GREThA
Groupe de Recherche en
Économie Théorique et Appliquée

The Right Job and the Job Right: Novelty, Impact and Journal Stratification in Science

Nicolas CARAYOL

GREThA, CNRS UMR 5113, University of Bordeaux, France

nicolas.carayol@u-bordeaux.fr

&

Agenor LAHATTE

Observatoire des Sciences et Techniques of the HCERES, France

agenor.lahatte@hceres.fr

&

Oscar LLOPIS

Department of Management 'Juan Jose Renau Piqueras', University of Valencia, Spain

and

Rennes School of Business, Rennes, France

oscar.llopis@uv.es

Cahiers du GREThA

n° 2019-05

March

GRETHA UMR CNRS 5113

Université de Bordeaux

Avenue Léon Duguit - 33608 PESSAC - FRANCE

Tel : +33 (0)5.56.84.25.75 - Fax : +33 (0)5.56.84.86.47 - www.gretha.fr

The Right Job and the Job Right: Novelty, Impact and Journal Stratification in Science

Abstract

Though Science is traditionally associated with creative behavior, concerns have been raised on its professional procedures being suÿciently open to innovative re-search. Thanks to a new measurement of novelty based on the frequencies of pair-wise combinations of article keywords calculated on the set of all research articles published from 1999 to 2013 in the journals referenced by the WoS (more than ten million papers), we find no evidence of shrinking novelty in science over that pe-riod. Novel contributions are more often performed in larger teams that span more institutional boundaries and geographic areas. High novelty increases citations by more than forty percent and the odds of a "big hit" by about fifty percent. High novelty simultaneously reduces citational risk conditioned on being published to a large extent because it rises the odds of the problem remaining active in the future. As we document that novel papers match preferentially with top journals (even controlling for journal quality), the risk induced by novel research is more likely to materialize through the publication process.

Keywords: Novelty; Web of Science; Scientific Creativity; Journal Stratification; Science

The Right Job and the Job Right: Nouveauté, Impact et Stratification des Journaux dans la Science

Résumé

Alors que la Science est généralement associée à la créativité, des inquiétudes ont été soulevées quant à l'ouverture de ses procédures professionnelles vis-à-vis à de la recherche innovante. Grace à un nouvel indicateur de nouveauté basé sur les fréquences des combinaisons de mots-clefs calculées sur tous les articles de recherche (publiés entre 1999 et 2013 par les journaux référencés dans le WoS, plus de dix millions d'articles), nous ne trouvons aucun indice d'une décroissance de la nouveauté sur cette période. Les contributions plus nouvelles sont plus souvent réalisées dans par des équipes plus larges, qui s'affranchissent plus des frontières institutionnelles et géographiques. Un fort degré de nouveauté accroît les citations reçues de plus de quarante pourcent et la probabilité d'un "big hit" d'environ cinquante pourcent. La forte nouveauté réduit simultanément le risque associé aux citations conditionnellement à la publication dans une grande mesure car cela augmente la probabilité que le « problème » reste « actif » dans le futur. Nous montrons que les articles les plus novateurs s'apparient préférentiellement avec les journaux les plus en vue (même en contrôlant pour la « qualité » du journal). Ceci suggère que le risque associé à une recherche plus nouvelle se matérialise plutôt dans le processus conduisant à la publication.

Mots-clés: Nouveauté; Web of Science; Créativité Scientifique; Stratification des Journaux; Science

JEL: O31; C78

Reference to this paper: CARAYOL Nicolas, LAHATTE Agenor, LLOPIS Oscar (2019) The Right Job and the Job Right: Novelty, Impact and Journal Stratification in Science. *Cahiers du GREThA*, n°2019-05.

http://ideas.repec.org/p/grt/wpegrt/2019-05.html.

GRETHA UMR CNRS 5113
Université de Bordeaux
Avenue Léon Duguit - 33608 PESSAC - FRANCE
Tel : +33 (0)5.56.84.25.75 - Fax : +33 (0)5.56.84.86.47 - www.gretha.fr

1 Introduction[1]

Novelty is central to the whole scientific enterprise, not much as a goal or something to maximize, but more as the locus of a fundamental tension. This is illustrated in the intellectual legacy of R. K. Merton's as novelty or creativity are *not* explicit dimensions of the "*scientific ethos*"[2] (Merton, 1942), whereas they are key to the "*priority rule*" and central in the "*reward system of science*" (Merton, 1957) which provides clever incentives for the advancement of scientific knowledge (Dasgupta and David, 1994). Though that tension exacerbates in times of intellectual discontinuities ("*scientific revolutions*" according to Kuhn, 1962, arising between "*normal science*" eras), it enlivens the everyday professional life of most academic researchers.

Despite the ubiquity of the novelty tension in science, still little is known empirically on its global and evolving role in contemporary science, to the exception of few studies such as Uzzi et al. (2013).[3] How is novelty distributed in today's science? Does novelty pays off in terms of academic impact? Is it more risky? Which outlets publish more novel papers and why? As scientific outcomes double every ten to twenty years (Price, 1961; Olesen Larsen and von Ins, 2010), as teams' size enlarges (Jones, Wuchty and Uzzi, 2008) and as individuals increasingly specialize their research (Jones, 2009), do scientific communities maintain their standards of creativity and originality? The present paper offers a comprehensive study of novelty in science that addresses those questions. It builds upon a new measurement of novelty based on the frequency of pairwise combinations of Author Keywords. This allows us to provide new evidence on the evolution, the distribution, the citational impact and the matching with scientific outlets of all research articles issued in journals indexed by the Web of Science (WoS) over the 1999-2013 period.

The issues at stake may extend far beyond science itself as the degree of academic novelty is likely to affect the contribution of scientific knowledge to the society and the economy, in particular in a context of a deepening division of innovative labor between academia and industry (Arora, Belenzon and Patacconi, 2018). On the one hand, putting too much emphasis on novelty may have unintended consequences. Equating 'good science' to 'novel science' might damage proofs and falsification studies, thus worsening a potential reproducibility crisis in science (Baker, 2016) and in turn demeaning the social

[1]We would like to thank participants to conferences in Barcelona, Paris, Atlanta, Bath, Copenhagen, Bordeaux, and to seminars in Paris and Melbourne for their comments and suggestions. This research was funded thanks to the financial support of the ANR (Grant no ANR-15-CE26-0005). N. Carayol also thanks the support of the USA-France Fulbright Commission and Bordeaux IdEx (Grant no ANR-10-IDEX-03-02).

