Kim et al. *EPJ Data Science* (2016) 5:8
DOI 10.1140/epjds/s13688-016-0069-1

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

Technological novelty profile and invention's future impact

Daniel Kim[1,2,3], Daniel Burkhardt Cerigo[3], Hawoong Jeong[4,5,6] and Hyejin Youn[1,3,7*]

*Correspondence:
youn@maths.ox.ac.uk
[1]Institute for New Economic Thinking at the Oxford Martin School, University of Oxford, Oxford, OX2 6ED, United Kingdom
[3]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA
[7]Mathematical Institute, University of Oxford, Oxford, OX2 6GG, United Kingdom
Full list of author information is available at the end of the article

Abstract

We consider inventions as novel combinations of existing technological capabilities. Patent data allow us to explicitly identify such combinatorial processes in invention activities (Youn *et al.* in *J R Soc Interface* 12:20150272, 2015). Unconsidered in the previous research, not every new combination is novel to the same extent. Some combinations are naturally anticipated based on patent activities in the past or mere random choices, and some appear to deviate exceptionally from existing invention pathways. We calculate a relative likelihood that each pair of classification codes is put together at random, and a deviation from the empirical observation so as to assess the overall novelty (or conventionality) that the patent brings forth at each year. An invention is considered as unconventional if a pair of codes therein is unlikely to be used together given the statistics in the past. Temporal evolution of the distribution indicates that the patenting activities become more conventional with occasional cross-over combinations. Our analyses show that patents introducing novelty on top of the conventional units would receive higher citations, and hence have higher impact.

Keywords: invention; patent; patent citation; co-occurrence; standard score; technology code; technological novelty

1 Introduction

A new idea that advances science and technology is commonly recognised as an important source of wealth creation, economic growth, and societal change [2, 3]. The steam engine, transistor and lithium ion battery are all such examples. Therefore, it is of no surprise that understanding the dynamics of generation of new ideas sits at the centre of many disciplines [4–18].

With an increasing volume of electronic corpora available online, research on the systems of science and technology, once considered to be a domain of humanities, social sciences and economics, has expanded its realm to be a subject of data science [5]. The growing empirical literature in this respect is to identify the process of publication [19], to utilise Google n-gram to characterise scientific evolution [11], to delineate the boundary of science [20, 21], and to predict the future impact of scientific papers [22, 23] and authors [24, 25].

In the case of inventive activities, Youn and co-workers availed themselves of technology codes, classified by United States Patent and Trademark Office (USPTO), as countable units to identify the underlying dynamics of inventions as combinatorial processes in a

comprehensive and explicit way [1]. In this way, an invention yields either a new unit of technological capability, a new way of combining the already existing units making an innovative function, or a refinement of existing combinations. When inventive activities are viewed in this way, it is also found, the rate at which the new combinations are introduced has been *invariant* over two centuries, implying that a combinatorial process is the nature of invention [1].

Building on these previous findings, we delve into the temporal evolution of this combinatorial process using U.S. patent data from 1836 to 2014. We first describe the data structure, and elaborate our method to quantify technological novelty scores. We then show how the conventional and novel pairings within an invention would affect its future impact. Finally we will discuss the implications of our findings.

2 Data

The U.S. patent records began at July 31, 1790 with Samuel Hopkins' patent on pot ash [26]. Since then, there has been almost ten million inventions granted over two hundred years [27]. Among them, we only consider utility patents, which are those that pertain to new and useful inventions, omitting design and plant patents for instance [28]. Patents that are explicitly marked by utility patents only begin from 1836, and amount to 8,884,909 patents as of December 2014, taking up almost 90% of the entire record. Among them, we analyse patents that have two or more technology codes (80.3%).

In order for examiners to efficiently search for relevant prior arts, the U.S. patent office encode salient technological capabilities into six-position alphanumeric codes. Every patent is then tagged with a combination of codes that represent the technologies involved in the invention [29]. The classification codes are created in a nested structure: 473 classes at the highest level and 168,743 codes at the lowest (most detailed) level. These low-level codes ('codes' from herein) can lie on different levels of the hierarchy tree; some classes have deeper branching than others.

We used patent citation data provided by National Bureau of Economic Research (NBER) and considered patents' citations as a measure of their impact [30, 31]. The citation data, NBER, span only 31 years, from 1976 to 2006 unlike the co-occurrence data covering almost two hundred years. In order to cover as large data as possible, we use the co-occurrence data, spanning over one hundred years, for the Section 4.1, but we had to use NBER data (smaller dataset) which have citation information that is needed in the Section 4.2. In order to control the temporal effect on citation volume, we use the citation number only up to the first five years after publication because it is also known that only recent citation works well in prediction of future impact [32]. This leaves us data spanning 26 years (from 1976 to 2001) [8, 33].

3 Methods

We aim to assess the novelty of technological constituents in each patent, and then compare aspects of this novelty to the patent's impact. We measure how technology codes are combined in the empirical data and compare the observed combination to what would be expected if the combinations were randomly configured. In this way, we can discern recurring themes within invention space and also those combinations that are unconventional or novel. These features, measured by well established standard scores, are related to patent future impact.

3.1 Standard scores (z-scores) of code pairs

The patent data P can be represented by a collection of sets of classification codes, where each set corresponds to an individual patent and contains its classification codes. The z-score for a pair of codes, α and β is expressed as:

$$z_{\alpha\beta} = \frac{o_{\alpha\beta} - \mu_{\alpha\beta}}{\sigma_{\alpha\beta}}, \tag{1}$$

where $o_{\alpha\beta}$ is the observed number of times the code α appears together with β within a patent (a set) within the actual data. $\mu_{\alpha\beta}$ and $\sigma_{\alpha\beta}$ are the expected co-occurrences of the codes and its standard deviation, derived from a null model of the data which randomises code arrangement while preserving code usage and number of patents within the data (the Section 3.2 provides the detail).

The observed co-occurrences $o_{\alpha\beta}$ in the patent record is compared with $\mu_{\alpha\beta}$. If the two codes appear together more often than expected, then Eq. (1) results in a positive value, or if they are rarely paired within a patent relative to their expected occurrences then their z-score is negative. The degree to which the deviation is significant is derived by normalising the value by the expected standard deviation σ [8, 9, 34, 35]. We can thus associate high z-scores with very typical code pairing, and conversely, a negative z-score is indicative of an atypical or novel pairing of codes.

