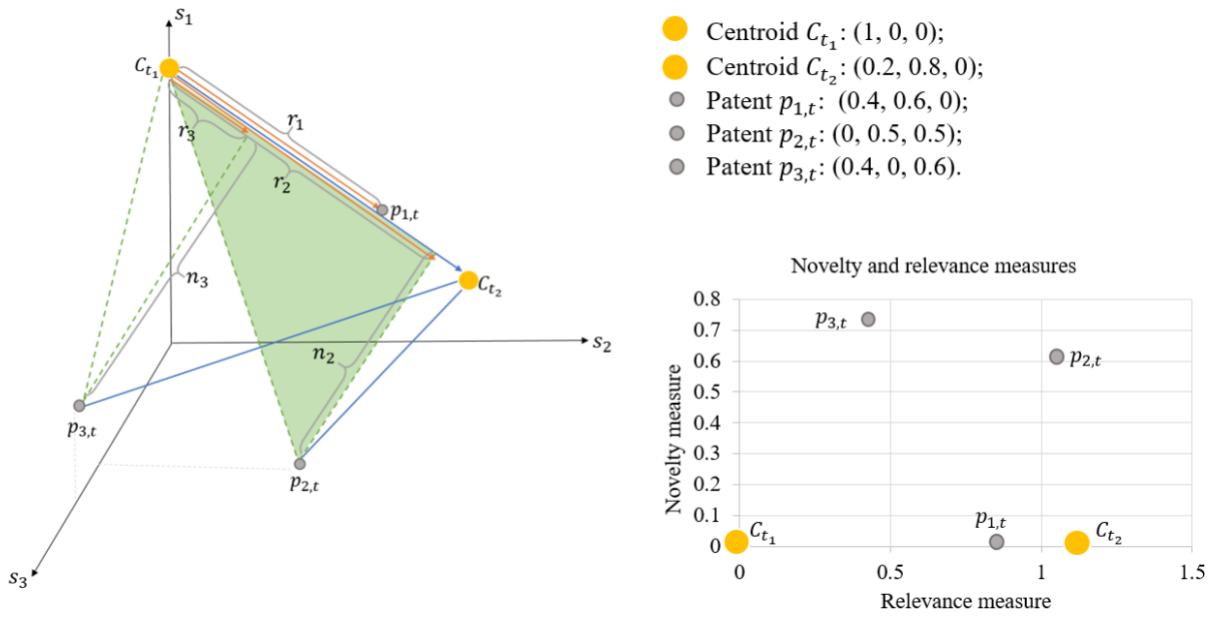


Figure 1. Example of novelty and relevance calculation for patents $p_{1,t}$, $p_{2,t}$ and $p_{3,t}$ in a three-dimensional technological space. Axes s_1, s_2, s_3 represent coordinate topic vectors. For simplicity, all three patents displayed are combinations of only two topics. Line segments n_2 and n_3 represent novelty of patents $p_{2,t}$ and $p_{3,t}$ (for $p_{1,t}$ novelty $n_1 = 0$), while r_1, r_2 and r_3 are the relevance of corresponding patents $p_{1,t}$, $p_{2,t}$ and $p_{3,t}$. The scatter plot shows the novelty and relevance values for the three patents.

Vector $\overrightarrow{C_{t_1}C_{t_2}}$ reflects both the direction and magnitude of change in the mainstream technological trajectory from year t_1 to t_2, as measured by the change in the position of the field centroid of these two years and the Euclidean distance between C_{t_1} and C_{t_2}. Novelty and relevance measures of patent $p_{i,t}$ relative to C_{t_1} and C_{t_2} are then defined as

$$\text{Novelty } n_{i,t}^{t_1,t_2} = \sin(\theta) \cdot \left|\overrightarrow{C_{t_1}p_{i,t}}\right|, \qquad (3)$$

$$\text{Relevance } r_{i,t}^{t_1,t_2} = \cos(\theta) \cdot \left|\overrightarrow{C_{t_1}p_{i,t}}\right|, \qquad (4)$$

where the superscripts t_1 and t_2 denote the two reference years, the subscript t denotes the application year of patent $p_{i,t}$ ($t_1 < t < t_2$), and θ is the angle between $\overrightarrow{C_{t_1}C_{t_2}}$ and $\overrightarrow{C_{t_1}p_{i,t}}$ and is between 0 and π.

Essentially, novelty measures how distant the knowledge in patent $p_{i,t}$ is to the main direction of technological evolution from C_{t_1} to C_{t_2}. Relevance reflects the amount of knowledge in $p_{i,t}$ that aligns with the direction of technological change from C_{t_1} to C_{t_2}.

Figure 1 gives an illustrative example with three patents to show how novelty and relevance measures are calculated in a three-topic space (s_1, s_2, s_3). The coordinates in the graph on the left side of the figure represent topic probabilities. Centroid C_{t_1}'s coordinates are (1,0,0), indicating that the mainstream technology in the year t_1 contains only topic 1. Centroid C_{t_2}'s coordinates are (0.2, 0.8, 0), showing that it is a mixture of content from topic 1 and 2 but mostly topic 2. Similarly, patents $p_{1,t}$, $p_{2,t}$ and $p_{3,t}$ are mixtures of topics 1-2, 2-3, and 1-3, respectively. The bottom right side of the figure plots the novelty and relevance measures of $p_{1,t}$, $p_{2,t}$ and $p_{3,t}$ calculated using formulas (3) and (4). The Euclidean distance between the centroids C_{t_1} and C_{t_2} is plotted on the x-axis as a reference. Patent $p_{1,t}$ has a novelty measure of 0 and a relatively high relevance measure because it is a combination of topics 1 and 2, similar to the centroid C_{t_2}, lying exactly on the centroid vector $\overrightarrow{C_{t_1}C_{t_2}}$. Patent $p_{2,t}$ has relatively high novelty and relevance measures because it contains topic 3, which is external to $\overrightarrow{C_{t_1}C_{t_2}}$, and topic 2, which might contribute to the ability to integrate topic 2's content in centroid C_{t_2}. Patent $p_{3,t}$ has the highest novelty value but relatively low relevance value, as it contains a high fraction of the external but irrelevant content from topic 3.

Equations (3) and (4) show that the value of novelty and relevance both depend on θ, the angle between $\overrightarrow{C_{t_1}C_{t_2}}$ and $\overrightarrow{C_{t_1}p_{i,t}}$. Figure 2 shows the range of novelty $n_{i,t}$ and relevance $r_{i,t}$ values as a function of the angle between the field centroid and the focal patent, θ. If θ is between 0 and π/2, both novelty and relevance are positive. If θ is between π/2 and π, novelty is positive while relevance is negative. Negative relevance means that the projection of the vector $|\overrightarrow{C_{t_1}p_{i,t}}|$ on vector $\overrightarrow{C_{t_1}C_{t_2}}$ is in the opposite direction of the $\overrightarrow{C_{t_1}C_{t_2}}$, indicating that there is little knowledge in $p_{i,t}$ that aligns with the pathway of technological evolution from C_{t_1} to C_{t_2}.

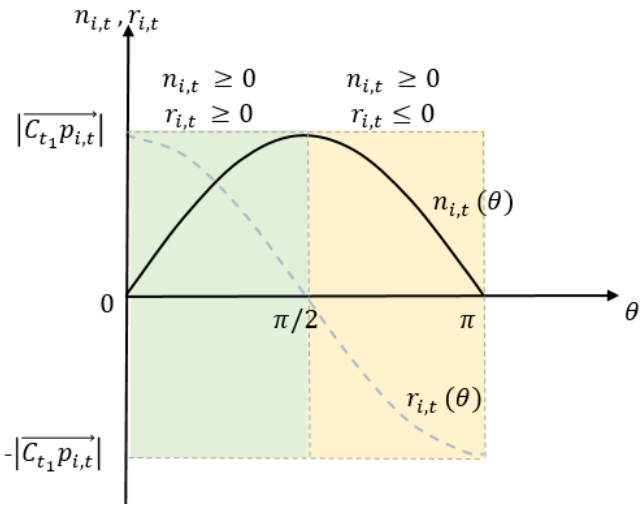


Figure 2. Novelty (solid line) and relevance (dashed line) as a function of the angle between the previous field centroid and the focal patent, θ.

Other studies have developed technological distance metrics to measure the magnitude of difference in knowledge bases between two patents (e.g. Jaffe, 1986; vom Stein, Sick and Leker, 2015; Aharonson and Schilling, 2016). These pairwise distances do not fully describe the positions of patents in a multi-dimensional technological space, let alone the impact of inventions on technological change. Our novelty and relevance metrics use vector decomposition of technological distances in multi-dimensional space to provide new information about the movement of the field relative to the content of a single patent.

Figure 3 illustrates how the same technological distance in different directions can have different novelty and relevance values. Patents $p_{1,t}$, $p_{2,t}$ and $p_{3,t}$ have the same technological distance to

the prior field centroid C_{t_1}. Because these three patents hold different positions in the technological space relative to the evolving field centroid vector, their vector decompositions yield different novelty and relevance values. As the angle between the vectors $\overrightarrow{C_{t_1}C_{t_2}}$ and $\overrightarrow{C_{t_1}p_{i,t}}$ gets smaller (from $p_{1,t}$ to $p_{2,t}$ to $p_{3,t}$), the perpendicular distance from $p_{i,t}$ to the centroid vectors (i.e. novelty) gets smaller and the projection of the vector $\overrightarrow{C_{t_1}p_{i,t}}$ onto the centroid vector's direction (i.e. relevance) gets larger. This gives us additional information on $p_{i,t}$'s potential influence on the magnitude and direction of mainstream technological change compared to a single technological distance measure.

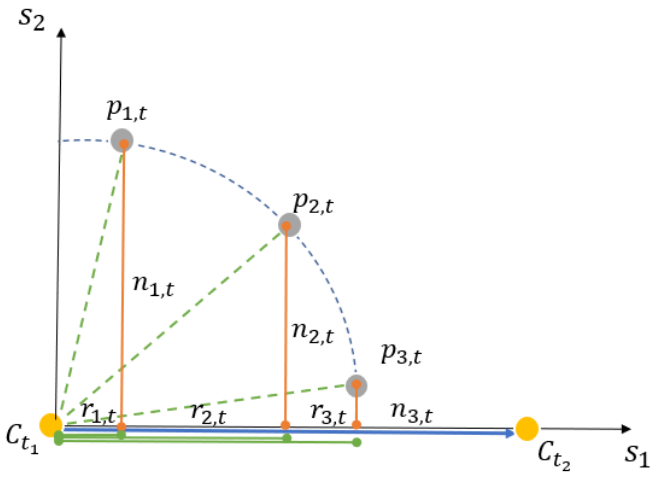


Figure 3. Vector decomposition of technological distance, novelty and relevance metrics in a two-topic space (s_1 and s_2) for three patents ($p_{1,t}$, $p_{2,t}$ and $p_{3,t}$). The blue dashed line represents the equidistant curve to the prior field centroid, C_{t_1}. Orange lines represent novelty values and green lines represent relevance values.

3.3. Identifying potential technological breakthroughs

Following theories of innovation dynamics in which technological breakthroughs (supported by incremental innovation) drive the evolution of a technological field (Mokyr, 1990), our methodology seeks to identify potential breakthrough inventions that may offer explanatory power for how the field evolves in the technological space. Our approach is focused on developing *candidate* breakthrough inventions, as we do not infer causality, other than noting the

time consistency with which the candidate breakthrough occurs before the subsequent technological change in the direction of the breakthrough; section 5.2 discusses potential inferential errors related to this issue. The power of the methodology described in this section is that it formalizes the abstract notion of a "technological space" through the topic model and proposes dynamic geometric interpretations in this space for the key theoretical concepts discussed in Section 2 that define technological breakthroughs.

When a patent representing a breakthrough invention is granted, its knowledge is relatively new to the mainstream of the technology, hence its initial novelty value should be relatively high, and its initial relevance value should be relatively low. As new ideas disseminate, more patents build upon the initially novel patent, and the mainstream technological trajectory represented by a sequence of field centroids should start to shift closer towards the location of the patent in the technological space, decreasing novelty and increasing relevance values of the patent. Changes in the novelty measure reflect the changing distance between the evolving technological trajectory and the breakthrough patent, while changes in the relevance measure indicate how closely the mainstream evolution aligns with the breakthrough patent over time.