[2]Merton (1942) rather focuses on opposed norms such as "*disinterestness*", "*communalism*" or "*organized scepticism*" (Merton, 1942).

[3]An important recent paper in the field is also Stephan, Veugelers and Wang (2017) through they study only one publication year.

value of scientific knowledge. On the other hand, a negative bias toward novelty in science could overly discourage scholars from designing innovative research agendas which may result in delayed exploration of new research areas (Carayol and Dalle, 2007) and in subsequent lags in society benefiting from welfare improving discoveries.

Many observers have recently expressed worries that originality and creativity could be under threat in science (Heinze et al., 2009). The peer review system is often criticized for its bias against groundbreaking research (Braben, 2004; Chubin and Hackett, 1990; Wesseley, 1998). A number of scholars have suggested that academic audiences reject novel contributions when they diverge too much from the dominant canon (Trapido, 2015; Shadish et al., 1995).[4] Kolata (2009) quotes a past acting director of the NIH who, after noting that the review system for grant proposals "*works over all pretty well, and is very good at ruling out bad things*", makes the point that the "*system provides disincentives to funding really transformative research*".[5] In the absence of a systematic analysis of novelty in science and its recent evolution, any statement on generic consequences of potentially declining (or rising) incentives for breakthrough remain however essentially speculative. The purpose of this article is to contribute to filling this gap.

The remainder of this article is organized as follows. In the next section, we discuss the notion of novelty and develop an original indicator of scientific novelty based on the frequency of pairwise combinations of Author Keywords. Our indicator is intended to capture the novelty of the most original angle of an article. We precisely study its behavior and show to what extend it differentiates from other measurements of novelty. In the third section, we study the evolution of novelty in science and we elaborate on the contexts in which novel research is undertaken, in particular the characteristics of the research team. The fourth section investigates the relation between novelty and impact in a number of dimensions. After evidencing the raw positive relation between novelty and citations, we assess the predictive nature of novelty on scientific impact and excellence, and explore the returns for being new and 'just before the crowd'. We also intend to

[4]Bias against novelty has also been discussed in other social contexts. In organizations, breaking with existing norms, rules or paradigms by introducing novel ideas creates tensions and paradoxes (Staw, 1995; Mueller, Melwani and Goncalo, 2012). Even in settings where novelty is depicted as a desirable outcome, employees may decide to 'play it safe' and avoid proposing novel ideas to prevent negative social evaluations from their peers (Yuan and Woodman, 2010; Mueller, Goncalo and Kamdar, 2011). Janssen (2003) points that novel ideas often challenge the established framework of task relationships by disrupting existing norms and routines, which is likely to be a source of peer conflict in organizations. Further, the uncertainty and questions about the practicality, usefulness and reproducibility of an idea increase with its degree of novelty (Amabile, 1996), thus making it more difficult for highly novel ideas to gain sufficient acceptance, resources and support to be effectively implemented (Baer, 2012).

[5]In fact, such tension between novelty and conservatism is not new, though as once famously stressed by Max Planck: "*a new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die, and a new generation grows up that is familiar with it*" (Planck, 1950).

better appreciate the individual citational returns to choose novel problems, and the relation between that choice and risk. In the fifth section, we investigate the role played by journals, in particular their stratification, in sustaining novelty. We show how journals stratification affects the relation between novelty and risk. The last section concludes and offers potential implications derived from our findings.

2 Measuring Novelty in Science

In this section, we firstly discuss how novelty has been previously defined and empirically captured in the literature and expose our forward looking notion of novelty. We then introduce a new indicator that intends to capture this idea based on the frequencies of pairwise Author Keywords combinations. In the second subsection, we study the behavior of the index on publication data, and compare it to the behavior of neighboring benchmark indexes.

2.1 An approach based on Pairwise Keyword Combinations

From Atypical Combinations to New Research Questions Henri Poincaré, an exceptionally productive and creative mathematician, firstly introduced the idea that invention in mathematics proceeds from re-combinations of distinct pre-existing ideas ("*mathematical entities*") in one's mind (Poincaré, 1910). Weitzman (1998) proposes a mechanism for the growth of ideas in the economy that results from binary random combinations of existing ideas. Building on this idea, Uzzi et al. (2013) employs pairwise journal co-citations in reference lists of articles to identify recombinations of previous knowledge. The degree of "*atypicality*" or conventionality of those re-combinations are computed through their frequencies of occurrence over the whole period. The authors find that high-impact articles are more likely to combine infrequent pairwise combinations of journal references with conventional ones.

Atypical combinations are not all fruitful, however. Creativity, as Poincaré himself also argued, "*consists precisely in not making useless combinations and in making those which are useful and which are only a small minority. Invention is discernment, choice*" (Poincaré, 1910, p. 325). In this process, "*the role of the preliminary conscious work [...] is evidently to mobilize certain of these [pre-existing mathematical entities], to unhook them from the wall and put them in swing... our will did not choose them at random; it pursued a perfectly determined aim*" (Poincaré, 1910, p. 333-334).[6] Even though serendip-

[6]This view is also reminiscent of the concept of "*abduction*" developed by Charles S. Peirce, which is a form of inference capable of generating new ideas as it proceeds from the effects to the causes, in the hope of generating fruitful new hypotheses.

ity, intuition or chance are obvious factors of breakthrough, fruitful combinations result from intentional exploration behaviors (in contrast to exploitation, March, 1991), a form of "*tinkering*" (a term popularized by Jacob, 1977) by which researchers intentionally address new scientific problems or questions by creatively assembling and re-purposing methods and concepts picked in the literature.

To identify the research directions of research articles (the very problem they address) we use the keywords given by the authors themselves, the ones they freely chose to describe their contribution. We suggest that pairwise keyword combinations capture the different "*angles*" of a scientific paper, to employ a notion introduced by Jacob: *"scientific advances often come from uncovering a hitherto unseen aspect of things as a result, not so much of using some new instrument, but rather of looking at objects from a different angle"* (Jacob, 1977, p. 1161). We focus on the most infrequent pairwise combination of keywords of the paper as capturing its novelty, its most original "angle".[7]

Pairwise Keyword Novelty Index We consider all pairwise keywords combinations by papers published in a given year and research field. Keyword combination frequencies are calculated within fields since the degree of novelty of a publication is likely to be interpreted within a given field or community, not across fields. Moreover, some terms can be interpreted differently across communities.