3.2 Expected co-occurrences

The null model acts as the baseline by which we deem an aspect of the data to have statistical significance, beyond what would occur by random, or with no underlying pattern or law. The aspect in consideration is the arrangement of the codes between the patents. The premise of the null model is that each of these arrangements of codes is equally likely. From these possible arrangements the expected pairing counts can be computed.

Consider codes α and β, with the number of occurrences n_α and n_β within the set of patents P. Noting that patents cannot be classified with the same code twice, the number of possible configurations of the α and β into the $|P|$ possible patents is $\binom{|P|}{n_\alpha}\binom{|P|}{n_\beta}$. Now consider those arrangements which contain exactly x co-occurrences, within a patent, of α and β. There are $\binom{|P|}{n_\alpha}\binom{n_\alpha}{x}\binom{|P|-n_\alpha}{n_\beta-x}$ possible configurations; first distribution the α into $|P|$, then x of the β into those n_α patents already assigned an α, finally distribute the remaining β into the patents without an α. Thus giving a hypergeometric probability distribution for the number of co-occurrences:

$$p(o_{\alpha\beta} = x) = \frac{\binom{n_\alpha}{x}\binom{|P|-n_\alpha}{n_\beta-x}}{\binom{|P|}{n_\beta}} \tag{2}$$

thus the expected number of patents that have both α and β is:

$$\mu_{\alpha\beta} = \frac{n_\alpha n_\beta}{|P|} \tag{3}$$

and the variance of $\mu_{\alpha\beta}$ is:

$$\sigma^2_{\alpha\beta} = \mu_{\alpha\beta}\left(1 - \frac{n_\alpha}{|P|}\right)\left(\frac{|P| - n_\beta}{|P| - 1}\right). \tag{4}$$

3.3 Incorporating temporal evolution

As new technologies become successful, so they may subsequently become established areas of inventive activity. Following z-scores in time allows us to observe the case where an invention may have been exceptionally novel in its time of creation, but its novelty would 'wash-out' with many similar inventions subsequently follow it over time.

To capture this time variance we consider z-scores specific to time-ordered subsets of the entire data. We choose cumulatively increasing subsets in yearly steps, letting $P(t)$ be the sub-collection of patents up to the year t in P. So $P(2000)$ contains all patents issue up to the year 2000, and the z-scores calculated using this set are specific to this year. Thus for a given year, the newly added patents' z-scores are discerned based on all the patents that precede them, and the older patents' z-scores continue to evolve and change based on subsequently issued inventions.

3.4 A schematic for three cases: atypical, typical and neutral

We provide a schematic to aid in our understanding of how atypicality embedded in code combinations is captured and expressed by z-score measure.

Suppose 40 inventions P at time t, indexed by its entering order i. Each invention is expressed as a combination of codes:

$$\begin{aligned} P(t) &= \{P_1, P_2, P_3, \ldots, P_{40}\} \\ &= \big\{\{A, C\} \times 19, \{C, D\}, \{B, E, F\} \times 4, \\ &\quad \{B, E\} \times 6, \{E, F\} \times 5, \{B, F\} \times 5\big\}. \end{aligned}$$

Figure 1 illustrates the collection of patents at t, $P(t)$, represented as a network structure where pairwise combinations are represented as weighted links (solid lines). We then consider three cases where a new patent $P(t+1)$ arrives with two codes, that is, (i) $P(t+1) - P(t) = \Delta P(t) \supset \{A, B\}$, denoted as a black dash line, (ii) $\Delta P(t) \supset \{B, E\}$, as a red dash line, or (iii) $\Delta P(t) \supset \{A, D\}$, as a green dashed line. Simply put, links are solid when they are present at time t, and dashed when they are added at time $t+1$.

In the case (i), the link a bridges the two most frequently used codes A and B that are yet combined together until time t. We therefore find the appearance of link *a atypical* given the current statistics, and naturally expect a *negative* z-score. Indeed, calculated z_a exhibits a negative value, that is, –4.3, with $\mu_{AB} = 7.8$, and $\sigma_{AB} = 1.6$ in the Eq. (1). Note that the frequency of A and B are, respectively, $n_A(t+1) = 20$ and $n_B(t+1) = 16$. On the other hand, the link b reinforces the existing pair that are already well connected, or established, hence, becoming a *convention*, yielding a positive z-score, 3. This indicates that they are combined more than expected by three times of standard deviation derived by the random

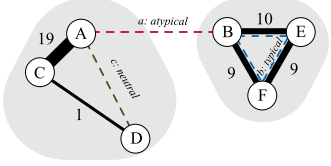

Figure 1 The shades guide two partitioned segments in a network representation. The solid lines denote the existing weighted pairs up to t while the dotted lines (a, b, c) possible additions introduced by newly created combinations (a set of codes used in a patent) at $t+1$. Depending on where a new combination is overlaid in the network structure, the invention at $t+1$ consists atypical, typical or neutral pairs relative to the random chance.

choices. Finally, the link c yields a statistically *neutral* pair, around zero, indicating the occurrence is indistinguishable from the random configurations.

This method was employed successfully by Uzzi *et al.* [8] for academic paper citation rather than classification codes for inventions. It is worth noting the difference that technology code combinations in patent records bring different implications than citation relationships do, in that technology codes define full uses of discrete units of technological capabilities to create an output; every code represents a definite constitute of the whole. In contrast, paper citations can imply a much broader range of relations, for instance, direct reliance on the previous work, to show correspondence with previous results, to negatively point out flaws in the cited paper, and parallel methods applied to a different subject. Thus when it comes to an invention, we believe codes would more accurately capture the parts of an invention.