To operationalize these conditions, we propose the following steps to identify candidates for patents representing technological breakthroughs in the focal field represented in the technological space by a pre-developed topic model:

1. Specify the set of patents granted in year *t* to be evaluated: $(p_{1,t}, p_{2,t}, \dots, p_{N_t,t})$, where N_t is the total number of patents in the focal domain granted in year *t*. Each patent $p_{i,t}$ has a topic distribution vector $\boldsymbol{p_{i,t}} = (p_{1,i,t}, \dots, p_{m,i,t}, \dots, p_{M,i,t})$, where *M* is the total number of topics.

2. Calculate the centroid of all patents in the focal domain granted in year (*t-1*) as the beginning field centroid (point C_{t_1} in Figure 1): $\overline{\boldsymbol{C}}_{t-1} = (\bar{p}_{1,t-1}, \bar{p}_{2,t-1}, \dots, \bar{p}_{M,t-1})$, where $\bar{p}_{m,t-1} = \frac{1}{N_{t-1}} \sum_i p_{m,i,t-1}$, which is the average probability of topic *m* among all the patents granted in year (*t-1*) in the focal domain.

3. Calculate the field centroids of all the patents granted in the same field in subsequent years since year *t*: C_{t+1}, C_{t+2}, ..., C_{t+T}, where *T* is the window of measurement. These

field centroids in subsequent years serve as the series of ending points representing the evolution of the centroid vector (point C_{t_2} in Figure 1).

4. Apply equations (3) and (4) to calculate novelty and relevance measures for patent $p_{i,t}$ relative to a series of consecutive field centroid vectors with the same starting point C_{t-1} but different ending points ($\overrightarrow{C_{t-1}C_{t+1}}$, $\overrightarrow{C_{t-1}C_{t+2}}$, …, $\overrightarrow{C_{t-1}C_{t+T}}$): $\boldsymbol{n_{i,t}} = (n_{i,t,1}, n_{i,t,2}, \dots, n_{i,t,T})$ and $\boldsymbol{r_{i,t}} = (r_{i,t,1}, r_{i,t,2}, \dots, r_{i,t,T})$.[2]

5. Use linear or second-order polynomial regression to calculate the trends in $\boldsymbol{n_{i,t}}$ and $\boldsymbol{r_{i,t}}$ over the period $(t+1, \dots, t+T)$ for each patent $p_{i,t}$.

6. Select patents with statistically significant negative trends in novelty and statistically significant positive trends in relevance over the window of measurement *T* as potential breakthrough candidates. Additional criteria may be added, such as restricting the length of the period during which the trends in novelty / relevance should remain negative / positive and significant.

Figure 4 presents three possible scenarios of a patent's novelty and relevance behavior relative to the mainstream technological trajectory after steps 1 to 4 are completed: (1) novelty and relevance stay relatively constant as the mainstream evolves (not an integrated patent); (2) novelty increases while relevance declines (a divergent patent); (3) novelty declines and relevance increases (an integrated patent). Other possible combinations are presented in Table 1.

Let $\theta_{i,t,\epsilon}$ denote the angle between $\overrightarrow{C_{t-1}p_{i,t}}$ and $\overrightarrow{C_{t-1}C_{t+\epsilon}}$ $(\epsilon \in [1,T])$.[3] All three scenarios share the same positions of the prior and first-post year field centroids relative to the focal patent. In scenario 1, subsequent centroid vectors are parallel with the initial centroid vector $\overrightarrow{C_{t-1}C_{t+1}}$ $(\theta_{i,t,1} = \theta_{i,t,2} = \theta_{i,t,3})$, suggesting that the mainstream technology in the field evolves in the same direction since year *t+1*. Hence, the level of novelty and relevance remains at the initial

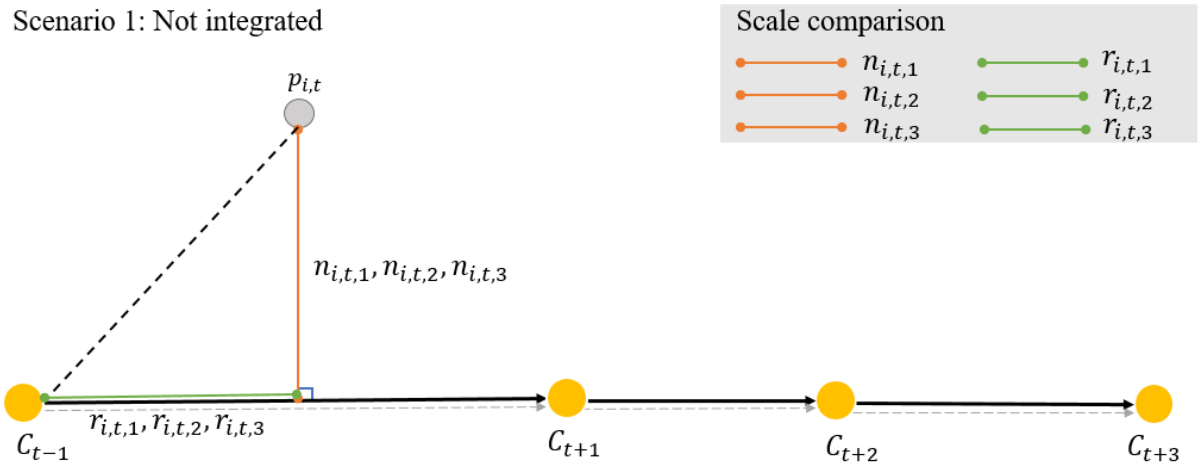
[2] We calculate novelty and relevance measures relative to the starting point of the trajectory C_{t-1}, the year before the patent is granted, rather than any sequential segment $\overrightarrow{C_kC_{k+1}}$ of the technological trajectory itself, because we are interested in the *direction and magnitude of change* on the technological trajectory relative to the starting point rather than the absolute position of the field centroid trajectory in the topic space. We also use traditional "static" version of LDA topic modeling algorithm rather than dynamic versions of it, because we have to measure novelty and relevance over time relative to the same static starting point.

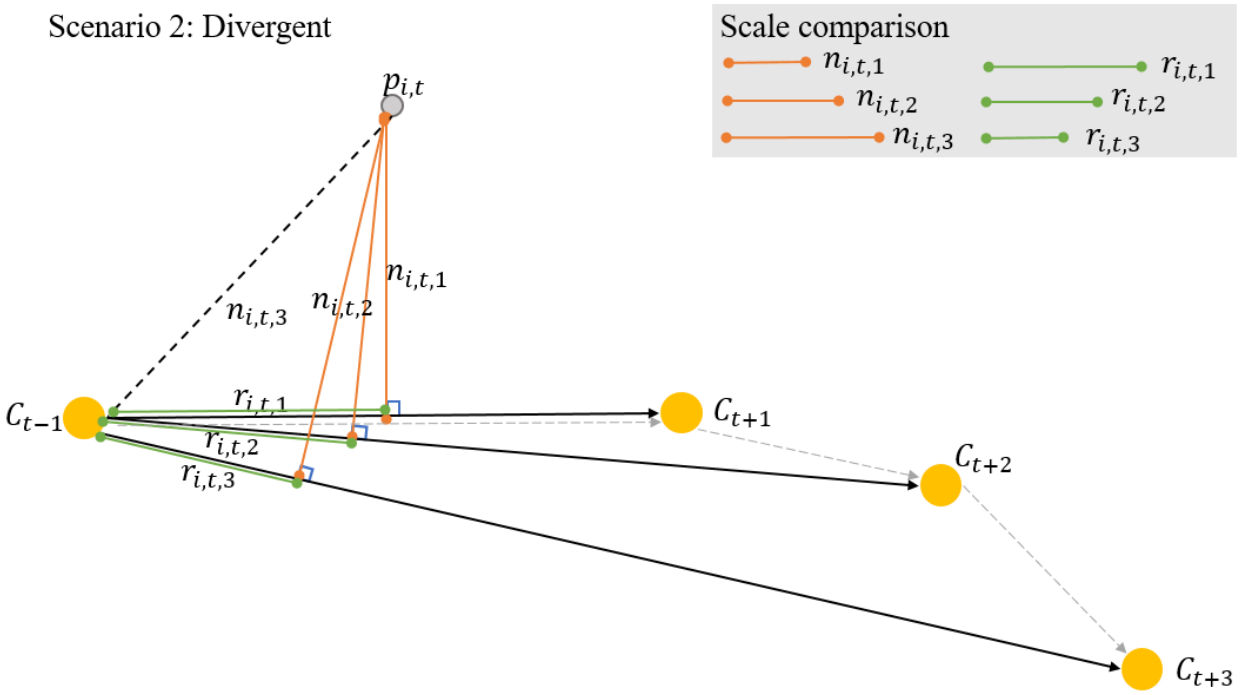
[3] $\theta_{i,t,1}, \theta_{i,t,2}, \theta_{i,t,3}$ are not indicated in Figure 4 due to the lack of space.

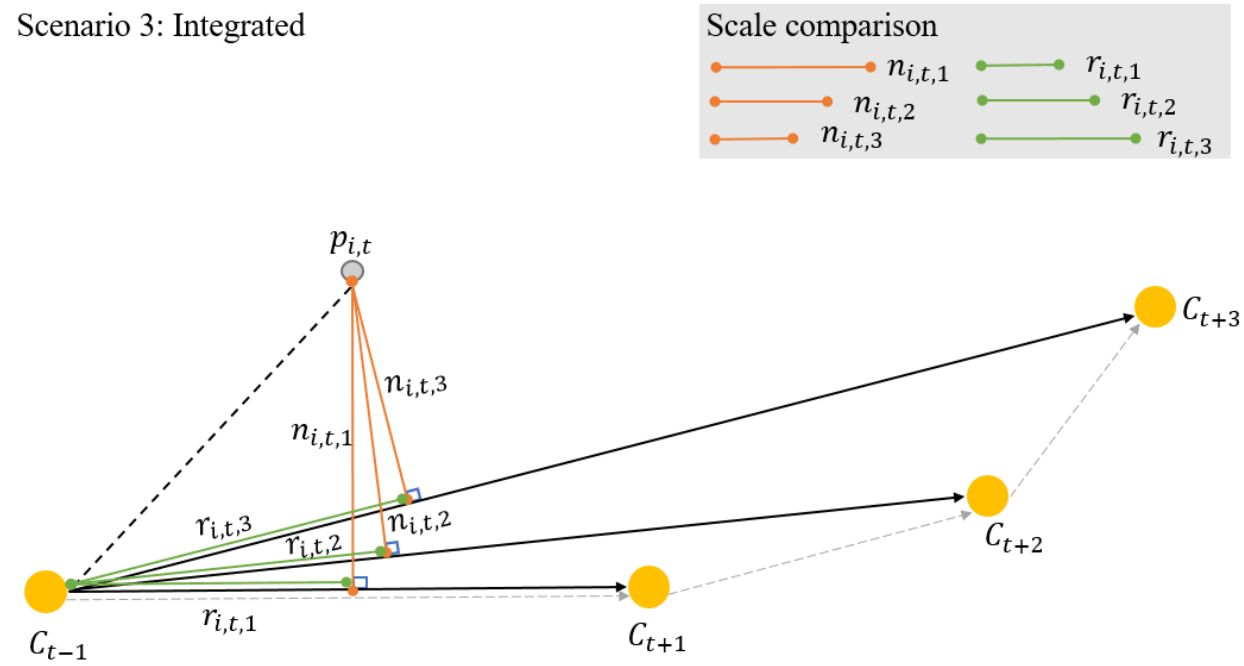
level and this technology is not integrated into the mainstream (i.e. $n_{i,t,1} = n_{i,t,2} = n_{i,t,3}$; $r_{i,t,1} = r_{i,t,2} = r_{i,t,3}$).

In scenario 2, the centroids $(C_{t+1}, C_{t+2}, C_{t+3})$ are getting further away from patent $p_{i,t}$ with each year, which implies $\theta_{i,t,1} < \theta_{i,t,2} < \theta_{i,t,3}$. In turn, the increasing angles suggest that the subsequent technological trajectory gradually moves away from $p_{i,t}$, so that $p_{i,t}$ becomes less relevant and more novel relative to the mainstream over time. In this scenario, the mainstream field centroid trajectory diverges from the focal patent (i.e. $n_{i,t,1} < n_{i,t,2} < n_{i,t,3}$; $r_{i,t,1} > r_{i,t,2} > r_{i,t,3}$).