Formally, the commonness of keywords i and j combination, in field $c \in C$ the set of all fields, and year t, is computed as follows:

$$Com_{ijct} = \frac{N_{ijct}/N_{ct}}{\frac{N_{ict}}{N_{ct}} \times \frac{N_{jct}}{N_{ct}}} = \frac{N_{ijct} \times N_{ct}}{N_{ict} \times N_{jct}}, \tag{1}$$

with N_{ct} the number of (non-distinct) keyword combinations in papers published in field c and year t. The terms N_{ict}, N_{jct} and N_{ijct} give the number of articles which use respectively keyword i, keyword j, and both keywords i and j.[8] Equation (1) manifests itself simply as the share of keyword pairwise combinations that use i and j in the domain c, divided by the expected share of such pairs given the number of times keywords i and j are used in c.

When articles have keywords, they are likely to have more than two, which is the minimal possible value for pairwise keyword frequencies to be calculated. For a given article a, let K_a denote its set of distinct (unordered) pairs of keywords and C_a ($\subset C$) its

[7] In fact, we use (see below in this section) a combination of the ninetieth percentile (within fields) and of the maximum (across fields). However, we show that variations in a number of dimensions on the precise way of computing the novelty index makes no difference in the results.

[8] Some Author Keywords may be misreported. Since most of the errors are very rare, we drop all keywords that appear in only one document, and thus all combinations implying at least one of these keywords.

set of associated fields. We focus on the value of the 10th percentile of the distribution of pairwise commonness values as follows:

$$com_c(a) = 10th\text{Percentile}\left(Com_{ijct} \,|\forall ij \in K_a\right), \tag{2}$$

for each paper a and field $c \in C_a$. We use the tenth percentile because it avoids extreme value problems we would have encountered had we taken the minimum. But the underlying idea is very similar.We want to select the least common combination of keywords, that is the most original "*angle*" of the paper, in each field. In Equation (2), the indexing of commonness to the publication year t drops out as it is unambiguous and unnecessary for what follows.

We then use the inverse logarithmic transformation of commonness to obtain the novelty of paper a in field c:

$$nov_c(a) = -log\, com_c(a) \tag{3}$$

Since journals, and thus articles, may be attached to multiple fields, we attribute the maximal novelty over all associated fields:

$$nov(a) = \max_{c \in C_a} nov_c(a). \tag{4}$$

This means the field of research in which the novelty is to be considered is the one in which it is found to be the most original.

The definition of any quantitative indicator imposes computational choices that balance various goals (robustness, clearness, simplicity...) besides effectively capturing the phenomenon that it intends to measure. We however do not want any of the results to depend on those choices. Thus we consider a number of alternative specifications for Equations (1) to (4).[9] We show later in the paper and in the Online Appendix that the results of the paper are robust to substituting any of those variants to the main Pairwise Author Keywords Novelty indicator.

2.2 Data, Pairwise Author Keywords Novelty Index and Benchmarks

In this subsection, we first study the behavior of the proposed index of novelty, before successively comparing it to two benchmark indexes. We make this comparison to show it is important to use pairs of Author Keywords rather than pairs of predefined keywords or pairs of journal references to identify top novelty.

[9]Detailed explanations of the variants specifications are presented in Online Appendix A.

Data and the Behavior of the Pairwise Keywords Novelty Index Our dataset groups together all research articles published from 1999 to 2013 and indexed in the Web of Science (WoS).[10] Therefore, review papers, letters and conference proceedings in particular are not considered. We use Author Keywords –this choice is important and is discussed later in the present section. Research fields are identified by the subject categories, according to the classification scheme developed by ISI which assigns to scientific journals at least one of the 251 subject categories (and potentially to several ones). As our novelty indicator is based on keyword combinations, we exclude all papers with less than two Author Keywords. The sample comprises 10,229,644 research articles. These papers are classified in three major research areas: humanities and social sciences (7%), life sciences (47%) and hard sciences and engineering (46%).

We display, in Figure 1–Graph (a), the distribution of Pairwise Author Keywords Novelty of the articles. The distribution is "well-shaped", slightly asymmetric. Broad field differences are minor: the distribution is slightly sharper for the sub-sample of papers in humanities and social sciences and the one of life sciences. The novelty of hard sciences and engineering papers is slightly both less concentrated and lower.

For a given paper, as the number of keywords grows, so does the number of pairwise keyword combinations for purely mathematical reasons. Lets us recall that our novelty indicator is defined as the maximum over research fields of each field 90th percentile novelty among all pairwise combinations of keywords (see Equations 2–4) which often coincides with taking the max. Therefore, a positive correlation between novelty and the number of keywords is expected. This is verified on the data as the correlation coefficient between Pairwise Author Keywords Novelty and the number of keywords equals .37. All correlation coefficients are presented in Table 1. Figure 11 in the Online Appendix shows however that, if Pairwise Author Keywords Novelty is positively related to the number of keywords from the first to the third quartile, above the third quartile, increasing further the number of keywords no longer correlates with changes in novelty. Broad fields differences are not behind such results as they extend to within-broad-fields analyses.

Is there a relation between the degree of novelty of keywords themselves and the degree of novelty of pairwise keyword combinations? Phrased differently, are articles using newer keywords also more likely to use newer pairwise keyword combinations, or are keyword novelty and pairwise keyword novelty substitutes? To answer this question, we create a keyword novelty indicator (see Online Appendix, Equations 7–9) and study its correlation with Pairwise Author Keywords Novelty. We find a significant and negative correlation ($-.23$, Table 6) but which essentially applies to articles having low levels of keyword novelty (three first deciles, see Figure 12 in the Appendix). This negative relation is much less clear with papers that have higher keyword novelty (after the fourth decile),

[10]Data come from the full WoS database provided by ISI Thomson Reuters, maintained and enriched in house at the OST-HCERES.

and, when present, seems mainly driven by the hard sciences and engineering. These results support the idea that pairwise combinations of keywords provide a much more sophisticated indicator of novelty, distinct from keyword novelty.

We also wonder whether our novelty index could be correlated with the research field size. Two alternative proxies for field size are employed: i) the number of articles that were published in each subject category (Figure 13, graph *a*) and ii) the number of articles published in the same journal (Figure 13, graph *b*). In both cases, results suggest the existence of a slightly positive relation between field size and Pairwise Author Keywords Novelty, which is however very limited and mainly explained by field-level differences.