3.5 Coarse-graining over classification codes

The classification codes are created in a nested structure (hierarchical tree). The number of classes at the highest level is 473, with 16,087 subclasses at the first level down, and 168,743 codes at the deepest and most detailed level. The codes can be extremely detailed in their content, making results pertaining to specific codes very narrow in scope, or it can be quite broad. For example, the class 257, 'ACTIVE SOLID-STATE DEVICES', has the longest depth up to 16 at the end of which 'Floating gate layer used for peripheral FET (EPO)' and 'Floating gate dielectric layer used for peripheral FET', while the class 245, 'Wire fabrics and structure', has the shortest depth up to 2 at the end of which 'Chain' and 'Coil'. As shown in the above examples, the level of differentiation for two codes in a class can be qualitatively different according to the depth and classes.

In addition, for a code pairing to appear novel the codes must have been used enough, relative to the number of patents, for the null model to predict a high number of co-occurrences by chance, it is then the relative lack of co-occurrences in the actual data that signifies a novel combination. The individual code usages (or frequencies) are far smaller than the total number of patents up to t; $|P(t)|$. This means that the expected co-occurrence value $\mu_{\alpha\beta}$ between two different codes, α and β, becomes increasingly small; $\mu_{\alpha\beta} \ll 1$. If two codes do co-occur, then by definition $o_{\alpha\beta} \geq 1$ thus giving a positive z-score. Hence using fine-grained codes gives us almost entirely positive z-scores and we cannot identify novel combinations.

To create a consistent level of detail in the analysis, and gain broader and more intelligible insights, we coarse-grain over the codes. This method is also employed by Uzzi *et al.* [8], who coarse-grain over individual papers up to the journal level. We look at pairings at the highest and the second highest level of the code hierarchy [36].

We consider each patent as a combination of classes - the highest level. We also consider them as a combination of subclasses, i.e. at one level below the class level, to gain insight in a slightly more fine-grained technology space, as well as for comparison with the class level results. At the code level a patent is never assigned the same code twice, and so no self-pairing is possible. This is not the case for class and subclasses, as multiple codes from the same class are often assigned to the same patent.

Let α_i and β_j represent a code each, where α and β denote the class and i and j represent the rest of the code, specifying the low level detail. Rather than coarse-grain over the data before computing the resultant statistics, we first compute $\mu_{\alpha_i\beta_j}$ and $o_{\alpha_i\beta_j}$ for the

most detailed structure at the code level, and only after which do we coarse-grain these values up the higher levels of the subclass and class pairings. Calculating the observed co-occurrences $o_{\alpha\beta}$ and expected co-occurrences $\mu_{\alpha\beta}$ of a class/subclass pair is carried out by summing over all code pairings that result in the considered class/subclass pairing. Hence for the observed co-occurrences:

$$o_{\alpha\beta} = \left(1 - \frac{\delta_{\alpha\beta}}{2}\right) \sum_{i,j} o_{\alpha_i\beta_j}, \tag{5}$$

where the bracketed term accounts for double counting when the classes considered are the same. Similarly for the expected co-occurrences:

$$\mu_{\alpha\beta} = \left(1 - \frac{\delta_{\alpha\beta}}{2}\right) \sum_{i,j} \mu_{\alpha_i\beta_j} \tag{6}$$

and the variance:

$$\sigma^2_{\alpha\beta} = \left(1 - \frac{\delta_{\alpha\beta}}{2}\right) \sum_{i,j} \sigma^2_{\alpha_i\beta_j} \tag{7}$$

from which using these the class/subclass pair z-score can be calculated.

4 Results

Technological constituents of a patent are translated into a set of pairwise z-scores that characterise its novelty or typicality. In this way, we are able to capture how inventors combine technological units by analysing the summary statistics of technology class and subclass co-occurrences.

In the following sections, we will look at the distribution of z-scores derived from the entire set of patents, and compare it with that of newly created combinations, in each year. Then we will relate the observed compositional features of an invention at the time of its creation to its future impact. All analysis is carried out at both the class and subclass level (one level down from the class) to ensure that our findings and insights are persistent across different levels of detail.

4.1 Decomposition of new combinations

It has been shown that the rate at which inventors create new combinations is invariant, and that they create new ones more often than not [1]. This result alludes to a ceaseless introduction of new ways of combining technological units, and thereby a constant reshaping of technology space. By just considering the number of new combinations occurring, the dynamics of novelty creation looks temporally independent, which conforms with the possibility that new inventions occur at random.

On the contrary, a new combination is not simply concocted by randomly choosing technological units, but, although novel it may be, it is either built on the existing body of knowledge accumulated, or discovered by the expansion of the adjacent possible in the technology space [37–40]. The new combination may not be composed of an entirely novel membership [10]. It may contain a set of codes that have been frequently combined, such that they can be considered as an established unit, or building block [41]. Therefore,

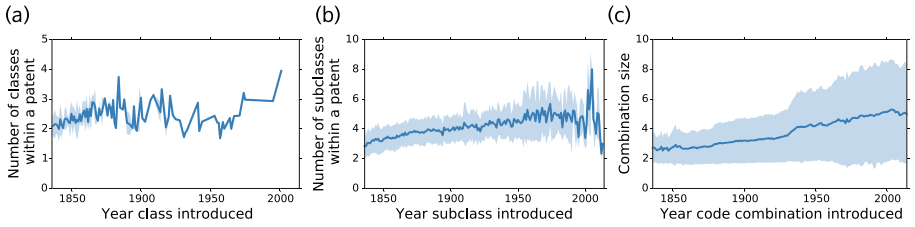

Figure 2 Basic statistics. (a)-**(c)** each *x*-axis indicates year when **(a)** a class, **(b)** a subclass, or **(c)** a code combination was introduced. **(a)**-**(b)** each *y*-axis represents the average number of **(a)** classes and **(b)** subclasses combined in those patents containing a new code. **(c)** *y*-axis displays the number of codes in a new combination. Shaded area denote standard deviation.

binary classification - a combination can be either absolutely novel if it was previously unseen, or otherwise not novel - misses the subtleties of an invention's novelty by lacking the complexity to capture it in any detail such as a combination with small novel addition to the conventional subset.

We decompose combinations into a novelty profile in terms of pairwise z-scores, and assess the extent to which the multiple aspects of a combination are novel or conventional in more detail [8, 9]. In this way, a new combination can both reinforce the current technological conventions, and introduce new ways of combining codes. As elaborated in the Method section, the z-scores are measures to compare the observed occurrences to the random counterpart.