Finally, scenario 3 depicts a successfully integrated breakthrough patent. In this case, the field centroids $(C_{t+1}, C_{t+2}, C_{t+3})$ move closer to the patent p_i over time, which means $\theta_{i,t,1} > \theta_{i,t,2} > \theta_{i,t,3}$, and the mainstream trajectory appears to gradually incorporate the invention represented in $p_{i,t}$ over time, leading to decreasing novelty and increasing relevance (i.e. $n_{i,t,1} > n_{i,t,2} > n_{i,t,3}$; $r_{i,t,1} < r_{i,t,2} < r_{i,t,3}$). This scenario is of main interest to us as it identifies patents potentially representing technological breakthroughs.







Notes:

1. Solid arrows represent centroid vectors and dashed arrows indicate the direction of centroid movement.

2. Figures are simplified two-dimensional representations of multi-dimensional topic space.

Figure 4. Three scenarios of novelty and relevance behavior relative to the field centroid trajectory over time. In all three scenarios, the focal patent ($p_{i,t}$) has the same initial position relative to the centroids of the prior year (C_{t-1}) and the first subsequent year (C_{t+1}).

According to Figure 2, novelty and relevance change in opposite directions with θ when θ is between 0 and π/2. A successfully integrating breakthrough patent is one in which novelty decreases over time and relevance increases over time, while for a diverging patent novelty increases and relevance decreases over time. Further two cases shown in Table 1 are possible when θ is between π/2 and π and novelty and relevance change in the same direction. In the case where both novelty and relevance values increase over time ($\theta_{i,t,\epsilon} \in (\pi/2, \pi]$), the invention, despite being radical or experimental, is unlikely to have been incorporated into the main technological field; its novelty value increases over time and the relevance value is negative, showing the irrelevance of the invention to the mainstream of the technology. In the case where both novelty and relevance values decrease ($\theta_{i,t,\epsilon} \in (\pi/2, \pi]$), the novel component of the invention may have been adopted by the mainstream over time as other components of the invention became more obsolete, suggesting that it is unlikely to have contributed significantly to the overall evolution of the field.

Table 1. Novelty and relevance trends and corresponding types of inventions

		Novelty		
		Increasing	**Decreasing**	**Constant**
Relevance	**Increasing**	Irrelevant $\theta_{i,t,\epsilon} \in (\pi/2, \pi]$	Integrating $\theta_{i,t,\epsilon} \in [0, \pi/2]$ (Scenario 3)	—
	Decreasing	Divergent $\theta_{i,t,\epsilon} \in [0, \pi/2]$ (Scenario 2)	Irrelevant $\theta_{i,t,\epsilon} \in (\pi/2, \pi]$	—
	Constant	—	—	Not integrated (Scenario 1)

Notation: impossible cases (—).

Figure 4 and Table 1 illustrate "ideal" scenarios, where field centroid trajectories are smooth lines. In practice, the field centroid trajectory is often noisy, as it can be influenced by local sub-technology trends or various sources of noise in patent data. As a result, it is possible that a patent's novelty and relevance trends may change direction at a certain moment in time, or there can even be no clear novelty and relevance trend at all. We exclude the latter cases from our consideration in Steps 5 and 6 by selecting the patents with statistically significant integrating trends as candidate breakthrough patents. Section 4.3 further discusses how the trends in novelty and relevance can be considered over different periods of time and with non-linear changes.

4. Application of patent novelty and relevance framework in silicon solar photovoltaics

In this section, we demonstrate how our framework and operationalization of patent novelty and relevance can be used to identify candidates for technological breakthroughs in a particular focal field: silicon-based solar photovoltaics (PV). This is a technology field where other innovation researchers have conducted significant work with alternative methods, thus allowing for some validation of our approach. Solar PV also has significant potential for public benefit due to the role it is anticipated to play in the low-carbon energy transition and climate change mitigation.

4.1. Data sources and processing

We demonstrate our framework using full-text patent data from the USPTO PatentsView database of patents filed in the United States. Solar PV-related patents are compiled using the Cooperative Patent Classification (CPC), which uses the "Y02" class tags to identify climate change-mitigation technologies related to energy generation, transmission, or distribution (Haščič and Migotto, 2015). Within the "Y02E" class, several subgroups under "Y02E 10/5" are related to the silicon-based subset of PV technologies. To construct the corresponding focal field, we extract data for patents granted between 1976 (the first year digitized full text of patents is available in the USPTO PatentsView database) and 2016 using the CPC subgroups "Y02E 10/545" (Microcrystalline silicon PV cells), "Y02E 10/546" (Polycrystalline silicon PV cells), "Y02E 10/547" (Monocrystalline silicon PV cells) and "Y02E 10/548" (Amorphous silicon PV

cells)[4]. Further, we focus on granted patents, hence excluding ungranted patent applications, as well as other types of intellectual property, such as utility models and design patents. The resulting dataset contains 3,126 patents. We use the full texts of these patents as the corpus for constructing the topic model. The analysis of novelty and relevance then focuses on characterizing the sample of 765 patents in this dataset granted between 1977 and 1996, after which 20 years of technological evolution in the field can be observed.

The process of constructing the topic model, including corpus pre-processing and hyperparameter selection, is described in Appendix A. Our novelty and relevance metrics can be sensitive to the choice of the number of topics in the model because of the inherent instability of LDA topic modeling algorithms (Rieger, Rahnenführer and Jentsch, 2020). In order to reduce potential bias introduced by using a single number of topics, we construct topic models with 175, 200 and 225 topics, which are selected by the LDA tuning process (see Appendix A for details). As an outcome of the analysis, we have vectors of novelty and relevance in three topic models for each of the 765 patents in the sample.

4.2. Breakthrough crystalline silicon PV patents from Nemet and Husmann (2012)

As the first step to demonstrate the validity of our empirical results, we explore the novelty and relevance metrics and integration patterns for a small subsample of 10 patents. These patents were previously identified as representative of important technological breakthroughs for crystalline silicon PV (c-Si PV) by Nemet and Husmann (2012), who compiled a list of 79 breakthroughs from 1951 to 2000 using literature on the history of c-Si PV technology. Using a citation-based method developed by Dahlin and Behrens (2005), they also identified 181 potential breakthrough patents. By matching these patents to the first list of 79 breakthroughs, they were able to identify 39 patents associated with 23 breakthroughs. Among these 39 patents, 10 patents related to eight breakthroughs are in our sample. We analyze the novelty and relevance of these 10 patents to test how our method compares to the citation-based approaches for the identification of technological breakthroughs.

[4] There are other solar PV patents in the broader Y02E 10/5 subgroup that are related to silicon PV technologies but are not included in the CPC subgroups described above. For the purpose of illustrative analysis presented in this paper, these patents are not included in the sample.

Table 2 presents the patent numbers, granted years, and titles of these 10 patents, as well as the count of all forward citations received by them from USPTO patents (*Cit*)[5]. Table 2 also lists the metrics we calculate for these patents' initial novelty values, initial novelty percentiles among all the patents in the sample granted in the same year, linear regression coefficients *N* and *R* (for novelty and relevance trends, respectively) in the 20 years after its granted year, *p* values for corresponding linear regressions that demonstrate the statistical significance of the regression coefficients, and scatter plots of their novelty and relevance values over this period. The novelty and relevance metrics shown in Table 2 are calculated using the topic model with 200 topics.

All patents, except for one on metallization (US4348546), display general integrating trends with a negative novelty slope *N* and a positive relevance slope *R*, although their integrating patterns (shown in their novelty and relevance plots) are often noisy, resulting in only 5 out of 9 integrating *N* and *R* patterns being statistically significant. The initial novelty percentile of these patents varies significantly; some breakthrough patents are not very novel compared to other c-Si PV patents within our dataset granted in the same year (e.g. 0.12 for US4086102, 0.10 for US4322253), while others are in relatively high novelty percentiles among their peers (e.g. 0.70 for US4171997, 0.77 for US4643913). The relatively low initial novelty of some breakthrough patents may be caused by the inclusion of technological sub-trajectories in our patent corpus. Apart from c-Si PV, our dataset also includes patents related to amorphous silicon (a-Si) PV and thin-film PV technologies, which became a target of active patenting later than c-Si PV. The dynamics of these sub-fields within our definition of silicon solar PV could cause the c-Si PV patents to appear less novel. Through this example, we see that setting a threshold on initial novelty in the process of screening for breakthrough patents would not necessarily yield the most impactful innovations.

The novelty and relevance plots in Table 2 also indicate that results are sensitive to the sample size. In several cases (e.g., US4643913, and the non-integrating patent US4348546) overall slopes of novelty and relevance are heavily influenced by early outliers. The size of annual subsamples in early years is roughly 20 patents, such that even a single closely related patent (e.g., a patent on a device when the original patent was on the method of its preparation) can skew the novelty downward and affect the overall slope of novelty and relevance trends. Pre-

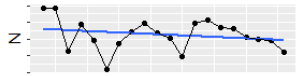
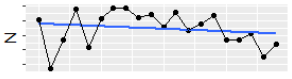
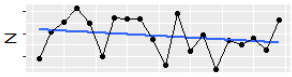
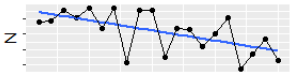
[5] The forward citation count is based on the PatentsView datasets downloaded in October 2019.

processing of the patent corpus to remove duplicate patents or analyzing moving averages of novelty and relevance could mitigate the impact of noise and outliers.

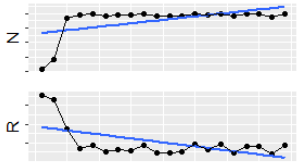
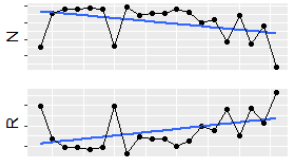
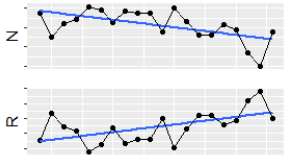
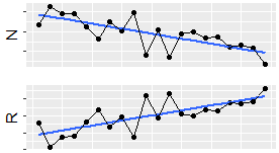
When the same novelty and relevance calculations are performed using the topic models with 175 or 225 topics, only 6 out of 10 patents display integrating patterns, compared to 9 in the case of the 200-topic model. Clearly, the integrating patterns are sensitive to the choice of the LDA topic model. We conduct further sensitivity tests, compare the results from applying three different topic models to the full patent sample, and suggest an operational approach to constructing the list of breakthrough candidate patents that takes into account the issue of topic model sensitivity in section 4.4.

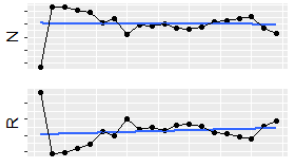
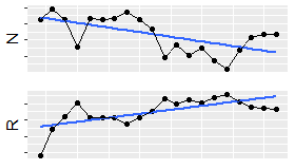
Overall, our first observations suggest that our novelty and relevance measures can capture the dynamics of patent integration associated with breakthrough inventions into the technology mainstream of the focal field quite well. We can also see that technological breakthroughs do not necessarily emerge from the most novel patents; it is the subsequent decrease in novelty and increase in relevance that turns an initial invention into a technological breakthrough.

Table 2. Selected breakthrough patents from Nemet and Husmann (2012), number of topics = 200[6]

Breakthrough patents	Grant year	Initial novelty	Initial novelty percentile	*Cit*	Patent title	Novelty and relevance plots since patent granting	Estimated slope of novelty *N* and relevance *R*	*p* values of *N* and *R* regressions
(1) Oxide surface passivation								
US4086102	1978	23.15	0.12	41	Inexpensive solar cell and method therefor		*N*: -0.014 / yr	0.345
							R: +0.088 / yr	0.158
US4171997	1979	37.04	0.70	14	Method of producing polycrystalline silicon components, particularly solar elements		*N*: -0.018 / yr	0.447
							R: +0.078 / yr	0.352
(2) Hydrogen-plasma passivation								
US4322253	1980	26.34	0.10	98	Method of making selective crystalline silicon regions containing entrapped hydrogen by laser treatment		*N*: -0.006 / yr	0.346
							R: +0.032 / yr	0.313
US4557037	1985	35.64	0.55	30	Method of fabricating solar cells		*N*: -0.007 / yr	0.005***
							R: +0.134 / yr	0.005***

[6] Novelty and relevance values in this table are scaled by 100 for operational convenience.