Keywords Chosen by the Authors vs. Assigned Keywords The Web of Science provides two different types of keywords: KeyWords Plus and Author Keywords. KeyWords Plus are assigned to articles by a computer algorithm which extracts words from the titles of the reference list articles (Garfield and Sher, 1993). Author Keywords are given by the authors themselves when submitting their papers. Both types of keywords have been previously employed to identify research trends (Li et al., 2009). Azoulay, Graff Zivin and Manso (2011) and Boudreau et al. (2016) are reliant on a set of predefined keywords provided by Medline database to identify research themes and their novelty, namely the MESH Keywords of the National Library of Medicine's controlled vocabulary thesaurus. Assigned keywords, such as KeyWords Plus or MESH Keywords, have the advantage of being less sensitive to the strategic choices of the authors. We however refrain from using such externally defined categories as keywords provided by the authors themselves are more likely to precisely capture the original idea their paper carries to the community. Comparing the use of KeyWords Plus and Author Keywords in a specific research field, Zhang et al. (2016) argue that KeyWords Plus more often describe methods and techniques whereas Author Keywords are more comprehensive in representing article content.

Actually, throughout this paper, we do document our main results are reliant on choosing Author Keywords instead of predefined keywords such as KeyWords Plus. We calculate the novelty of all papers using KeyWords Plus instead of Author Keywords and study to what extend the two differentiate in the data. We note first that Pairwise Author Keywords Novelty and Pairwise KeyWords Plus Novelty are positively correlated (.41, see Table 1). The originality of authors' choices of keywords is more apparent in the way they associate keywords. Pairwise novelty indexes based on Author Keywords and on KeyWords Plus are positively correlated to a more limited extent (.22). Interestingly, this correlation reverses when focusing on top novel pairs of keywords. The dummy of top 10% Pairwise KeyWords Plus Novelty is negatively correlated with Pairwise Author Keywords Novelty ($-.10$) as well as with the top 10% dummy of this variable ($-.04$). The negative relation rises when increasingly focusing on top Keywords Plus novel papers (top

5%, top 1%). This tells us that top novel papers relying on Pairwise Author Keywords Novelty and according to Pairwise KeyWords Plus Novelty are likely to *not* be the same.

How do the patterns of choosing and associating keywords by authors differ from the ones of the algorithm allocating KeyWords Plus? Any article adds in average .93 new Authors Keywords to the pool and 7.26 new pairs whereas it brings .63 new KeyWords Plus and 14.88 distinct KeyWord Plus pairs. This shows authors' choices of keywords are more idiosyncratic whereas their pairwise association of keywords is less heterogeneous. This apparent paradox may be explained by purely mathematical reasons as the average number of KeyWord Plus pairs in papers is much larger than the number of Author Keyword pairs. There are on average 27 pairs of KeyWords Plus for a given paper, vs. 8.8 Authors Keyword pairs. Figure 1–Graph (d) exposes the distributions of articles with respect to their number of Author Keyword pairs and KeyWord Plus pairs. About 27% of papers have less than 10 pairs of KeyWords Plus (that is five KeyWords Plus), against 80% having less than 10 pairs of Author Keywords. As the average paper has much more KeyWords Plus than Author Keywords, it is more likely that it brings in completely original pairwise combinations (distinct ones) of KeyWords Plus rather than of Author Keywords. This does not mean that the most original combination of KeyWords Plus is more likely to characterize the paper though. On the contrary, the large number of Keywords Plus pairs may introduce noise in the indicator. Some pairs of Keywords Plus may end out being very original for very artificial reasons. The fact that more than 40% of the papers have more than 45 pairs of KeyWords Plus (10 KeyWords Plus), raises concerns on the reliability of the information on the lowest frequency among those pairs. Further, Figure 1–Graph (b) shows that the distribution of Pairwise KeyWords Plus Novelty is much more concentrated than Pairwise Author Keywords Novelty, with a particularly sharp decline just above the mode, thus rendering small differences decisive to be included in the sets of top novel articles. That may lead to an even noisier identification of top novel articles.

Author Keywords are more original though they are less numerous in the average article. Therefore their pairing is much more precise and likely to better characterize the research directions of the paper.

Pairwise Journal References Novelty Benchmark Previous literature has mainly computed the novelty of articles using the frequency of journal reference combinations (Uzzi et al., 2013; Lee, Walsh and Wang, 2015; Wang, Veugelers and Stephan, 2016; Stephan, Veugelers and Wang, 2017). Though we believe an approach based on the frequency of keyword combinations permits a more direct measurement of scientific novelty than one based on journal reference combinations and is more closely aligned with neighboring concepts, such as creativity or knowledge exploration, we also compute a benchmark indicator of novelty based on pairwise journal references combinations. This

indicator is, like ours, time-variant, and close in spirit to the indicator developed by Lee, Walsh and Wang (2015) based on the atypicality of journal reference combinations in year t. The precise computation of this indicator is presented in Online Appendix C (see Equations 10–12).

We expect a positive correlation is expected between a pairwise keyword combinations indicator and alternative indicators based on pairwise journal co-citations, because investigating new research questions may often require combining pre-existing pieces of knowledge in new ways. However, the correlation between our novelty indicator and the indicator based on journal references appears to be positive but rather small (.12, see Table 1). Figure 17 in the Online Appendix shows that Pairwise Author Keywords Novelty increases with Pairwise Journal References Novelty essentially for the lowest deciles of that indicator. The relation is weak and non monotone for all other deciles. This implies that highly novel papers are often not the same depending on which indicator is used. This relationship is confirmed by the correlation analysis involving top Pairwise Journal References Novelty. We find that all such variables (top 10%, 5%, 1%) are negatively correlated with Pairwise Author Keywords Novelty indexes.

Actually, Pairwise Journal References Novelty has a number of common traits with Pairwise KeyWords Plus Novelty. It is sharply distributed around its mode with a very steep slope on its right side as it is apparent from Figure 14–Graph (c).[11] Further, as the median article has 29 references, that is more than four hundred distinct pairwise combinations, we expect this indicator to be noisy as pairwise KeyWords plus novelty. In fact top novel Pairwise KeyWords Plus Novelty and top Pairwise Journal References Novelty are positively correlated, which suggests these dummies capture close phenomena, distinct from Pairwise Author Keywords Novelty. An explanation may be that these two indicators basically hinge on the same source of information because KeyWords Plus are extracted from the titles of the references (Garfield and Sher, 1993).