Although there are no limits to how many codes may be assigned to a patent, Figure 2 shows the number of codes in a patent hovers around three to four in average with a tendency that the number increases in time, indicating parsimonious code usages. Every pair within a combination is then assigned a z-score (see, the Method section), and the composition of an invention can be captured by three statistics: its median z_{med} and minimum z_{min}, and the difference between the two $\Delta z \equiv z_{\text{med}} - z_{\text{min}}$ [9]. We used *numpy.median* in a Python library for numerical computations, equivalent to *numpy.percentile* with '$q = 50$' [42]. In the case of three z-scores sorted in ascending order, $z_{ab} < z_{bc} < z_{ac}$, $z_{ab} = z_{bc} < z_{ac}$, or $z_{ab} < z_{bc} = z_{ac}$, the z_{med} is z_{bc}.

The z_{med} indicates the degree to which the main body of a patent conforms to technological conventionality, while z_{min} indicates the extent to which the invention contains an element which is novel when combined with its other parts. The difference between the two, Δz, captures aspects of both in a single measure; whether the patent has a conventional core *and* a novel addition.

Figure 3(a) shows z-score of new class pairings, in which it is seen that their average z-score remain zero, that is, indistinguishable from the random incidence on average, and then gradually become negative after 1980. This implies, new class pairings neither strengthen nor join any modular structures of the class network when they firstly appear, and then new atypical class pairings after 1980 gradually join two different technological domains. In addition, new class pairs gradually being atypical may dispute a claim that 1880s was more innovative than now [43].

When a new pair, or combination was introduced, it is normally the case that its z-score is negative, or neutral as shown in Figure 1. Occasionally however, there is also a case where codes involved in the combination have rarely been used, that the expected co-occurrences, μ, and the standard deviation, σ, is relatively low.

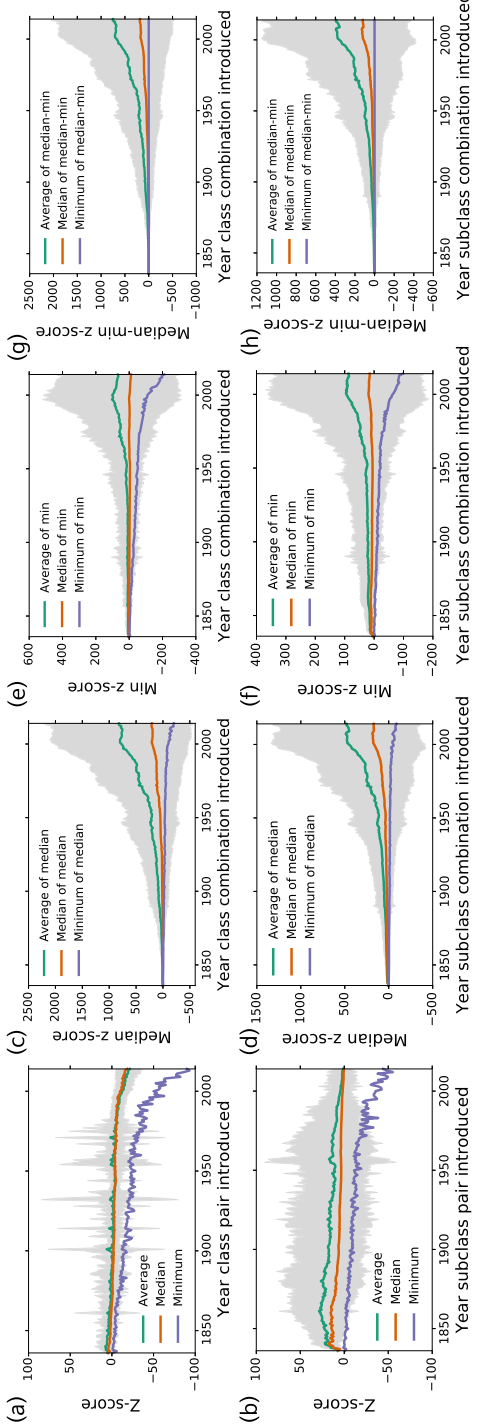

Figure 3 Newly introduced pairs and combinations versus z-score. (a) and **(b)** show *z*-score of a new pair when it was firstly combined. **(c)**-**(h)** each *x*-axis indicates a year that the new combination was firstly used, and each *y*-axis represents z_{med} **(c)**-**(d)**, z_{min} **(e)**-**(f)**, and $z_{\mathrm{med}} - z_{\mathrm{min}}$ **(g)**-**(h)**. **(a)**, **(c)**, **(e)**, and **(g)** are for classes. **(b)**, **(d)**, **(f)**, and **(h)** are for subclasses. Shaded regions represent standard deviation.

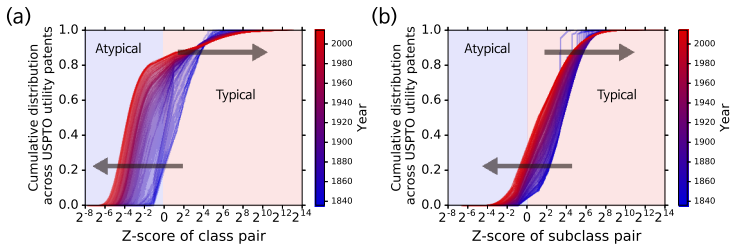

Figure 4 Cumulative distributions of z-scores over year. (a) for class pairs. (**b**) for subclass pairs.

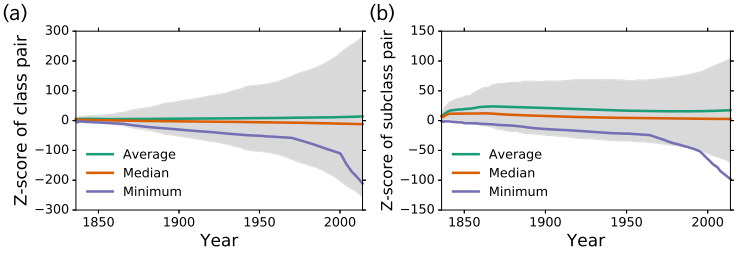

Figure 5 Statistical properties of cumulative distributions of z-scores over year. (a) for class pairs. **(b)** for subclass pairs. Shaded regions represent standard deviation.