Breakthrough patents	Grant year	Initial novelty	Initial novelty percentile	*Cit*	Patent title	Novelty and relevance plots since patent granting	Estimated slope of novelty *N* and relevance *R*	*p* values of *N* and *R* regressions
(3) Metallization								
US4348546	1982	83.28	0.97	42	Front surface metallization and encapsulation of solar cells		*N*: +0.024 / yr *R*: -0.429 / yr	0.015** 0.008***
(4) Reduced metallization resistance								
US4395583	1983	35.56	0.60	7	Optimized back contact for solar cells		*N*: -0.021 / yr *R*: +0.161 / yr	0.085* 0.055*
(5) MINP Cell								
US4404422	1983	24.76	0.17	66	High efficiency solar cell structure		*N*: -0.016 / yr *R*: +0.102 / yr	0.007*** 0.008***
(6) Plasma deposition of SiN passivation								
US4640001	1987	28.32	0.31	17	Solar cell manufacturing method		*N*: -0.101 / yr *R*: +0.238 / yr	0.0002*** 0.0001***

Breakthrough patents	Grant year	Initial novelty	Initial novelty percentile	*Cit*	Patent title	Novelty and relevance plots since patent granting	Estimated slope of novelty *N* and relevance *R*	*p* values of *N* and *R* regressions
(7) Low-contact resistance with AR films								
US4643913	1987	46.84	0.77	18	Process for producing solar cells		*N*: -0.002 / yr *R*: +0.052 / yr	0.904 0.535
(8) PESC cell								
US4589191	1986	26.75	0.14	76	Manufacture of high efficiency solar cells		*N*: -0.023 / yr *R*: +0.245 / yr	0.004*** 0.001***

Note: * indicates statistical significance of novelty *N* and relevance *R* linear regression coefficients at p < 0.1 level, ** at p < 0.05, *** at p < 0.01.

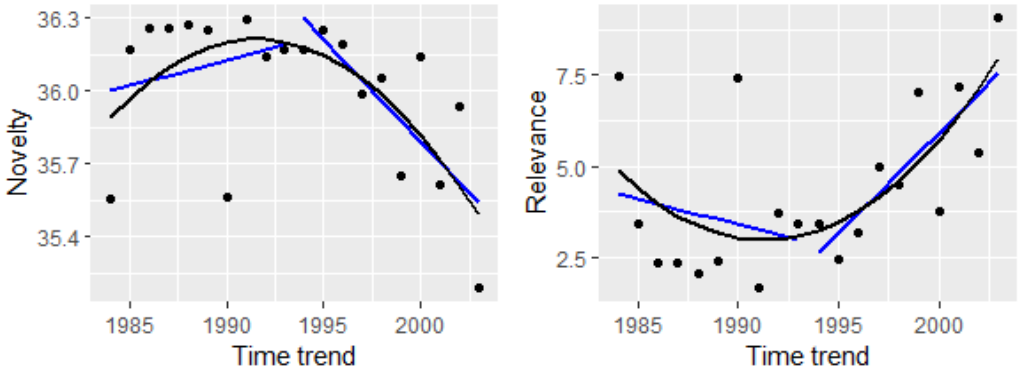
4.3. Time of integration

One useful insight from the novelty and relevance plots in Table 2 is that patents may integrate into the mainstream of the focal field at different rates or after a significant delay. Linear regressions over 20 years may not capture these scenarios; patent US4171997, for example, has a clearly integrating pattern in the second half of the 20-year window, despite linear *N* and *R* slopes being statistically insignificant. This is a common phenomenon in our patent data, consistent with a delay between an invention and the response to it by other inventors.

In Panel A in Figure 5, novelty and relevance trends of patent US4395583 (one of the breakthrough patents in Table 2) are shown with piecewise linear regressions instead of a single linear regression over the 20-year window. In the first 10 years, the calculated novelty remains relatively constant; novelty then decreases in the second 10-year period as the patent is integrated into the mainstream. Hence, a longer evaluation period can help identify patents with delayed integration.

For the patents that are faster to integrate, however, a longer time horizon might result in a flatter estimated slope after an initial substantial change. Moreover, if a breakthrough patent is rapidly integrated into the mainstream, its novel content may become so mainstream—or even obsolete—that it is no longer explicitly mentioned in the text of future patents that build on this breakthrough. Such patents might have an increase in novelty and a decline in relevance as the topics they describe no longer appear in subsequent patent texts. Panel B in Figure 5 demonstrates an example of this pattern for patent US4843451. Here, it appears that the trends of novelty and relevance measures reversed 15 years after the patent was granted.

Panel A. Patent US4395583 (1983): Optimized back contact for solar cells



Panel B. Patent US4843451 (1989): Photovoltaic device with O and N doping

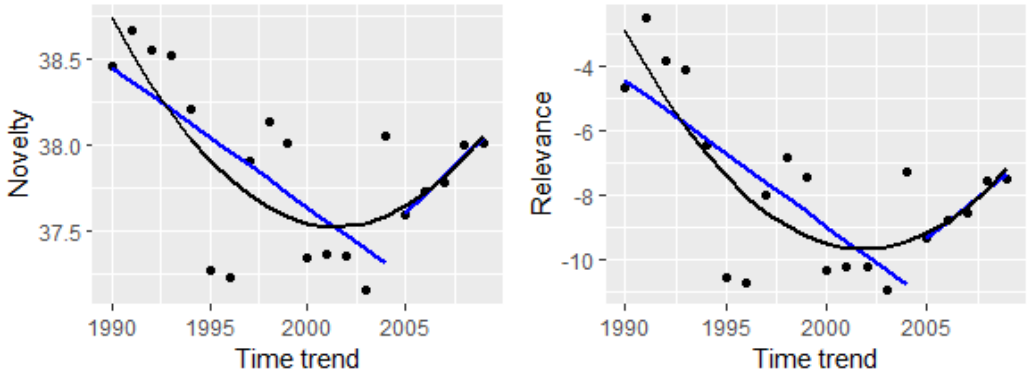


Figure 5. Illustrative examples of non-linear patent integration patterns. The 200-topic model is used for novelty and relevance calculations.[7]

The non-linear trends demonstrated with the patents in Figure 5 suggest that it may be important to graphically plot the novelty and relevance metrics in borderline cases to examine specific technology integration patterns. In addition to visualization, one simple approach is to compare the slopes of linear models constructed over different periods, such as 10 years vs. 20 years after the patent is granted. For example, if the slope in the shorter term is higher than the slope in the longer term, then the patent is likely to embody fast-integrating technologies.

A more sophisticated method to endogenously identify the "turning points" of the novelty and relevance trends is to fit a second-order polynomial function to the novelty and relevance values instead of a line. If the second-order polynomial coefficient is statistically significant, the trend is likely to change direction at some point in time. Local maxima or minima, as well as the timing of the trend change, can be calculated for additional insight into any potential delay or non-linear pace of integration. In the case of Panel A in Figure 5, the fitted polynomial functions for both novelty and relevance values have statistically significant second-order term coefficients (– 0.005, p = 0.009 for novelty; 0.036, p = 0.013 for relevance). Using the estimated polynomial coefficients, we find that the novelty and relevance trends changed direction approximately 8-9 years after the patent was granted. Similarly, in Panel B in Figure 5, the second-order terms are statistically significant for both novelty and relevance measures (0.009, p = 0.002 for novelty; 0.049, p = 0.0008 for relevance), and the reversal in trends happened around 12-13 years after

[7] Novelty and relevance values are scaled by 100 for operational convenience.

the patent was granted. Thus, polynomial trend estimates can account for more subtle patterns of integration than linear estimates alone.

4.4 Screening for breakthrough patents in crystalline silicon solar photovoltaics

Next, we apply our framework to a subsample of c-Si PV patents to create a full list of potential breakthrough patents in the c-Si PV subfield.

Starting from the initial sample covering all CPC subclasses relevant to silicon-based PV technology, we restrict the subsample to only three out of four initial CPC subclasses that explicitly mentioned crystalline silicon (Y02E 10/545, Y02E 10/546, Y02E 10/547), and then manually removed irrelevant misclassified patents. The size of the resulting subsample was reduced to 376 patents out of an initial 765.

Based on the observations in Section 4.2, we recognize that the integration pattern of a patent can be sensitive to outliers in centroid calculation given the small annual subsamples in the early years. Therefore, in addition to annual centroids, we also calculate three-year moving average centroids. Specifically, the field centroid in year t is the average topic distribution of all patents granted in years t, (t-1), and (t-2). Since our sample starts in 1976, the field centroid for 1977 is the average topic distribution for patents granted in 1976 and 1977, and the field centroid for 1976 is the average topic distribution in 1976.

Using three topic models with a different number of topics (i.e. 175, 200 and 225), as well as both annual and three-year moving average centroids, we calculate six sets of novelty and relevance metrics for each patent and linear and second-order polynomial fits for each set. Then we include the following patents in the candidate breakthrough patent list:

1. Patents with integrating and significant linear novelty and relevance trends, i.e. their novelty decreases and relevance increases over time.
2. Patents with integrating but non-significant linear novelty and relevance trends that have significant second-order polynomial novelty and relevance trends
3. Patents with non-integrating and non-significant linear novelty and relevance trends that nevertheless have significant second-order polynomial novelty and relevance trends and show continuous integration (decreasing novelty and increasing relevance) for 7 or more years.

Patents not falling under one of these three conditions are discarded. The third condition has been developed as a heuristic based on the analysis of real novelty and relevance patterns with very quick or delayed integration described in section 4.3.

Following this selection approach, we create six sets of potential breakthrough patents. These lists include between 179 and 190 patents (48-51% of the initial c-Si PV subsample of 376 patents). Due to the sensitivity of the results to the topic model specification, as well as the small sample size and various sources of noise in patent data, the pairwise overlaps between these lists include 74% to 84% of each set.

In addition, our methodology cannot distinguish breakthrough inventions that *cause* changes to a technological field from those inventions *caused by* changes to a technological field or those otherwise contemporaneously *associated* with changes to a field. The possible lack of a causal link between integration trends and the pathway of technological evolution precludes deductive patent screening to conclusively identify breakthroughs. Instead, we view our results as providing a candidate patent list that requires further validation (see Section 5.3).

A patent that is present in all variations of the candidate list obtained with various topic model parameters more plausibly represents a technological breakthrough with a successful integration pattern.[8] Each step of our method is easily automated; therefore, analysis of multiple candidate list variations ideally should not cause burdensome extra work for researchers compared to using a single candidate list. After applying this approach to the candidate lists created for six framework specifications (175, 200 or 225 topics; one-year or three-year moving average field centroids), the final "short list" of candidate breakthrough patents present in all six lists contains 98 patents (26% of the initial c-Si PV subsample). It is presented in Appendix B[9].

[8] We investigate the robustness of the candidate patent "short list" by varying the topic model specifications and observing the overlap of identified patent lists. The initial cross-validation we conducted suggested an optimal number of topics between 175-225 to efficiently capture the information content of the solar PV patent corpus (see Appendix A). We repeated our analysis with many topic models with different numbers of topics as well as moving average centroids and compared the overlap of patents identified as breakthrough candidates. We found marginal changes in the number of patents repeatedly selected by the set of topic models after 5-6 models were compared, which stands in contrast to a pattern of asymptotically zero overlap if the topic modeling outcomes were unrelated. Thus, we select 6 topic models as a heuristic for the number of specifications that are needed to compile a robust "short list" of candidate patents in our case.

[9] The "long list" of patents (available as online supplement) included in at least a single list includes 277 patents (74% of c-Si PV subsample).

Among these 98 patents, two patents are included in Table 2 as breakthrough patents identified by Nemet and Husmann (2012). If we loosen the restriction to include all patents in at least two sets of results out of six, we arrive at a list of 238 patents that contains all 10 patents in Table 2. Of our test set of 10 breakthrough patents, the two patents included in the short list in Appendix B arguably represent some of the most important c-Si PV innovations of the 1980s with the clearest impact on the field. Plasma deposition of silicon nitride passivation (US4640001) was invented and commercialized by Kyocera in Japan in the early 1980s and had become by the 1990s a state-of-the-art process in solar PV manufacturing (Green, 2005; Husmann, 2011). The passivated emitter solar cell (PESC, US4589191) was invented at the University of New South Wales (UNSW) in 1984 and had the highest efficiency of all PV cells at the time at 20%. It was not adopted in PV manufacturing directly but led to a development of PERL (passivated emitter, rear locally-diffused) cells, also at UNSW in 1993 (Green, 1995; Husmann, 2011). Thus, the initial breakthrough idea of passivated emitter architecture originated by the PESC cell was successfully integrated into the domain through PERL cells, which eventually entered the market.