3 The Evolution, Distribution and Contexts of Novelty in Science

In this section, we first look at the expansion of scientific inquiry as a whole through the evolution of distinct keyword combinations before examining the evolution of the novelty of research articles using the Pairwise Author Keywords Novelty indicator. Finally, we investigate the relation between novelty and team characteristics.

[11] Note that this indicator is subject to significant broad fields differences that render it much less appealing.

3.1 The Expansion of "Scientific Inquiry"

There has been growing concern that, though scientific production is globally increasing, the degree of creativity embedded in published research may be actually shrinking. We document these dynamics by looking at the relation between the growth in the number of research articles and the growth in the explored "knowledge space", approximated by the number of distinct keyword combinations that are used each year. In years 1999 and 2013, about five and fourteen millions distinct keyword combinations are respectively employed, while approximately fifteen times fewer distinct keywords are used and research papers are produced. As the orders of magnitude of the different variables are so different, year 1999 is taken as reference in the graphs of Figure 2 so that our analysis focuses on growth rates.

In the left plot we see that the number of distinct pairs of Author Keywords follows a very similar growth pattern to the number of research articles. Both increase steadily (in log scale) to reach an overall growth of 290% over the thirteen years considered (which corresponds to an average 7.9% yearly growth rate). Note that as the vertical axis is in log scale, a straight line represents to a constant growth rate (exponential growth). These growth patterns contrast with the one of the number of keywords, which exhibits a decreasing growth rate over the period (less than linear growth in log scale).

The right plot of Figure 2 allows us to look at the expansion of scientific knowledge from a slightly different perspective, that is, with respect to the number of research articles – the horizontal axis. The number of keyword combinations used each year relative to year 1999 level very closely follows the straight line of the number of research articles (also relative to the number in year 1999). This confirms that, according to our indicator of novelty, the scientific community does not exhibit a decreasing rate of creativity, though the yearly production of knowledge increased exponentially over the thirteen years under investigation in our database of research articles.

This graph also traces the evolution of the number of possible pairwise combinations of keywords, calculated as half the square of the number of distinct keywords. Linear growth is also observed for the number of possible combinations. If we interpret the number of possible keyword combinations as prospective combinations that the scientific community could explore, we see that this exploration space increases at a rate nearly twice the growth rate in the combinations that are eventually explored. This is consistent with the idea that, as knowledge accumulates and as the frontier expands, the potential pairs to be formed out of the used keywords increase at a significantly greater rate.

The number of distinct keywords is growing sub-linearly with the number of articles. This can be interpreted as a supplementary argument supporting the idea that pairwise keyword combinations provide a better proxy to reflect the exploration of new scientific

areas, as compared to the consideration of keywords only.

3.2 The Evolution of Novelty at the Article Level

Most article-level calculations use data for the 1999-2011 period because later on in this article we need a sufficient time lag to collect three-year forward citations of publications and we want to remain fully consistent throughout the paper. This dataset contains 7,896,301 articles which received more than 26 million citations.[12]

Figure 3–Graph (a) depicts the evolution of mean article novelty for the full sample of publications and by broad field of science. Mean novelty is pretty stable, very slightly increasing over the complete period for the full sample. This overall stability hides a more pronounced increase in the life sciences. However, such variations still remain very limited (between 1% and 2% of variation over the whole period). In times of apparent changes in the way science in performed, the degree of novelty remains stable. This is fully consistent with our results displayed above showing that keyword combinations and articles grow at the same rate. Note that the average number of Author Keywords per paper remain pretty constant over the period so that an underlying decrease in novelty is not hidden by an artefactual increase in the number of keywords.

Life sciences exhibit the highest degree of novelty, and hard sciences and engineering the lowest degree of novelty. This confirms the intuition that life sciences explore a higher number of more diverse problems while hard sciences focus on a smaller number of problems.[13] When we look in more detail at novelty by discipline (Figure 3–Graph (b)) we find that there is a group of disciplines composed of medicine, humanities, fundamental biology, social sciences, chemistry and sciences of the universe that are characterized by a higher level of novelty. Then comes applied biology, followed by physics and engineering which are very close, and then maths. Some disciplines are characterized by an increasing average novelty, such as medicine, chemistry or physics, while others follow a decreasing path such as the sciences of the universe or the humanities. Absolute differences across disciplines should be taken with caution as they may be influenced by field-specific academic norms.

The time evolution of the benchmark indicator (Pairwise Journal References Novelty) indicator depicts a very different pattern (see Figure 16, Appendix Cof the Online Appendix). Pairwise Journal References Novelty increases from 1999 to 2011 by 50%, a

[12]Note that all analyzes and regressions are also performed using a five-year window of forward citations that thus need to restrict to the 1999-2009 period (less than seven million papers).

[13]Similar cross-broad fields differences have been also found in recent studies on interdisciplinarity. For instance, Millar and Dillman (2012) found that life sciences have the highest proportion of interdisciplinary dissertations, and Leahey, Beckman and Stanko (2017) found that biotechnology and medicine were the most interdisciplinary fields.

sharp variation which seems difficult to explain. Such variation concerns all large fields of science, and is more marked for social sciences and humanities, and engineering; and less marked for life sciences. When looking at the degree of novelty by scientific discipline (Figure 16b), we find significant differences from what we found for our indicator based on pairwise keyword novelty. Physics is for instance found to be significantly more novel than the sciences of the universe, the engineering sciences, the social sciences and the humanities. This does not fit with the intuition that physics, as an older science, deals with a more limited number of problems and hence should exhibit lower levels of novelty.

3.3 Which Teams Produce More Novel Papers?

There has been an increasing interest in exploring the factors behind breakthrough scientific contributions. Heinze et al. (2009) looked in particular at the institutional factors. Jones, Wuchty and Uzzi (2008) and Jones (2009) connect radical scientific contributions to age. However, the relation between team composition and novel science has been underinvestigated (an exception is Lee, Walsh and Wang, 2015). With respect to scientific teams, we know that their size is increasing over time, as well as their institutional and geographical span (Adams et al., 2005; Wuchty, Jones and Uzzi, 2007; Jones, Wuchty and Uzzi, 2008; Adams, 2013). The literature mainly explains these evolutions by the falling costs of remote collaboration (thanks to the www and new information technologies) or as a response to knowledge specialization (Jones, Wuchty and Uzzi, 2008). Adams (2013) argues that there is a rising stratification in collaborations, in the sense that the best universities develop longer-range and higher-quality collaborations.