We then capture the compositional features of how newly created combination by these two summary statistics of each year. Figure 3(c), (d), and (e) show z_{med}, z_{min}, and Δz. It shows that these z-scores steadily grow over time in a more or less margin of increase. The new combinations mostly contain the conventional pairings. Additionally, the gap between z_{med} and z_{min} increases (Figure 3(g) and (h)). In other words, a new combination becomes both more conventionality and non-conventionality.

We now look at how these compositional features of new pair and new combinations reshape the landscape of technology space. We characterising this phenomena by analysing the distribution of z-scores for entire pairs accumulated up to the year. Figure 4 shows the cumulative distribution of z-scores for every year from 1836 to 2014, respectively denoted by a colour scale (from blue to red). Broadening of this distribution across time indicates that the network is becoming more ingrained, with increasingly highly connected subsets of codes, hence higher z-scores, while pairs that span between these conventional units are thus increasingly perceived as atypical. Thus if two codes are used together more often than random expectation, it is probable that they are used together again. This broadening is also explicitly shown in the Figure 5 where the standard deviation (grey shade) widens over time and the minimum z-scores of the year cohort becomes increasingly negative, especially around the recent decades.

4.2 Compositional features predicting future impact

Predicting which new invention will have a high impact is an obviously wanted goal, both for attempting to predict profitability, but also as a signifier of future societal changes caused by new technologies [12, 32]. Further to just assessing new inventions, an understanding of the qualities related to invention impact enables one to optimise their inventive strategy to maximise such qualities. We show that a patent's success is predictable using its novelty profile.

As we discussed in the previous sections, an invention is interpretable as pairwise z-scores, quantifying statistically significance of code pairings in inventing activities. Built on the previous research, suggesting that the compositional feature is key to a patent having a high impact, we delve into the temporal dynamics of this relationship given our patent records [8, 9].

We define high impact inventions as those patents in the upper 5th percentile,[a] within each year, of citations gained within 5 years from their publication year [32, 33]. We categorise the patents accordingly: whether (i) z_{med} of a patent belongs to either the top quartile z_{med} of a year, middle half, or bottom quartile, (ii) similarly z_{min} in the top quartile, middle half, or bottom, and (iii) Δz in the top, middle, or the bottom. We abbreviated top quartile to high, middle to mid, and bottom to low.

Panels (a) and (d) of Figure 6 show that high z_{med} has a small but positive influence on future impact, and vice versa for low z_{med}, indicating that inventions that are primarily based on established prior work do marginally better in the future. Meanwhile, Figure 6(b) and (e) shows that a high z_{min}, signifying more typical, has a noticeable negative effect on a patents future. Thus if all the pairings of an invention become conventional, it is less likely to be influential. On the other hand, it is evident that when measuring against core conventionality *and* a novel element together, the results are both more consistent and more significant, as seen in Figure 6(c) and (f). The high Δz has a clear positive influence, whereas the mid Δz has no influence and the low Δz has negative influence. Thus these results indicate that it is neither of the two aspects on their own to have the most influence but the combination of the two.

We can further elaborate on this through a differing classification of the patent set: whether (i) z_{med} of a patent is above or below the quartile z-score within the entire period and (ii) z_{min} of a patent is above or below quartile z-score within the entire period. These classifications directly capture (i) whether the patent has a conventional core and (ii) whether it includes a novel aspect. We also redefine high impact inventions as being in the top 5th across *all* the patent records (1976-2001). These criteria split patents into one of four categories: (high z_{min}, high z_{med}) being those patents with a conventional core, but without a novel addition, namely marginal improving; (high z_{min}, low z_{med}) as neither having a conventional core nor a relatively novel addition; (low z_{min}, high z_{med}) as those patents with the success signifier of a conventional core and a novel twist; and lastly (low z_{min}, and low z_{med}) as those patents which are entirely novel, or oddball.

Figure 7(a) and (b) show the 'hit' patent probability of a patent throughout the period depending on the four categories. Instead of quartile that was used for (a) and (b), z_{min} and z_{med} are now chosen to optimise to achieve the highest hit patent probability, resulting the maximum probability as high as almost 9% shown in Figure 7, the full extent of the influence of the categories, with an almost doubling over the background hit patent rate for (c). The results corroborate those in Figure 6(c) and (f); those patents with conventional cores and a novel addition do notable better then the background rate, and do best of all four categories. Also, it is again shown that entirely novel inventions fair the worst. These analyses also suggest that conventionality does not collide over novelty, but conventionality *illuminated* by novelty can help an invention's influence [8, 9].

5 Conclusion

In this paper, we quantitatively studied the novelty distribution of technology pairs, and a connection between the novelty profile of an invention and its future impact, by using