For other breakthroughs in our test sample in Table 2, the relationship between initial invention and patent integration might not be as straightforward, as there is often not a one-to-one relation between patents, inventions, and technology. For example, two breakthroughs in Table 2 (oxide surface passivation and hydrogen plasma passivation) are each represented by two patents in the test sample. Each patent may have represented different aspects or variations of the technology at different stages of its development, some of which may have been less relevant to the eventual mainstream than others, resulting in relatively higher sensitivity of integration trends for these patents to the topic model variation, which, in turn, may explain why they are captured not in every framework specification.

5. Discussion

In this paper, we propose a framework for characterizing technological breakthroughs based on the position of inventions in a technological space. Specifically, we develop two metrics, novelty and relevance, that describe the position of a patent relative to a technological field's trajectory in a topic space. We posit how these metrics change over time if an invention represented by a

patent integrates into the technological field and becomes a technological breakthrough. We apply this framework as a retrospective screening tool for identifying potential technological breakthroughs in silicon-based solar photovoltaics patents granted between 1977 and 1996.

One clear advantage of our framework is its flexibility. Since the novelty and relevance metrics are continuous, researchers can create their own heuristic process to define breakthrough criteria based on their search goals. They may also study the specific topics driving changes in aggregate novelty and relevance within a specific technological context. Integration plots such as those presented in Table 2 offer a rich source of insights about the dynamics of innovation in each patent case. In this section, we discuss further applications of the framework beyond its use as a retrospective screening tool, as well as approaches to the validation of results, limitations to their applicability, and avenues for further research.

5.1. Patent citations

Previous studies have used patent citations extensively to measure the influence of inventions on subsequent innovation. Although our empirical approach is partially motivated by the noisy nature of citation data and its limitations, it is still useful to compare how our framework fares against traditional metrics of influence such as the number of forward citations a patent receives. Regression analysis[10] indicates that higher initial novelty and negative novelty slopes are both significantly correlated with higher forward citations; the correlation between negative novelty slopes and forward citations is particularly strong for patents with lower initial novelty compared to the full sample. These results support the basic intuition behind traditional citation metrics: that many influential inventions are also relatively novel at the time of invention. These results also show an important dimension of impactful innovation that simple citation-based metrics often fail to capture: the possibility that even a relatively less novel patent can develop into an impactful innovation by integrating into the mainstream.

Another strength of our approach in comparison with traditional citation metrics is that our approach is able to capture breakthrough inventions associated with relatively fewer cited

[10] Regressions were performed on the patent dataset using novelty and relevance values from the 200-topic model, scaled by 100 for operational convenience. Full regression analysis is presented in Appendix C.

patents, such as US4640001 on plasma deposition of silicon nitride passivation that has just 17 total forward citations in the USPTO patent database as of October 2019, despite its historical significance for PV manufacturing technology (see Table 2). Our breakthrough candidate patent short list in Appendix B contains 21 patents with less than 10 USPTO citations. While inclusion of some of these patents may be indicative of the limitations of forward citations as an indicator of a breakthrough, a lack of forward citations could also indicate the absence of actual adoption of novel ideas in a patent by subsequent inventions, which is necessary for the integration of inventions into the focal technology domain (see discussion of "false positives" in Section 5.3). Therefore, extra attention should be paid to the validation of such breakthrough candidate patents.

5.2. Citation networks

Our novelty and relevance framework defines the technological trajectory of a focal field by the movement of annual field centroids. One alternative approach is called "main path analysis," which identifies the technological trajectory as the main path through a patent citation network (Hummon and Dereian, 1989). This method, which has been used to identify essential innovation trajectories (Verspagen, 2007; Bekkers and Martinelli, 2012; Huang et al., 2015; Huang et al., 2017; Huenteler et al., 2016a, 2016b), assigns weights to edges (i.e. citation relationships) and nodes (i.e. patents) based on their importance in connecting other patents in the network.

Novelty and relevance metrics can be applied to main path analysis; in this case, they are calculated relative to the main patent path instead of a series of field centroids. These metrics can help identify potential breakthrough inventions along the main path and explain how they contribute to the main path moving from one node to the next. Potential breakthrough candidates can be identified as the cited patents that are (1) novel relative to the main path segment *before* they were cited by the main path patent; (2) increasingly relevant and decreasingly novel relative to the main path segment *after* they were cited by the main path patent.

Another way to supplement our method using citation networks is to calculate novelty and relevance for patents that are cited within the focal technological field but are not themselves considered inside the field. This can help identify important channels of external knowledge flow into the technological field from other technologies or sectors. Developing methodological

approaches for the identification of such "technology spillovers" could be an important focus of future work.

5.3. Limitations of methodology and validation of breakthroughs

A limitation of our framework is that it is based on observed associations and does not directly estimate the impact of a patent on subsequent innovation in a field. Establishing the causality of a patent's impact is beyond the scope of most quantitative studies of innovation dynamics, including this one, due to the difficulty in constructing counterfactual scenarios of innovation trajectories, as there are few credible approaches to constructing long-run estimates of alternative paths of innovation had a breakthrough not occurred. Hence, accurate identification of technological breakthroughs with our approach requires exploration of these potential causal relationships with more in-depth process tracing of knowledge integration, utilizing expert knowledge and historical data outside of the patent record.

Table 3 lists possible reasons how our proposed method might yield "false positive" or "false negative" results upon validation, in addition to correctly identified breakthrough patents.

Table 3. Validation scenarios and explanations for identification errors

		Validated breakthroughs	
		False	**True**
Breakthrough candidates	**False**	Correctly identified as a non-breakthrough patent	False negative (missed breakthrough): 1. Measurement errors 2. Not fully integrated or integrated too slow 3. Integrated too fast
	True	False positive (incorrectly identified as a breakthrough): 1. Similar content to an earlier breakthrough 2. Missed opportunities	Correctly identified as a breakthrough patent

False negative patents, i.e., breakthroughs not identified through our method, can occur for at least three reasons. First, measurement errors in the methodology, such as the impact of outliers, imperfect definition of the focal field, or instability of the topic model, may prevent the topic model from fully capturing the integration of the invention represented by each patent. Second, the patent may not be integrated into the mainstream yet; more recent inventions, in particular, may need more time to be understood and adopted by other inventors. Third, patents that are integrated into the mainstream too fast may not show the expected trends of novelty and relevance; from the perspective of our methodology an invention that is very quickly adopted by the mainstream could become instantly highly relevant and not novel.

False positive patents, which are flagged as breakthrough candidates by our method but are not validated as such, can occur because the relationship between a patent and an observed change in a technological trajectory may not be causal. First, the patent may simply appear similar in the topic space to a breakthrough patent if it represents closely related technical content (e.g., patents on a device and on a method of its preparation). Second, a patent may have been granted before a similar, more influential patent but it may not be salient to other inventors and thus not directly impactful to the field through knowledge flows. This can be considered a "missed opportunity" for innovation, as a technology that could have accelerated the integration of new knowledge does not end up impacting the field, despite occupying the same position in the "technological space" as patents that did lead to direct breakthrough-causing knowledge flows. To reduce the occurrence of false positives identified by our methodology, we order patents according to their grant date in the technological trajectory calculations; this ordering is more likely than the filing order to represent causal knowledge flows.

Special attention should also be paid to the focal technology domain definition in reducing identification errors. If the domain is too narrow, it might miss potentially relevant breakthrough patents and evidence of their future integration. This may be why our test sample includes only 10 out of 28 breakthrough c-Si PV patents previously identified by Nemet and Husmann (2012) in the same 1976-1996 period. Some inventions directly relevant to the silicon-based PV technology are patented in the main CPC group for PV but outside of the silicon-based CPC subgroups. On the other hand, a domain that is too expansive and includes irrelevant patents can significantly affect both the topic model and field centroid calculations. Even in our narrow

domain definition, several of the patents on our short list of 98 c-Si PV breakthrough candidate patents (Appendix B) are of questionable relevance to PV technology (e.g. US4435897 and US4579626, both on photoimaging devices). A domain that includes several distinct technological sub-trajectories can also provide skewed results, as seen in our case for the inclusion of c-Si, amorphous silicon, and thin film PV technology. These factors must be all be weighed against each other, and even the most careful domain definition will produce results that require validation.

A final note should be made about the applicability of our method. While our method is able to identify a novel patent very soon after it is granted, its relevance to the future trajectory of innovation cannot, by definition, be known in the short run. Therefore, the method is applicable for retrospective analysis instead of forward-looking predictions of successful technological breakthroughs. Nevertheless, without understanding the dynamics and sources of past innovation, it is impossible to make informed strategic decisions and design policies for the development of future technological innovation. Our framework offers a new approach to studying technological breakthroughs that can complement traditional citation-based and classification-based methods and qualitative approaches.

6. Conclusion

The notion of "technological distance" can be traced back to at least Griliches (1979) who studied the technological closeness between firms or industries. Subsequent studies in the theories of technological change expanded the concept of "technological distance" to "technological space," and empirical metrics have been proposed utilizing product characteristics (Dodson, 1985), industry SIC codes (Teece, et al., 1994) and patent classes (Jaffe, 1986; McNamee, 2013; Aharonson and Schilling, 2016). In this paper, we apply natural language processing algorithms to systematically analyze the textual content of patents and represent their substance in technological space without relying on patent classifications or citation patterns to identify relatedness. We then apply simple geometric interpretations of patents in technological space to develop dynamic measures that indicate a patent's importance as a potential technological breakthrough in a field, following the theoretical work on innovation dynamics and recombinant innovation. We use these measures to create a list of potential breakthrough patents in crystalline silicon photovoltaics.

Our results are consistent with the observation that the initial novelty or "radicalness" of an invention is not enough to ensure that it becomes a successful technological breakthrough; breakthrough inventions must integrate into the technology through a series of subsequent relevant inventions and improvements. In the current age of high-powered computing and machine learning, our paper thus provides a timely contribution to operationalizing some of the theoretical contributions of past innovation research. We hope that this work can inform future research for "giving empirical content to theorizing about the role of knowledge in the modern economy" (Hall, Jaffe, and Trajtenberg, 2002).

Funding Statement:

This research is supported by the grant from the Alfred P. Sloan Foundation titled "What factors drive innovation in energy technologies? The role of technology spillovers and government investment."

Acknowledgement:

We greatly appreciate the feedback received on early versions of this paper at the 2019 Global TechMining Conference as well as the 2019 Association for Public Policy Analysis & Management Fall Research Conference. We are also thankful for the comments from the anonymous reviewers and the guest editor.

References:

Aharonson, B. S., & Schilling, M. A. (2016). Mapping the technological landscape: Measuring technology distance, technological footprints, and technology evolution. *Research Policy, 45*(1), 81-96.

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

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

Arthur, W. B. (2007). The structure of invention. *Research Policy*, *36*(2), 274-287.

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, *1*(1), 17-35.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*(Jan), 993-1022.

Bekkers, R., & Martinelli, A. (2012). Knowledge positions in high-tech markets: Trajectories, standards, strategies and true innovators. *Technological Forecasting and Social Change*, *79*(7), 1192-1216.

Briggs, K., & Buehler, D. L. (2018). An analysis of technologically radical innovation and breakthrough patents. *International Journal of the Economics of Business*, *25*(3), 341-365.

Callon, M., Courtial, J. P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Social Science Information*, *22*(2), 191-235.

Cascini, G., Fantechi, A., & Spinicci, E. (2004). Natural language processing of patents and technical documentation. In *International Workshop on Document Analysis Systems* (pp. 508-520). Springer, Berlin, Heidelberg.

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

Dincer, I. (2000). Renewable energy and sustainable development: a crucial review. *Renewable and Sustainable Energy Reviews*, *4*(2), 157-175.

Dodson, E. N. (1985). Measurement of state of the art and technological advance. *Technological Forecasting and Social Change*, *27*(2-3), 129-146.

Dosi, G. (1982). Technological paradigms and technological trajectories: a suggested interpretation of the determinants and directions of technical change. *Research Policy*, *11*(3), 147-162.

Fleming, L. (2001). Recombinant uncertainty in technological search. *Management Science*, *47*(1), 117-132.

Fleming, L. (2007). Breakthroughs and the "long tail" of innovation. *MIT Sloan Management Review*, *49*(1), 69.

Funk, R. J., & Owen-Smith, J. (2017). A dynamic network measure of technological change. *Management Science*, *63*(3), 791-817.