We wonder whether, to deliver more novel science, scientists are more likely to assemble in larger (vs. smaller) teams or to form cross-institutional (vs. within closed walls) collaborations. Our data show (see Figure 4) that novelty increases with team size (approximated by the number of co-authors). This is consistent across all broad scientific fields, with the exception of social sciences and humanities, where there is a decreasing relation after five co-authors. We also find similar results for the number of distinct institutions. Further, novelty increases when the teams involve members located in different regions of the world. All these results support the idea that "break things and think different" is more frequent in larger and more dispersed teams. This could be due either to a "diversity effect" or to a selection effect (better teams are more likely to span boundaries) that can not be easily disentangled –a natural experiment would be in order. Novelty may have been preserved by the increasing size and span of scientific teams. We also find that novelty is higher when articles involve co-authors from North America and, to a lesser extent, Europe. Interestingly, the gap between the novelty of North American teams and that of European teams has widened since 2005.

We perform a series of regressions to consider whether the correlations of the different

team characteristics with our novelty indicator remain controlling for a number of dimensions. The number of keywords, keyword novelty, year dummies as well as sub-domain dummies are included as controls in logistic regressions on top novel paper dummies (top 10%, top 5% and top 1% dummies of Pairwise Author Keywords Novelty). As the different characteristics of the teams are highly correlated, we do not include the number of authors, the number of distinct institutions, and the number of word regions simultaneously in the regressions. These variables are included one at the time, together with with all controls and with world region dummies.

Results can be found on Online Appendix F. They confirm the positive effect of the number of distinct institutions on the odds of a paper to be highly novel. On average, an additional institution participating in a paper increases the odds of a paper being a top 10% highly novel by 5.5%. This remains positive and significant when we split our sample by scientific domains, or when using more restrictive definitions of what a highly novel paper is (top 5% and top 1%). The positive relation between the number of co-authors on the odds of a paper to be highly novel is preserved as well. The effect is larger though, as an additional author, on average, increases the odds of a top 10% novel paper by 10%. Again, results remain consistent by scientific domains and for the different specifications of highly novel articles. Turning to the role of the number of geographical regions, our multivariate analysis comes from logit regressions excluding geographical locations dummies that are strongly correlated to this variable. We find that an additional world region in the team increases the odds of a top novel paper by more than 20%. This is robust to the various definitions of high novelty papers and more pronounced in the hard sciences and engineering. The positive relation between North America and top novelty is also confirmed with an impact on the odd of top novelty ranging from 14 to 20%. Europe as well with an impact ranging from 8 to 15%. The Asia dummy has a positive relation to the odds of top novelty ranging from 4 to 12%. It is thus likely that the initial univariate analysis underestimated the odds of high novelty in Asia, mainly because of its specific distribution of research over scientific disciplines in those countries (in particular in China) that needs to be controlled for, as we do in those regressions.

4 Novelty and Scientific Impact

This section examines whether novel articles are more or less cited (Subsection 4.1), and to what extent novelty can be conceived as a predictor of impact and excellence (Subsection 4.2). We are also interested in the future conventionality of the pairs of keywords used as a predictor of scientific impact and its complementarity/substituability with novelty (Subsection 4.3). We further tackle the individual rewards of novelty by

introducing more controls (Subsection 4.4) and conclude the section by analyzing how novelty relates to the notion of risk (Subsection 4.5).

4.1 Are Novel Articles more Cited?

To explore the relationship between novelty and academic impact, we rank all papers according to their Pairwise Author Keywords Novelty. For each centile of novelty, we calculate the average number of forward citations received over a three-year period.[14] We observe in Figure 5 that citations increase significantly with novelty. A paper in the last centile of Pairwise Author Keywords Novelty receives, on average, two to three times more citations than a paper in the first centile. This applies for each broad field of science taken separately, though to a lesser extent in the social sciences and humanities.

We use the percentile-based approach (Waltman and Schreiber, 2013) to define "big hit" papers, namely the ones which are among the top 10% most cited in their subject category and publication year. Similar work is done for the top 5% most cited articles. This allows us to focus on the most cited papers, knowing that the distribution of citations is very skewed: half of the research articles in our dataset receive (within a three-year time frame) fewer than 2 citations while the top 10% most cited articles receive on average 11 citations more than the "mean" article.

We find (see Figure 5) that the proportion of papers categorized as "big hits" rises with the centiles of novelty. While the proportion of "big hit" papers is slightly over 6% for the lowest centiles of novelty, this rate rises up to 14% for the highest centiles of novelty (again, two to three times more). This result is quite robust across broad scientific fields. Only social sciences and humanities exhibit a somewhat non-linear pattern so that the share of top 10% papers tends to decrease slightly above the seventh to eighth deciles of novelty. Very similar results are found when we define a "big hit" article as a paper in the top 5% of the citation distribution or when looking at 5-year forward citations (see Online Appendix D, Figure 18).

Even if these first results clearly support the idea that keyword combination novelty strongly correlates with citations (and in particular with high impact), it is not yet clear whether novelty is a predictor or lever of academic impact. We address these questions more directly in the following subsection.

4.2 Novelty as a Predictor of Impact and Excellence

Let's consider the following purely conceptual experiment: a scientific team is picked randomly in a given discipline, and one wants to calculate the probability that its ongoing

[14] A 5-year citation window is also employed as a robustness check in Online Appendix D.

research becomes a "big hit", conditional on being published in some journal referenced by the WoS. To what extent does a supplementary conditioning on high novelty of the paper increases future academic impact?

To answer this question, we have performed a series of regressions whose results are summarized in Table 2. Generalized negative binomial regressions allow us to estimate the impact of high novelty (top 10% highest novelty) on citations.[15] We find a 38% impact in the 3-year window. Moreover, logit regressions show that picking a highly novel article increases the chances of a "big hit" (top 10% most cited articles in a 3-year window) by 42%. That effect is slightly larger in the hard sciences and engineering and in the life sciences, and lower in the humanities and social sciences (25%). Similar results hold when considering an additional specification of "big hit" (top 5% most cited). A slight increase in the odds of a "big hit" is observed when citations are considered over a time window enlarged to five years, but the correlation with the number of citations does not rise when citations are recorded over this larger time scale. Interestingly, the incidence rate ratios increase when high novelty is more sharply defined: Citations increase by 50% for top 1% most novel articles (See Tables 86-97 in the Online Appendix).