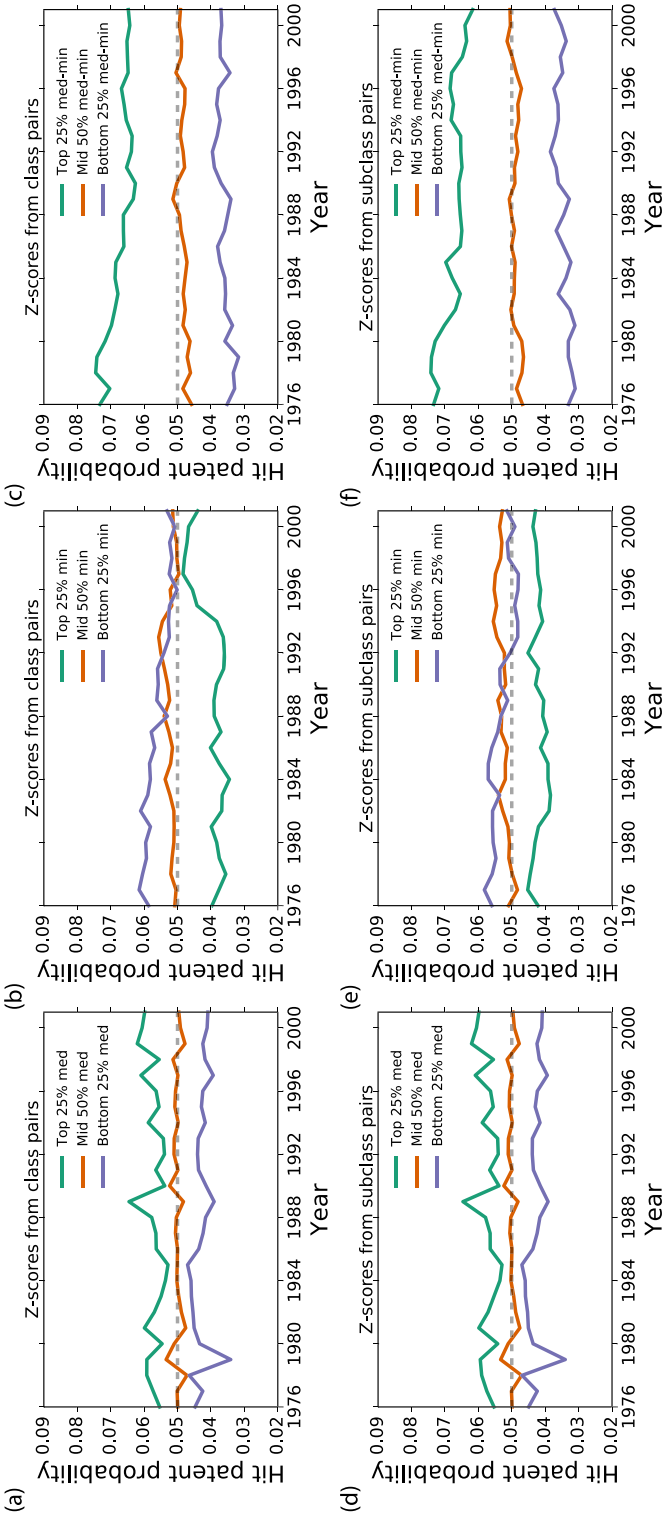

Figure 6 Year versus 'hit' patent probability. (a)-(f) show the probability that a patent is in top 5th percentile by citations among the grant year cohort. Patents are partitioned into three categories by the top and the bottom quartile, and in between, with **(a)** z_{med}, **(b)** z_{min}, and **(c)** Δz. **(a)-(c)** are for classes. **(d)-(f)** are for subclasses.

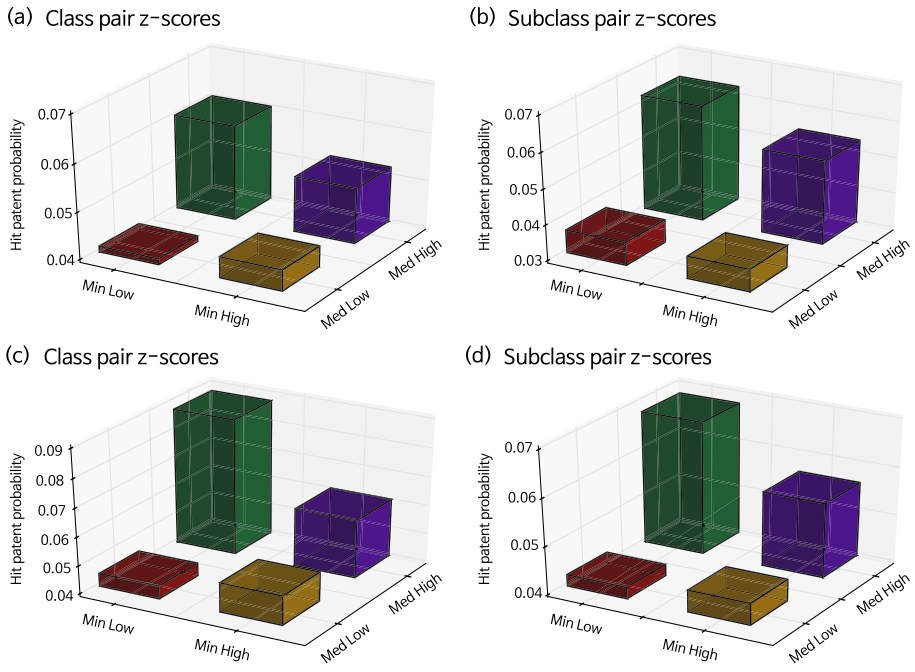

Figure 7 The probability of a 'hit' patent (top 5th percentile), conditional on four categories. The red bar is for low z_{med} and low z_{min}; the green is for high z_{med} and low z_{min}; the gold for low z_{med} and high z_{min}; and the purple for high z_{med} and low z_{min}. The thresholds of *high* and *low* are the quartile for **(a)** and **(b)**; the 95th percentile of z_{med} and the 15th percentile of z_{min} for **(c)**; and the 50th percentile of z_{med} and the quartile of z_{min} for **(d)**. The fraction of patents in each resulted category are: **(a)** 17.66% (red), 7.33% (green), 7.33% (gold), and 67.66% (purple); **(b)** 16.77% (red), 8.22% (green), 8.22% (gold), and 66.77% (purple); **(c)** 14.9% (red), 0.1% (green), 80.1% (gold), and 4.9% (purple); and **(d)** 22.47% (red), 2.5% (green), 27.53% (gold), and 47.47% (purple). The left (right) column are for class (subclass) pairs.

technology code pairings in the U.S. patent spanning 179 years (1836-2014) [27]. We show inventions assemble technological units in a way to reinforce the already conventional pairs, thereby some components become increasingly entrenched within the inventive repertoire with increasing z-scores, such that they become a further building block for future combinations. Yet still combinations will occasionally bridge between these code-cliques, a set of codes frequently co-occurred together, exhibited as increasingly negative z-scores in time.

This result implies that the technology space forms units of tightly co-occurring codes with occasional inter-unit combinations to change that structure, and that inventors always require components which are familiar to them, or available in the industry [38, 39, 44–46].

We also show how technological composition can effect the future impact of an invention, by associating the patents' citation count as a measure of that impact [30]. Through analysis of citation relationships across the U.S. patents (1976-2006), our analysis shows the statistically significant technology pairings are correlated with future influence of an invention. In line with the previous research, our findings demonstrate that conventional combinations, enlightened by proper novelty, are more likely to be influential in future, alluding to that there is an optimal balance between conventionality and novelty for influential inventions, and that influence is associated with knowledge transfer between technological domains [8–10, 47].