Gerken, J. M., & Moehrle, M. G. (2012). A new instrument for technology monitoring: novelty in patents measured by semantic patent analysis. *Scientometrics*, *91*(3), 645-670.

Geum, Y., Jeon, J., & Seol, H. (2013). Identifying technological opportunities using the novelty detection technique: A case of laser technology in semiconductor manufacturing. *Technology Analysis & Strategic Management*, *25*(1), 1-22.

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.

Green, M. A. (1995). *Silicon solar cells: advanced principles & practice*. Centre for Photovoltaic Devices and Systems.

Green, M. A. (2005). Silicon photovoltaic modules: a brief history of the first 50 years. *Progress in Photovoltaics: Research and applications*, *13*(5), 447-455.

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

Hall, B., Jaffe, A.B. & Trajtenberg, M. (2002). The NBER patent citations data file: Lessons, insights and methodological tools, in Jaffe, A.B. & Trajtenberg, M. (eds.), *Patents, citations, and innovations: A window on the knowledge economy*, MIT press, 403-451.

Haščič, I., and Migotto, M. (2015). Measuring environmental innovation using patent data. *OECD Environment Working Papers*, No. 89, OECD Publishing, Paris. DOI: https://doi.org/10.1787/5js009kf48xw-en.

Huang, Y., Zhu, D., Qian, Y., Zhang, Y., Porter, A. L., Liu, Y., & Guo, Y. (2017). A hybrid method to trace technology evolution pathways: a case study of 3D printing. *Scientometrics*, *111*(1), 185-204.

Huenteler, J., Schmidt, T. S., Ossenbrink, J., & Hoffmann, V. H. (2016a). Technology life-cycles in the energy sector—Technological characteristics and the role of deployment for innovation. *Technological Forecasting and Social Change*, *104*, 102-121.

Huenteler, J., Ossenbrink, J., Schmidt, T. S., & Hoffmann, V. H. (2016b). How a product's design hierarchy shapes the evolution of technological knowledge—Evidence from patent-citation networks in wind power. *Research Policy*, *45*(6), 1195-1217.

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

Husmann, D. (2011). *U.S. Photovoltaic Breakthroughs from 1947 To 1993: their Identification, Origin, and Commercialization*. [Master's thesis, University of Wisconsin-Madison].

Jaffe, A. B. (1986). Technological opportunity and spillovers of R&D: evidence from firms' patents, profits and market value. *NBER Working Paper*.

Jaffe, A.B., Newell, R. & Stavins, R. (2005). A tale of two market failures: Technology and environmental policy. *Ecological Economics*, *54*, 164-174.

Jaffe, A.B., Trajtenberg, M. (2002). *Patents, Citations, and Innovations: A Window on the Knowledge Economy*. MIT Press, Cambridge, Mass.

Jaffe, A. B., Trajtenberg, M., & Fogarty, M. S. (2000). Knowledge spillovers and patent citations: Evidence from a survey of inventors. *American Economic Review*, *90*(2), 215-218.

Joung, J., & Kim, K. (2017). Monitoring emerging technologies for technology planning using technical keyword based analysis from patent data. *Technological Forecasting and Social Change*, *114*, 281-292.

Kaplan, S., & Vakili, K. (2015). The double‐edged sword of recombination in breakthrough innovation. *Strategic Management Journal*, *36*(10), 1435-1457.

Kay, L., Newman, N., Youtie, J., Porter, A. L., & Rafols, I. (2014). Patent overlay mapping: Visualizing technological distance. *Journal of the Association for Information Science and Technology*, *65*(12), 2432-2443.

Kelley, D. J., Ali, A., & Zahra, S. A. (2013). Where Do Breakthroughs Come From? Characteristics of High-Potential Inventions. *Journal of Product Innovation Management*, *30*(6), 1212-1226.

Kim, D., Cerigo, D. B., Jeong, H., & Youn, H. (2016). Technological novelty profile and invention's future impact. *EPJ Data Science*, *5*(1), 1-15.

Kim, Y. G., Suh, J. H., & Park, S. C. (2008). Visualization of patent analysis for emerging technology. *Expert systems with applications*, *34*(3), 1804-1812.

Kuhn, J. M., Younge, K. A., & Marco, A. C. (2020). Patent citations reexamined. *The RAND Journal of Economics*, *51*(1), 109-132.

Lee, C., Kang, B., & Shin, J. (2015). Novelty-focused patent mapping for technology opportunity analysis. *Technological Forecasting and Social Change*, *90*, 355-365.

Lee, S., Yoon, B., & Park, Y. (2009). An approach to discovering new technology opportunities: Keyword-based patent map approach. *Technovation*, *29*(6-7), 481-497.

Levinthal, D. A., & March, J. G. (1993). The myopia of learning. *Strategic Management Journal*, *14*(S2), 95-112.

McNamee, R. C. (2013). Can't see the forest for the leaves: Similarity and distance measures for hierarchical taxonomies with a patent classification example. *Research Policy*, *42*(4), 855-873.

Mokyr, J. (1990). Equilibria and Technological Progress. *The American Economic Review*, *80*(2), 350-354.

Momeni, A., & Rost, K. (2016). Identification and monitoring of possible disruptive technologies by patent-development paths and topic modeling. *Technological Forecasting and Social Change*, *104*, 16-29.

Nemet, G. F. (2012). Inter-technology knowledge spillovers for energy technologies. *Energy Economics*, *34*(5), 1259-1270.

Nemet, G. F., & Husmann, D. (2012). PV learning curves and cost dynamics. In *Semiconductors and semimetals* (Vol. 87, pp. 85-142). Elsevier.

Phene, A., Fladmoe-Lindquist, K., & Marsh, L. (2006). Breakthrough innovations in the US biotechnology industry: the effects of technological space and geographic origin. *Strategic Management Journal, 27*(4), 369-388.

Pilkington, A., Dyerson, R., & Tissier, O. (2002). The electric vehicle: Patent data as indicators of technological development. *World Patent Information, 24*(1), 5-12.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program, 14*(3), 130-137.

Popp, D., Santen, N., Fisher-Vanden, K., & Webster, M. (2013). Technology variation vs. R&D uncertainty: What matters most for energy patent success? *Resource and Energy Economics*, 35(4), 505-533.

Ranaei, S., & Suominen, A. (2017). Using machine learning approaches to identify emergence: Case of vehicle related patent data. In *2017 Portland International Conference on Management of Engineering and Technology (PICMET)* (pp. 1-8). IEEE.

Rieger, J., Rahnenführer, J., & Jentsch, C. (2020). Improving Latent Dirichlet Allocation: On Reliability of the Novel Method LDAPrototype. In *International Conference on Applications of Natural Language to Information Systems* (pp. 118-125). Springer, Cham.

Rosenberg, N. (1994). *Exploring the Black Box: Technology, Economics, and History*. Cambridge University Press, Cambridge.

Rosenkopf, L., & Nerkar, A. (2001). Beyond local search: boundary-spanning, exploration, and impact in the optical disk industry. *Strategic Management Journal, 22*(4), 287-306.

Rotolo, D., Hicks, D., & Martin, B. R. (2015). What is an emerging technology?. *Research Policy, 44*(10), 1827-1843.

Schmoch, U. (2008). Concept of a technology classification for country comparisons. *Final report to the world intellectual property organisation (WIPO), WIPO*.

Shane, S. (2001). Technological opportunities and new firm creation. *Management Science, 47*(2), 205-220.

Schoenmakers, W., & Duysters, G. (2010). The technological origins of radical inventions. *Research Policy*, *39*(8), 1051-1059.

Strumsky, D., & Lobo, J. (2015). Identifying the sources of technological novelty in the process of invention. *Research Policy*, *44*(8), 1445-1461.

Suominen, A., & Newman, N. C. (2017). Exploring the fundamental conceptual units of technical emergence. In *2017 Portland International Conference on Management of Engineering and Technology (PICMET)*. IEEE.

Suominen, A., Toivanen, H., & Seppänen, M. (2017). Firms' knowledge profiles: Mapping patent data with unsupervised learning. *Technological Forecasting and Social Change*, *115*, 131-142.

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.

Teece, D. J., Rumelt, R., Dosi, G., & Winter, S. (1994). Understanding corporate coherence: Theory and evidence. *Journal of Economic Behavior & Organization*, *23*(1), 1-30.

Tseng, Y. H., Lin, C. J., & Lin, Y. I. (2007). Text mining techniques for patent analysis. *Information Processing & Management*, *43*(5), 1216-1247.

U.S. Department of Energy. (2004). *The History of Solar*. Available: https://www1.eere.energy.gov/solar/pdfs/solar_timeline.pdf

Usher, A.P. (1954). *A History of Mechanical Inventions: Revised edition*. Harvard University Press, Cambridge, MA.

Verhoeven, D., Bakker, J., & Veugelers, R. (2016). Measuring technological novelty with patent-based indicators. *Research Policy*, *45*(3), 707-723.

Verspagen, B. (2007). Mapping technological trajectories as patent citation networks: A study on the history of fuel cell research. *Advances in Complex Systems*, *10*(01), 93-115.

Vom Stein, N., Sick, N., & Leker, J. (2015). How to measure technological distance in collaborations—The case of electric mobility. *Technological Forecasting and Social Change*, *97*, 154-167.

Wang, J., & Chen, Y. J. (2019). A novelty detection patent mining approach for analyzing technological opportunities. *Advanced Engineering Informatics*, *42*, 100941.

Yau, C. K., Porter, A., Newman, N., & Suominen, A. (2014). Clustering scientific documents with topic modeling. *Scientometrics*, *100*(3), 767-786.

Yoon, J., & Kim, K. (2011). Identifying rapidly evolving technological trends for R&D planning using SAO-based semantic patent networks. *Scientometrics*, *88*(1), 213-228.

Yoon, J., & Kim, K. (2012). Detecting signals of new technological opportunities using semantic patent analysis and outlier detection. *Scientometrics*, *90*(2), 445-461.

Yoon, B., & Park, Y. (2004). A text-mining-based patent network: Analytical tool for high-technology trend. *The Journal of High Technology Management Research*, *15*(1), 37-50.

Youn, H., Strumsky, D., Bettencourt, L. M., & Lobo, J. (2015). Invention as a combinatorial process: evidence from US patents. *Journal of The Royal Society Interface*, *12*(106), 20150272.

Zhang, Y., Porter, A. L., Hu, Z., Guo, Y., & Newman, N. C. (2014). "Term clumping" for technical intelligence: A case study on dye-sensitized solar cells. *Technological Forecasting and Social Change*, *85*, 26-39.

Zhang, Y., Zhang, G., Zhu, D., & Lu, J. (2017). Scientific evolutionary pathways: Identifying and visualizing relationships for scientific topics. *Journal of the Association for Information Science and Technology*, *68*(8), 1925-1939.

Appendix A: Topic model construction and parameter tuning

In this appendix, we describe the steps performed to construct the topic model.

1. Constructing topic model corpus

We first construct our patent corpus using *quanteda*, an R package for quantitative analysis of textual data (https://quanteda.io). The corpus contains 3,126 USPTO patent full texts downloaded from PatentsView database (https://www.patentsview.org/download).

We apply Porter's stemming algorithm (Porter, 1980) to the corpus and complemented the built-in stop-word list of common English-language words with an additional list of generic scientific and technical terms that are common among all patents and do not offer substance to a patent's technological content. Examples of such stop-words are "table(s)", "description(s)", "solution". We also exclude measurement units, such as "mol", "mg", "g", "cm", "mm" from the corpus. Only unigrams (i.e. single terms) are used to construct topic models.

Finally, to reduce the size of the document-feature matrix (DFM) and mitigate the effect of outlier terms influencing the topic model, we trim the matrix to only include terms that appear in the corpus 10 times or more (i.e. appearing in a single patent at least 10 times or appearing in multiple patents more than 10 times in total).

The size of the final DFM is 3126 × 10398, indicating 3125 documents and 10398 word tokens used to construct the topic model.

2. Selecting optimal number of topics

The most important exogenous parameter we need to select to construct the topic model is the number of topics. We use the R package *ldatuning* (https://github.com/nikita-moor/ldatuning) by Nikita Murzintcev to select the optimal number of topics. It constructs four metrics by Arun et al. (2010), Cao et al. (2009), Deveaud et al. (2014) and Griffiths and Steyvers (2004). Figure A.1 shows the results from the LDA tuning function, where the optimal number of topics should minimize the metrics in the top graph and maximize the metrics in the bottom graph.