4.3 Novelty and the Crowd

Uzzi et al. (2013) highlight that citation impact is significantly enhanced when a publication simultaneously couples "high median conventionality" with "high tail novelty". Rather than looking at conventionality and novelty simultaneously, but on different dimensions as they do, the time-variant nature of our novelty indicator allows us to conceive articles' novelty and conventionality in a sequential manner on the same dimension. Our contention is that those papers that anticipate further interest in the very dimension on which they innovate, will have greater impact. In other words, breakthrough contributions in directions that keep attracting interest from colleagues are likely to be much more cited. On the contrary, novelty without a sustained future interest may pay relatively little in terms of citations. It is often argued, based on anecdotal evidence, that science has some common features with Keynes' "beauty contest" idea or with finance, in the sense that it would pay more to address problems in new areas of science which very soon become trendy.

Testing these conjectures with our data implies creating a new variable of future commonness. We create a dummy that equals one if the same keyword combination employed to assess novelty in period t is still used by papers published in periods $t+1$ and $t+2$ in the same field. Otherwise, it takes the value 0. In other words, for a paper to be considered as common in the near future, its most original "angle" at the time of publication needs

[15] In this part, we only discuss the main coefficients of the generalized negative binomial estimation. The determinants of the dependent variable dispersion are discussed in Subsection 4.5.

to remain active in the two following years after it was published. From now onwards, commonness will always refers to this definition without recalling the time lag.

We identify four different scenarios to categorize each article in terms of novelty and commonness: highly-novel-and-common, highly-novel-and-*not*-common, *not*-highly-novel-and-common and, *not*-highly-novel-and-*not*-common. As previously, we employ two different regression methods for estimation, depending on the sort of dependent variable: i) generalized negative binomial regressions for models with a count-based dependent variable (number of publications after 3 years), and ii) logistic regressions for models with a percentile-based dummy dependent variable ("big hits"). Figure 6 summarizes our results on the estimated coefficients. The left-side graph (a) reports coefficient estimates of each considered dummy from generalized negative binomial models, and the right-side graph (b) is based on similar estimates from logistic regressions.[16]

We find that the most successful papers are those that are highly novel at the time of publication (t) and which are published in a field still active later on. Novel-today-and-common-later-on papers receive, on average, 62% (see left-side graph) more citations than papers located at the reference category (*not*-highly-novel-and-*not*-common). When turning to "big hit" papers, estimates results show a similar effect, although even more pronounced. The odds ratios of a paper being a top 10% "big hit" are 74% greater for those highly-novel-and-common papers. This rises up to 82% on top 5% "big hit" papers (see Table 50 in the Appendix).[17]

The dummies' coefficients are the largest for hard sciences and engineering. Publishing a highly novel paper on a research angle that remains active two years after publication receives 70% more citations, and has 79% and 85% more chances of becoming top 10% and top 5% "big hit" papers respectively. The incidence rate ratios are also very large for life sciences where highly-novel-and-common dummy raises the odd of a top 5% "big hit" by 88%.

The papers in the two "mixed" categories, highly-novel-and-*not*-common and *not*-highly-novel-and-common papers have similar performances in terms of citations (approaching 40% more than the baseline). However, in terms of likelihood of becoming a "big hit", the latter category performs significantly better (58% vs. 41% for top 10% papers and 64% vs. 41% for top 5% papers).

Overall, we can observe that novelty without commonness increases citations and the probability of becoming a "big hit" by 40%. This effect is enlarged by 50% on the number of citations and even nearly doubled for the odds of a "big hit" when coupled with future

[16]Detailed regression results can be found in Tables 46 and 47.

[17]Robustness checks employing a 5-year citation frame and additional definitions of top papers were also performed and can be found in the Online Appendix. No qualitatively significant differences were found. Detailed results for the robustness checks can be found in Tables 48, 49, and 51.

commonness. However, it is important to note that the anticipation of future interest is key for citations in general, not specifically for more novel articles. If there is a "beauty contest" reward dimension in science, it is not specific to highly novel papers. In other words, we do not find any complementarity between high novelty and commonness as, jointly, they do not increase citations more than they do separately. Indeed, the sum of the coefficients of highly-novel-and-*not*-common and of *not*-highly-novel-and-common is always more than the coefficient of highly-novel-and-common. In a nutshell, *novelty and commonness appear to be more substitutes than complementary in rising citations.*[18]

4.4 Does it Pay to Address more Novel Research Questions?

Merton (1957) and Dasgupta and David (1994) described well how priority is key for distributing credit in the Open Science community. Because highly novel papers are more likely to open up new knowledge areas, we expect those papers to more directly influence subsequent work and to more often become "citation classics". As we know scientific credit and (monetary and non-monetary) rewards correlate with citations (Garfield, 1984; Gomez-Mejia and Balkin, 1992; Evans, 2008; Lynn, 2014), we are, from a normative point of view, very interested in novel articles being more cited as a sign of preserved incentives to engage in novel research.

A neat identification of the direct individual rewards of performing novel research in terms of forward citations is clearly beyond the scope of this paper, mainly because article quality is not observable. Indeed, as article quality is likely to be positively correlated with novelty and to raise citations, the impact of novelty on citations is likely to be over-estimated in a naive approach. We can however include a number of additional covariates to partially capture article quality. A number of previous studies have documented a strong relationship between team composition and citations (Adams et al., 2005; Wuchty, Jones and Uzzi, 2007; Jones, Wuchty and Uzzi, 2008; Adams, 2013). We thus include the number of co-authors, the number of distinct institutions and the geographical regions dummies as additional co-variates. We also include a number of other controls such as year and discipline dummies, the number of keywords, and keyword novelty which, as we have shown in Subsection 3.3, correlate with keyword combination novelty. Because those controls may jointly only imperfectly capture article quality, we are inclined to interpret the odds ratios of novelty on citations as upper-bounds. The results are synthesized in Table 3.

Our first evidence is that highly novel papers receive 32% more citations than other papers, a number which actually remains close to the one obtained in Section 4.2 (38%,

[18] As a robustness check, we also performed a set of regressions including novelty, future commonness and the interaction between those two variables. Results (Tables 52-54) confirm this point (the odds ratios of the interaction term are always below unity).

see Table 2). The probability of becoming a "big hit" is more sensitive as the odds ratios drop to 25% (vs. 42% initially). These results are consistent across broad scientific fields. It is interesting to note that the odds ratios of a "big hit" are larger when citations are recorded over 5 years, indicating that time plays slightly in favor of novelty.