Yet there still remains much research to be resolved to more rigorously quantify statistical significance of code pairings. The proposed z-score measure to capture novelty profile within an invention is limited. First, it does not account for inventions of a single code, not to mention codes that appear first time. When a new code is created it must mark an extreme novelty of invention, but the current z-score measure does not capture this attribute, by its nature (null model does account for newly created codes). In our paper, during the periods that our data analysis relates novelty to impact spans does not contain many newly created classes (only two classes are created, making up 0.4% of classes), although the fine detailed codes continue to be created within the period (3% of subclasses were added during the time), and our coarse-graining method over the codes up to the classes level (Section 3.5) mitigates the effects of creations underneath. This leaves much room to improve a quantitative assessment of novelty measure to capture such extreme novelty introduced by inventions in a more systematic and clear way.

In addition, it is worth noting that excluding citations to outside the data or academic papers may miss the important role of scientific research in guiding inventors to search the technological space more efficiently, hence resulting highly novel content [48–50], data of which can be complemented in the future research. Nonetheless, our study may provide valuable insights into how technology combinations give rise to boundary-spanning breakthroughs in technology as well as science, and how innovative technology combinations become influential. Furthermore, our approach has potential in other creative activities beyond scientific knowledge and inventions such as culinary, garment, and journey combinations [18, 35, 51–56].

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
DK, DC and HY participated in all methodological decisions. DK collected and preprocessed the data, and simulation. DK, DC and HY analysed the result. The manuscript was written by DK, DC, HJ and HY. All authors read and agreed on the final version.

Author details
[1]Institute for New Economic Thinking at the Oxford Martin School, University of Oxford, Oxford, OX2 6ED, United Kingdom. [2]Natural Science Research Institute, Korea Advanced Institute of Science and Technology, 291 Daehak-ro, Daejeon, 34141, Republic of Korea. [3]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA. [4]Department of Physics, Korea Advanced Institute of Science and Technology, 291 Daehak-ro, Daejeon, 34141, Republic of Korea. [5]Asia Pacific Center for Theoretical Physics, 67 Cheongam-ro, Pohang, 37673, Republic of Korea. [6]Institute for the BioCentury, Korea Advanced Institute of Science and Technology, 291 Daehak-ro, Daejeon, 34141, Republic of Korea. [7]Mathematical Institute, University of Oxford, Oxford, OX2 6GG, United Kingdom.

Acknowledgements
DK, DC and HY acknowledge the support of the National Science Foundation (no. SMA-1312294). This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (No. 2011-0028908) (DK). This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2012-S1A3A-2033860) (HJ). DK, DC and HY acknowledge the support of Institute for New Economic Thinking at Oxford Martin School, Santa Fe Institute, and CABDyN. The authors thank James McNerney and François Lafond for helpful comments on an earlier version of the paper.

Endnote
[a] The upper 5th percentile can be considered as a statistical significance with p values of 0.05.

Received: 16 November 2015 Accepted: 24 February 2016 Published online: 10 March 2016

References
1. Youn H, Strumsky D, Bettencourt LMA, Lobo J (2015) Invention as a combinatorial process: evidence from US patents. J R Soc Interface 12:20150272
2. Schumpeter JA (1939) Business cycles. McGraw-Hill, New York
3. Nelson RR (1993) National innovation systems: a comparative analysis. Oxford University Press, London