According to Figure A.1 and different sets of metrics taken together, the optimal number of topics most likely lies between 175 and 225. Therefore, we decide to construct the topic models

with 175, 200 and 225 topics and compare the results for these topic models in a separate sensitivity analysis presented in section 4.4 of the main text.

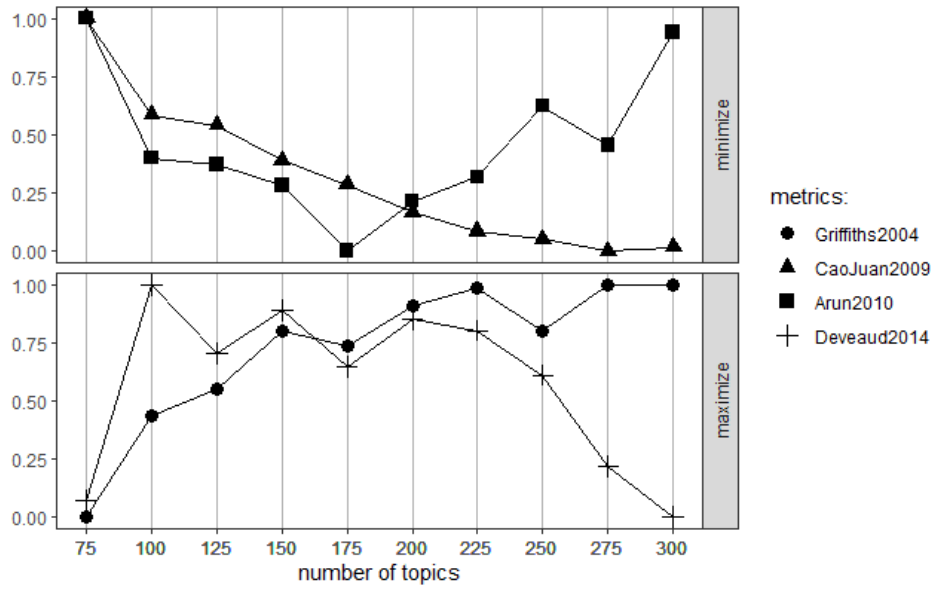


Figure A.1. LDA tuning results

3. Hyper-parameter alpha selection

Hyper-parameter alpha is selected using training and test sets. Specifically, 1000 patents are randomly selected as the training set and another 250 patents are randomly selected as the test set. We construct the LDA topic models using the same training dataset and a range of alpha values as the initial value. We use R package *topicmodels* (https://cran.r-project.org/web/packages/topicmodels/index.html) to estimate the alpha value based on the initial alpha value supplied. Then we apply the topic model to the same test dataset to calculate perplexity value, which measures the fit of the topic model. The lower the perplexity, the better the model. Results shown in Table A.1 show that the optimal alpha value is estimated at around 0.01 for 175, 200 and 225 topics.

Finally, we build topic models on the original DFM consisting of 3125 patents using the selected hyper-parameters and the *topicmodels* package.

Table A.1. Perplexity values of topic models given the number of topics and initial alpha value

	Initial alpha = 0.001	Initial alpha = 0.01	Initial alpha = 0.1	Initial alpha = 1	No initial alpha assigned
Number of topics = 175	582.96	579.45	572.95 Estimated alpha = 0.01	576.78	573.34
Number of topics = 200	574.59	566.67	557.25	566.98	556.39 Estimated alpha = 0.01
Number of topics = 225	570.21	560.52	554.08	555.14	553.50 Estimated alpha = 0.01

References:

Arun, R., Suresh, V., Madhavan, C. V., & Murthy, M. N. (2010). On finding the natural number of topics with latent Dirichlet allocation: Some observations. In *Pacific-Asia conference on knowledge discovery and data mining* (pp. 391-402). Springer, Berlin, Heidelberg.

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing*, *72*(7-9), 1775-1781.

Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique*, *17*(1), 61-84.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, *101*(suppl 1), 5228-5235.

Appendix B. Short list of 98 breakthrough patent candidates in crystalline silicon solar PV

Index	Patent number	Grant year	Full USPTO citation count	Title
1	4059461	1977	160	Method for improving the crystallinity of semiconductor films by laser beam scanning and the products thereof
2	4053326	1977	9	Photovoltaic cell
3	4104091	1978	48	Application of semiconductor diffusants to solar cells by screen printing
4	4140545	1979	3	Plural solar cell arrangement including transparent interconnectors
5	4187126	1980	37	Growth-orientation of crystals by raster scanning electron beam
6	4214920	1980	22	Method for producing solar cell-grade silicon from rice hulls
7	4195067	1980	13	Process for the production of refined metallurgical silicon
8	4255208	1981	57	Method of producing monocrystalline semiconductor films utilizing an intermediate water dissolvable salt layer
9	4302763	1981	21	Semiconductor device
10	4313254	1982	34	Thin-film silicon solar cell with metal boride bottom electrode
11	4343830	1982	17	Method for improving the efficiency of solar cells having imperfections
12	4332838	1982	8	Particulate thin film fabrication process
13	4409423	1983	49	Hole matrix vertical junction solar cell
14	4377564	1983	26	Method of producing silicon
15	4390743	1983	23	Silicon layer solar cell and method of producing it
16	4402771	1983	6	Substrate for silicon solar cells
17	4427839	1984	77	Faceted low absorptance solar cell
18	4478879	1984	70	Screen printed interdigitated back contact solar cell
19	4451969	1984	52	Method of fabricating solar cells
20	4431858	1984	37	Method of making quasi-grain boundary-free polycrystalline solar cell structure and solar cell structure obtained thereby
21	4449286	1984	26	Method for producing a semiconductor layer solar cell
22	4435897	1984	9	Method for fabricating solid-state image sensor
23	4496788	1985	67	Photovoltaic device
24	4514580	1985	49	Particulate silicon photovoltaic device and method of making
25	4520552	1985	33	Semiconductor device with deep grip accessible via the surface and process for manufacturing same
26	4554203	1985	11	Method for manufacturing large surface silicon crystal bodies for solar cells, and bodies so produced

27	4497675	1985	6	Process for the manufacture of substrates from carbon-coated silicon dioxide fabric that can be used for large-surface silicon bodies
28	4571448	1986	84	Thin film photovoltaic solar cell and method of making the same
29	4589191	1986	76	Manufacture of high efficiency solar cells
30	4610077	1986	40	Process for fabricating a wraparound contact solar cell
31	4602120	1986	39	Solar cell manufacture
32	4607132	1986	30	Integrated PV-thermal panel and process for production
33	4599244	1986	16	Method large-area silicon bodies
34	4579626	1986	15	Method of making a charge-coupled device imager
35	4577588	1986	12	Device for process-type deposition of polycrystalline silicon on carbon film
36	4577393	1986	8	Process for the production of a solar cell
37	4618874	1986	6	Solid state imaging device
38	4602422	1986	5	Flash compression process for making photovoltaic cells
39	4667060	1987	65	Back junction photovoltaic solar cell
40	4703553	1987	39	Drive through doping process for manufacturing low back surface recombination solar cells
41	4681983	1987	26	Semiconductor solar cells
42	4676845	1987	18	Passivated deep p/n junction
43	4640001	1987	17	Solar cell manufacturing method
44	4690797	1987	15	Method for the manufacture of large area silicon crystal bodies for solar cells
45	4643797	1987	4	Method for the manufacture of large area silicon crystal bodies for solar cells
46	4652901	1987	2	Infrared sensitive silicon substrate with integrated electronic processing devices and method for producing same
47	4751191	1988	48	Method of fabricating solar cells with silicon nitride coating
48	4778478	1988	30	Method of making thin film photovoltaic solar cell
49	4758525	1988	26	Method of making light-receiving diode
50	4782028	1988	23	Process methodology for two-sided fabrication of devices on thinned silicon
51	4743955	1988	16	Photoelectric converting device
52	4780394	1988	13	Photosensitive semiconductor device and a method of manufacturing such a device
53	4824489	1989	31	Ultra-thin solar cell and method
54	4818337	1989	30	Thin active-layer solar cell with multiple internal reflections
55	4818357	1989	20	Method and apparatus for sputter deposition of a semiconductor homojunction and

				semiconductor homojunction products created by same
56	4927770	1990	148	Method of fabricating back surface point contact solar cells
57	4947219	1990	86	Particulate semiconductor devices and methods
58	4891325	1990	12	Method for re-using silicon base material of a metal insulator semiconductor (mis) inversion-layer solar cell
59	4961097	1990	7	High frequency photo detector and method for the manufacture thereof
60	5053083	1991	196	Bilevel contact solar cells
61	5011565	1991	119	Dotted contact solar cell and method of making same
62	5057439	1991	64	Method of fabricating polysilicon emitters for solar cells
63	4989059	1991	51	Solar cell with trench through pn junction
64	5053355	1991	47	Method and means for producing a layered system of semiconductors
65	5011782	1991	40	Method of making passivated antireflective coating for photovoltaic cell
66	5019176	1991	35	Thin solar cell and lightweight array
67	5067985	1991	26	Back-contact vertical-junction solar cell and method
68	5017243	1991	23	Solar cell and a production method therefor
69	5011567	1991	22	Method of fabricating solar cells
70	4992392	1991	17	Method of making a virtual phase CCD
71	5118362	1992	62	Electrical contacts and methods of manufacturing same
72	5131933	1992	31	Solar cell
73	5082791	1992	29	Method of fabricating solar cells
74	5147468	1992	9	Photovoltaic semiconductor device and method for manufacturing the same
75	5178685	1993	117	Method for forming solar cell contacts and interconnecting solar cells
76	5248621	1993	67	Method for producing solar cell devices of crystalline material
77	5258077	1993	34	High efficiency silicon solar cells and method of fabrication
78	5270221	1993	18	Method of fabricating high quantum efficiency solid state sensors
79	5264376	1993	15	Method of making a thin film solar cell
80	5215599	1993	13	Advanced solar cell
81	5279682	1994	60	Solar cell and method of making same
82	5326719	1994	12	Thin film growth using two part metal solvent
83	5360745	1994	4	Thin-film solar cell production method
84	5362978	1994	3	Method for establishing an electrical field at a surface of a semiconductor device
85	5468652	1995	143	Method of making a back contacted solar cell

86	5461002	1995	51	Method of making diffused doped areas for semiconductor components
87	5415700	1995	11	Concrete solar cell
88	5403439	1995	9	Method of producing same-sized particles
89	5385849	1995	8	Process of fabricating solid-state image pick-up device free from crystal defects in active region
90	5538564	1996	175	Three dimensional amorphous silicon/microcrystalline silicon solar cells
91	5494832	1996	56	Method for manufacturing a solar cell from a substrate wafer
92	5543333	1996	49	Method for manufacturing a solar cell having combined metallization
93	5508206	1996	13	Method of fabrication of thin semiconductor device
94	5510095	1996	11	Production of high-purity silicon ingot
95	5489555	1996	4	Method for forming a photoelectric conversion device
96	5504015	1996	1	Process for preparing photovoltaic modules based on crystalline silicon
97	5585283	1996	1	Method for establishing an electrical field by thinning an implanted region at a surface of a semiconductor device
98	5540183	1996	0	Zone-melting recrystallization of semiconductor materials

Appendix C. Regression analysis

In this Appendix, we examine the correlations between our set of metrics of novelty and relevance with forward citations count. Forward citation counts are frequently used as a proxy metric of a patent's influence despite the limitations discussed in section 2. Still, the goal of this analysis is to assess the extent to which the methodology we propose in this paper yields consistent results with existing metrics.

The following hypotheses based on our conceptual framework are tested using a patent's initial novelty value and its novelty and relevance slopes, calculated as linear regression coefficients of subsequent annual novelty and relevance values against the number of years since the patent was granted:

> H.1. A patent's initial novelty value and its forward citation count are positively correlated.
>
> H.2. A patent's initial relevance value and its forward citation count are positively correlated.
>
> H.3. Patents with *negative* novelty slopes are more likely to have higher forward citation counts (i.e. integrating patent), and among those, steeper slopes are associated with higher counts (i.e. rapid integration).
>
> H.4. Patents with *positive* relevance slopes are likely to have higher forward citation counts (i.e. integrating patent), and among those, steeper slopes are associated with higher counts (i.e. rapid integration).