4. Page SE (2008) The difference: how the power of diversity creates better groups, firms, schools, and societies. Princeton University Press, Princeton
5. Evans JA, Foster JG (2011) Metaknowledge. Science 331:721-725
6. Wagner A, Rosen W (2014) Spaces of the possible: universal Darwinism and the wall between technological and biological innovation. J R Soc Interface 11(97):20131190
7. Corominas-Murtra B, Goñi J, Solé RV, Rodríguez-Caso C (2013) On the origins of hierarchy in complex networks. Proc Natl Acad Sci USA 110(33):13316-13321
8. Uzzi B, Mukherjee S, Stringer M, Jones B (2013) Atypical combinations and scientific impact. Science 342:468-472
9. Della Malva A, Riccaboni M (2015) (Un) conventional combinations: at the origins of breakthrough inventions. doi:10.2139/ssrn.2610562
10. Strumsky D, Lobo J (2015) Identifying the sources of technological novelty in the process of invention. Res Policy 44(8):1445-1461
11. Yun J, Kim P-J, Jeong H (2015) Anatomy of scientific evolution. PLoS ONE 10(2):e0117388. doi:10.1371/journal.pone.0117388
12. Farmer DJ, Lafond L (2015) How predictable is technological progress? ArXiv preprint. arXiv:1502.05274
13. Wang J, Thijs B, Glänzel W (2015) Interdisciplinarity and impact: distinct effects of variety, balance, and disparity. PLoS ONE 10(5):e0127298. doi:10.1371/journal.pone.0127298
14. Sinatra R, Deville P, Szell M, Wang D, Barabási A (2015) A century of physics. Nat Phys 11(10):791-796
15. O'Neale DRJ, Hendy SC (2012) Power law distributions of patents as indicators of innovation. PLoS ONE 7:e49501. doi:10.1371/journal.pone.0049501
16. Bettencourt LMA, Samaniego H, Youn H (2014) Professional diversity and the productivity of cities. Sci Rep 4: 5393. doi:10.1038/srep05393
17. Youn H, Bettencourt L, Lobo J, Strumsky D, Samaniego H, West GB (2016) Scaling and universality in urban economic diversification. J R Soc Interface 13:20150937. doi:10.1098/rsif.2015.0937
18. Jagmohan A, Li Y, Shao N, Sheopuri A, Wang D, Varshney LR, Huang P (2014) Exploring application domains for computational creativity. In: Proceedings of the fifth international conference on computational creativity
19. Calcagno V, Demoinet E, Gollner K, Guidi L, Ruths D, de Mazancourt C (2012) Flows of research manuscripts among scientific journals reveal hidden submission patterns. Science 338:1065-1069
20. Rosvall M, Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. Proc Natl Acad Sci USA 105:1118-1123
21. Börner K (2015) Atlas of knowledge: anyone can map. MIT Press, Cambridge
22. Eom Y, Fortunato S (2011) Characterizing and modeling citation dynamics. PLoS ONE 6(9):e24926. doi:10.1371/journal.pone.0024926
23. Wang D, Song C, Barabási A-L (2013) Quantifying long-term scientific impact. Science 342:127-132
24. Deville P, Wang D, Sinatra R, Song C, Blondel VD, Barabási A-L (2014) Career on the move: geography, stratification, and scientific impact. Sci Rep 4:4770. doi:10.1038/srep04770
25. Petersen AM, Fortunato S, Pan RK, Kaski K, Penner O, Rungi A, Riccaboni M, Stanley HE, Pammolli F (2014) Reputation and impact in academic careers. Proc Natl Acad Sci USA 111(43):15316-15321. doi:10.1073/pnas.1323111111
26. USX1I1 - Google patents. https://patents.google.com/patent/USX1/en
27. Electronic bulk data products - USPTO. http://www.uspto.gov/learning-and-resources/electronic-bulk-data-products
28. Description of patent types. http://www.uspto.gov/web/offices/ac/ido/oeip/taf/patdesc.htm
29. United States Patent and Trademark Office (2012) Overview of the U.S. patent classification system (USPC) Washington, DC
30. Hall BH, Jaffe AB, Trajtenberg M (2001) The NBER patent citation data file: lessons, insights and methodological tools. NBER working paper 8498
31. Patent data project national bureau of economic research. https://sites.google.com/site/patentdataproject/Home
32. Benson CL, Magee CL (2015) Quantitative determination of technological improvement from patent data. PLoS ONE 10(4):e0121635. doi:10.1371/journal.pone.0121635
33. Valverde S, Solé RV, Bedau MA, Packard N (2007) Topology and evolution of technology innovation networks. Phys Rev E 76:056118. doi:10.1103/PhysRevE.76.056118
34. Tibély G, Pollner P, Vicsek T, Palla G (2013) Extracting tag hierarchies. PLoS ONE 8(12):e84133. doi:10.1371/journal.pone.0084133
35. Ahn Y-Y, Ahnert SE, Bagrow JP, Barabási A-L (2011) Flavor network and the principles of food pairing. Sci Rep 1:196. doi:10.1038/srep00196
36. McNamee RC (2013) Can't see the forest for the leaves: similarity and distance measures for hierarchical taxonomies with a patent classification example. Res Policy 42(4):855-873. doi:10.1016/j.respol.2013.01.006
37. Tria F, Loreto V, Servedio VDP, Strogatz SH (2014) The dynamics of correlated novelties. Sci Rep 4:5890
38. Arthur WB (2011) The nature of technology: what it is and how it evolves, reprint edn, Free Press, New York
39. Kauffman SA (1996) Investigations: the nature of autonomous agents and the worlds they mutually create. Santa Fe Institute
40. Alstott J, Triulzi G, Yan B, Luo J (2015) Mapping technology space by normalizing technology relatedness networks. ArXiv preprint. arXiv:1509.07285
41. Hidalgo CA, Hausmann R (2009) The building blocks of economic complexity. Proc Natl Acad Sci USA 106(26):10570-10575. doi:10.1073/pnas.0900943106
42. numpy.median. http://docs.scipy.org/doc/numpy-1.10.1/reference/generated/numpy.median.html
43. Smil V (2015) The miraculous 1880s. IEEE Spectr 52(7):26. doi:10.1109/MSPEC.2015.7131688
44. Arthur WB, Polak W (2006) The evolution of technology within a simple computer model. Complexity 11(5):23-31. doi:10.1002/cplx.20130
45. Thurner S, Klimek P, Hanel R (2010) Schumpeterian economic dynamics as a quantifiable model of evolution. New J Phys 12(7):075029. doi:10.1088/1367-2630/12/7/075029
46. Klimek P, Hausmann R, Thurner S (2012) Empirical confirmation of creative destruction from world trade data. PLoS ONE 7(6):e38924. doi:10.1371/journal.pone.0038924

47. Fleming L (2001) Recombinant uncertainty in technological search. Manag Sci 47(1):117-132. doi:10.1287/mnsc.47.1.117.10671
48. Fleming L, Sorenson O (2004) Science as a map in technological search. Strateg Manag J 25(8-9):909-928. doi:10.1002/smj.384
49. Koh H, Magee CL (2008) A functional approach for studying technological progress: extension to energy technology. Technol Forecast Soc Change 75(6):735-758. doi:10.1016/j.techfore.2007.05.007
50. McNerney J, Farmer JD, Redner S, Trancik JE (2011) Role of design complexity in technology improvement. Proc Natl Acad Sci USA 108(22):9008-9013. doi:10.1073/pnas.1017298108
51. Wagner C, Singer P, Strohmaier M (2014) The nature and evolution of online food preferences. EPJ Data Sci 3(1):38. doi:10.1140/epjds/s13688-014-0036-7
52. Spence C, Wang QJ (2015) Wine and music (I): on the crossmodal matching of wine and music. Flavour 4(1):34. doi:10.1186/s13411-015-0045-x
53. Pinel F, Varshney LR (2014) Computational creativity for culinary recipes. In: CHI'14 extended abstracts on human factors in computing systems. ACM, New York, pp 439-442. doi:10.1145/2559206.2574794
54. Pinel F, Varshney LR, Bhattacharjya D (2015) A culinary computational creativity system. In: Computational creativity research: towards creative machines. Atlantis thinking machines, vol 7, Springer, Berlin, pp 327-346
55. Liu Q, Chen E, Xiong H, Ge Y, Li Z, Wu X (2014) A cocktail approach for travel package recommendation. IEEE Trans Knowl Data Eng 26(2):278-293
56. Lin Y, Kawakita Y, Suzuki E, Ichikawa H (2012) Personalized clothing-recommendation system based on a modified Bayesian network. In: 2012 IEEE/IPSJ 12th international symposium on applications and the internet (SAINT). IEEE, New York, pp 414-417