Hypothesis H.1 suggests that patents with higher initial novelty are more likely to be highly cited compared to patents with lower initial novelty. In other words, patents that are different from the existing inventions are more likely to be cited by later inventions. Hypothesis H.2 suggests that patents that are more relevant to the subsequent technological trajectory are more likely to be highly cited. Hypotheses H.3 and H.4 propose that integrating patents are more likely to be highly cited, especially the patents with rapid integration patterns.

To test these hypotheses, the following empirical models are constructed to correspond to the four hypotheses:

H.1. $y = \beta_0 + \beta_1 N_{initial} + \delta + \epsilon, \quad (C.1)$

H.2. $y = \beta_2 + \beta_3 R_{initial} + \delta + \epsilon, \quad (C.2)$

H.3. $y = \beta_4 + \beta_5 |N| + \beta_6 Neg(N) + \beta_7 |N| \times Neg(N) + \delta + \epsilon, \quad (C.3)$

H.4. $y = \beta_8 + \beta_9 |R| + \beta_{10} Pos(R) + \beta_{11} |R| \times Pos(R) + \delta + \epsilon, \quad (C.4)$

where $N_{initial}$ and $R_{initial}$ are the initial novelty and relevance of a patent, $|N|$ and $|R|$ are the absolute values of the novelty and relevance slopes, $Neg(N)$ is a dummy variable indicating whether the novelty slope is negative while $Pos(R)$ is a dummy variable indicating whether the relevance slope is positive, δ are fixed effects representing the year in which a patent is granted and ϵ is the error term.

The reason for disaggregating the slopes into absolute values and their signs is that they might have different predictive power for forward citation count as the sign of the slopes represents a distinct phenomenon, as described in Table 1. According to hypotheses H.2 and H.3, only negative novelty and positive relevance slopes would be expected to be associated with high forward citation counts, and their steepness in this direction would indicate a more important role. As equations (3) and (4) in the main text show, novelty and relevance values both depend on $\left|\overrightarrow{C_{t_1} p_{\iota,t}}\right|$ and θ. Because of this co-determination of novelty and relevance based on the position of each focal patent, the novelty and relevance slopes are not included in a regression model simultaneously. While the correlation between novelty and relevance slopes need not be high in this dataset, the correlation between $|N|$ and $|R|$ is 0.781 and the correlation between $Neg(N)$ and $Pos(R)$ is 0.716.

If hypothesis H.1 is true, then the estimated coefficient of β_1 in equation (C.1) would be positive, and if hypothesis H.2 is true, then the estimated coefficient of β_3 in equation (C.2) would be positive.

To understand the implication of hypothesis H.3 on the regression coefficients, we write equation (C.3) as follows, which breaks the equation into a piecewise equation based on the dummy variable for the sign of the novelty slope:

$$\begin{cases} y = \beta_4 + \beta_6 + (\beta_5 + \beta_7)|N| + \delta + \epsilon, & for\ N \le 0; \quad (C.5) \\ y = \beta_4 + \beta_5|N| + \delta + \epsilon, & for\ N > 0. \quad (C.6) \end{cases}$$

If the first half of H.3 is true (i.e. patents with a negative novelty slope have higher forward citation counts), then $\beta_6 + \beta_7\overline{|N|}$ should be positive, where $\overline{|N|}$ is the average absolute value of N. This means either β_6 or β_7 or both should be positive. If the second half of H.3 is true (i.e. patents with negative and steep slopes have higher forward citation counts), then β_7 should be positive, so that the coefficient of $|N|$ is higher for patents with more negative slopes. Overall, if H.3 is true, then the estimated coefficient for β_7 should be positive.

Similarly, equation (C.4) can be written as a piecewise function:

$$\begin{cases} y = \beta_8 + \beta_9|R| + \delta + \epsilon, & for\ R < 0; \quad (C.7) \\ y = \beta_8 + \beta_{10} + (\beta_9 + \beta_{11})|R| + \delta + \epsilon, & for\ R \ge 0. \quad (C.8) \end{cases}$$

If the first half of H.4 is true (i.e. positive relevance slope is associated with greater forward citations), then $\beta_{10} + \beta_{11}\overline{|R|}$ should be positive, where $\overline{|R|}$ is the average absolute value of R. This means β_{10} or β_{11} or both should be positive. If the second half of H.4 is true (i.e. positive and steep relevance slopes indicate higher forward citations), then β_{11} should be positive, so that the coefficient for $|R|$ is higher for patents with positive relevance slopes. Overall, H.4 implies that the estimated coefficient for β_{11} should be positive.

Table C.1 shows the descriptive statistics of the variables used for the regression. 20-year forward citation count is the dependent variable, which is the number of forward citations the focal patent received within 20 years after it was granted. Granted year fixed effect accounts for year-specific innovative activities in the field and any other time-specific factors. In addition to running regressions on the whole sample, we split the dataset into subsamples based on initial novelty and relevance values to obtain additional insights on the validity of our hypotheses.

Table C.1 Descriptive statistics for silicon PV patents (1977-1996, N=765)

	Mean	Median	Std. Dev.	Min.	Max.
Initial novelty	43.08	35.00	20.74	19.17	100.48
Initial relevance	6.32	6.09	11.14	-51.90	70.08
Novelty slope, absolute	0.046	0.029	0.213	0.000	0.708

Relevance slope, absolute	0.206	0.150	0.068	0.000	1.555
Negative novelty slope (dummy)	0.465	0	0.499	0	1
Positive relevance slope (dummy)	0.490	0	0.500	0	1
Forward citation count (20 years)	14.01	9	17.30	0	179

Table C.2 Regression results for H.1 and H.2.

	Full sample		$n_{initial}$ ≤ median	$r_{initial}$ ≤ median	$n_{initial}$ > median	$r_{initial}$ > median
	(1)	(2)	(3)	(4)	(5)	(6)
Initial novelty ($\hat{\beta}_1$)	0.005*** (0.002)		-0.012 (0.012)		0.0049* (0.003)	
Initial relevance ($\hat{\beta}_3$)		0.0040 (0.003)		0.00051 (0.006)		0.0159** (0.005)
Constant	2.42***	2.54***	3.09***	2.87***	1.80***	1.628***
Granted year fixed effect	Yes	Yes	Yes	Yes	Yes	Yes
Observations	765	765	397	397	368	368
2×Log-likelihood	-5495	-5503	-2628	-2750	-2815	-2699

Table C.3 Regression results for H.3 and H.4.

	Full sample		$n_{initial}$ ≤ median	$r_{initial}$ ≤ median	$n_{initial}$ > median	$r_{initial}$ > median
	(1)	(2)	(3)	(4)	(5)	(6)
$\|N\|$ ($\hat{\beta}_5$)	1.11** (0.508)		0.28 (1.032)		1.26** (0.612)	
$\|R\|$ ($\hat{\beta}_9$)		0.51*** (0.172)		0.47 (0.390)		0.61*** (0.205)
$Neg(N)$ ($\hat{\beta}_6$)	-0.29*** (0.084)		-0.37** (0.123)		-0.30** (0.140)	
$Pos(R)$ ($\hat{\beta}_{10}$)		-0.14 (0.112)		0.005 (0.163)		-0.48*** (0.157)
$\|N\| \times Neg(N)$ ($\hat{\beta}_7$)	4.07 ** (1.850)		10.18*** (2.290)		0.49 (3.240)	
$\|R\| \times Pos(R)$ ($\hat{\beta}_{11}$)		0.23 (0.608)		-0.63 (0.844)		2.60*** (0.963)

Constant	2.57***	2.51***	2.69***	2.84***	2.04***	1.70***
Granted year fixed effect	Yes	Yes	Yes	Yes	Yes	Yes
Observations	765	765	397	397	368	368
2×Log-likelihood	-5485	-5487	-2610	-2754	-2805	-2684

Z statistics in parentheses *** p<0.01, ** p<0.05, * p<0.1.

Tables C.2-C.3 present the estimated coefficients of equations (C.1) to (C.4) using the full sample, and the subsamples with the initial novelty (or relevance) value below and above the median.

Overall, the signs and significance of the coefficients are consistent with all four hypotheses in at least one of the samples. The estimated coefficients of initial novelty ($\hat{\beta}_1$) are positive and significant in the full sample as well as the subsample where the initial novelty value is higher than the median level. This suggests that the positive correlation between initial novelty and forward citation counts only exists when the initial novelty is sufficiently high. Similarly, the positive correlation between initial relevance and forward citation counts only exists in the subsample with the initial relevance above the median ($\hat{\beta}_3$).

The estimated coefficient $\hat{\beta}_7$ is positive and significant for the full sample and the subsample with initial novelty value lower than median and is insignificant for the subsample with initial novelty value above the median. The estimated coefficient $\hat{\beta}_{11}$ is positive and significant only for the subsample where the initial relevance is above the median. Figure C.1 summarizes the empirical findings on hypotheses H.3 and H.4 using the estimated regression coefficients.

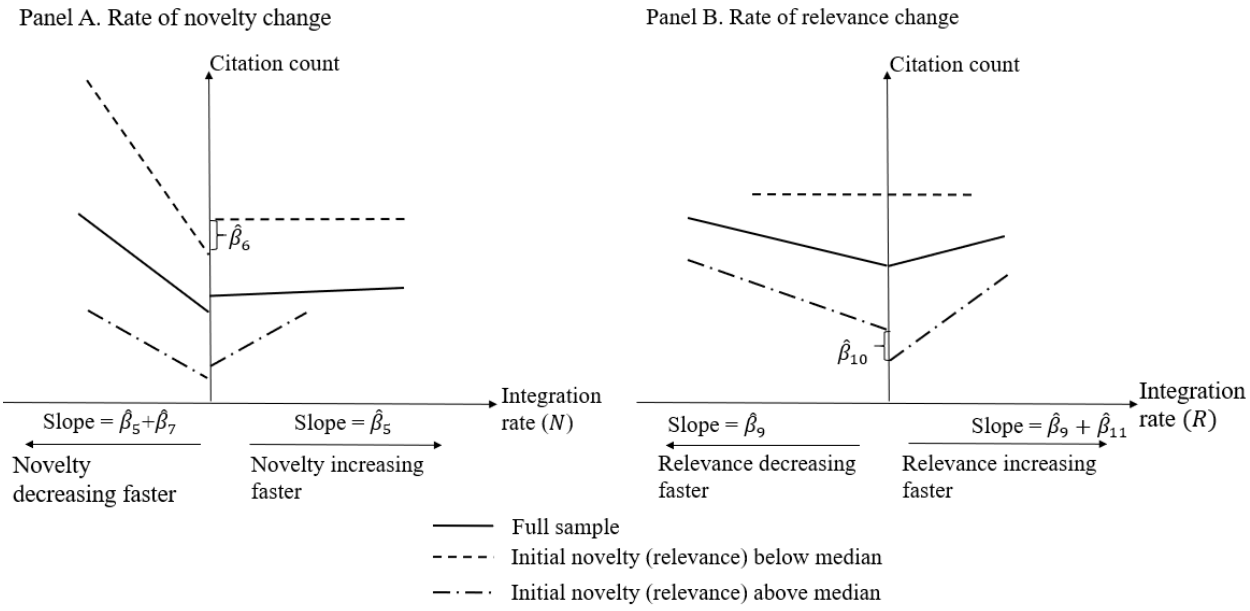


Figure C.1 Representation of the correlation between N (R) and forward citation count.

In Panel A of Figure C.1, the slopes of the lines are based on the estimated coefficients $\hat{\beta}_5$ and $\hat{\beta}_7$, and the difference in intercepts is $\hat{\beta}_6$. The graph shows the trend described in hypothesis H.3: integrating patents with declining novelty (i.e. $N < 0$) mostly have higher forward citation count, and the rate of integration is positively correlated with forward citation count. This trend is more predominant among patents with lower initial novelty compared to the full sample. In Panel B, the slopes of the lines are based on $\hat{\beta}_9$ and $\hat{\beta}_{11}$, and the difference in intercepts is $\hat{\beta}_{10}$. Hypothesis H.4 is also partially validated: it is unclear whether patents with increasing relevance (i.e. $R > 0$) have higher forward citations. But among the patents with increasing relevance and initial relevance above median value, the faster the integrating process is, the higher the forward citation count a patent has.

Validation of these hypotheses suggests that our measures do not contradict traditional forward citation counts as a measure of patent influence. They offer additional information and perspectives on understanding what constitutes breakthroughs and on the integration process of highly cited patents. Future analyses can examine high-residual outliers from the regression as potential breakthrough patents not reflected by their forward citation counts.