In short, the estimated conditional correlation of impact and novelty are reduced as compared to what we found in the previous subsection, but remain quite significant. However, as we are likely to interpret them as upper-bounds of the individual return to novelty.

4.5 Novelty and Risk

We have shown that novelty is associated with more citations even when controlling for a set of confounding variables that are likely to be correlated with team quality and that future commonness is an important part of the equation. In the current section, we explore the role of risk in the different tradeoffs.

Risk in the Prediction Picking a team working on a highly novel subject predicts a 38% rise in citations. An additional positive relation of novelty with the dispersion of citations would be consistent with the idea that a more novel research corresponds to an induced risk (a higher average return at the expense of a higher variance). As we employ a generalized negative binomial regression to regress citations on novelty and other covariates, we calculate the coefficient of each covariate not only with the mean but also with the dispersion parameter (Fleming, 2001; Verhoeven, Bakker and Veugelers, 2016). The risk induced hypothesis would be consistent with an estimated significant and positive coefficient of high novelty on the dispersion parameter as it would indicate a higher variance of the number of citations when articles are highly novel.

Interestingly, results go very consistently in the reverse direction as high novelty decreases citations dispersion (Table 2, last two columns "Gen. Neg. Bin. (ln(alpha))"). Picking a team working on highly novel research decreases the dispersion of citations by 15% over a 3-year period (4% after a 5-year period).

We look at the Lorenz curves of the number of citations in the two sets of papers (highly novel papers vs. the other papers) to check the validity of this statement.[19] The

[19] A Lorenz curve is a very complete way of representing the concentration of a distribution. It gives, for any proportion of the population considered, the share of total outcome that is associated with the "less wealthy". The Gini coefficient (also reported) is a more synthetic indicator, equal to the ratio of the area between the Lorenz curve and the line corresponding to a perfect equality (the 45 degree line) to the area between perfect equality and perfect inequality (basically the triangle below the 45 degree line). The higher this coefficient, the more unequal the distribution. A Lorenz curve everywhere below another Lorenz curve unambiguously indicates a more unequal distribution. When the distribution depicts a

comparison of the Lorenz curves (Figure 7) built for the citation distributions in the group of highly novel papers vs. in the complement group confirms that a higher novelty can never be associated with a higher dispersion. These results with respect to risk are not specific to the use of keyword combinations instead of journal reference combinations as the dispersion of citations parameter is also negatively correlated to Pairwise Journal References Novelty (cf. Online Appendix).

Risk Taking Wang, Veugelers and Stephan (2016) suggest that individual attitudes vis-a-vis novelty can be interpreted at the individual level as risk taking. The returns of novel research would be greater on average but more dispersed. Interestingly, none of our results is consistent with this view. Rather, in the negative binomial regressions which control for many potential confounders, the results on dispersion (see ln(alpha) column in Table 3) show that the forward citations of highly novel papers are 7-to-12% *less* dispersed than non-novel papers. The results are never positive whenever we employ a 3 or a 5-year citation window or focus on any given broad field of science.

Note that this estimation does not account for the part of the risk that is associated with ending up not being published in a journal referenced by the WoS. However, this part of the risk cannot explain the difference with the results of Wang, Veugelers and Stephan (2016) as they too are using publication data.[20] That difference also cannot be due to us using pairs of keywords, as regressing (generalized negative binomial) citations on Pairwise Journal References Novelty never reveals a positive correlation with the dispersion of citations (see Online Appendix, Tables 82–83).[21]

Risk, Novelty and Commonness We have seen earlier in this section that future interest in the most novel keyword combination (commonness) has a strong influence on citations. We wonder whether the risk could in fact be due to the unpredictability of others' future interest, potentially interacting with novelty. A way to address this question consists in checking, in the generalized binomial model on citations, if the dispersion of citations (ln(alpha)) is positively affected by future commonness and novelty dummies. The answer is negative as all three dummies that combine high novelty and/or commonness decrease citation dispersion. The highly-novel-and-common dummy decreases citation dispersion by 12.4%, a coefficient close to that of the highly-novel-and-*non*-common dummy (-10%). Note commonness alone (*not*-highly-novel-and-common dummy) does so

lottery (returns in the different states of the world), a more dispersed distribution corresponds to a more risky asset.

[20] In principle, it could be due to the fact they are using only one year of WoS publications while we use more than ten years, but this is not likely as numbers are very large and robust.

[21] We are inclined to believe that their result is in fact specific to their particular version of the journal references novelty index that they introduce in their paper, which does not rely on frequencies but on radically new journal pairwise combinations.

slightly less (-7.5%).Thus risk is actually only greater for the dummy taken in reference, that is *not*-highly-novel-and-*non*-common dummy [22]

In fact the negative relation between novelty and risk turns out to be linked to future commonness in a surprising way. The key observation is that novelty and (future) commonness are actually strongly positively correlated (correlation coeff. equals .49, see correlation Table 6 in the Online Appendix). *Prima facie*, this looks like an inconsistent statement as novelty and commonness are negatively related by construction. However, the time lag reverses the relation as novel contributions in t are likely to address problems which become common in years $(t+1)-(t+2)$. Why? Because contributions that are common in t are likely to address problems no longer visited in the following years $(t+1)-(t+2)$. The novelty of problems is a predictor of interest from the scientific community in the near future. The papers whose most original angle is already present in the literature, are more likely to *not* be common anymore in the near future. In a rapidly changing environment like the academic science system, avoiding addressing new research questions seems to be what in fact really puts agents at risk, not the reverse.

4.6 Robustness checks

We here first assess the insensitivity of our results to alternative ways of computing Pairwise Author Keywords Novelty. We explore three main variants for the system of Equations (1)–(4).[23] A first variant avoids making across fields maximization (cf. Equation 4). A second variant takes the minimum value of the distribution of keyword pairs frequencies in a given paper instead of the tenth percentile (Equation 1). Equation (1) implies that the computation of pairwise keyword frequencies is based only on the pool of papers that were published in the same year as the focal paper (publication year t). This way of proceeding is appealing as it is simple and easier to compute. Of course, novelty may need to be defined with respect to what has been done in the past as well. However, taking the past into consideration adds in complexity while it is not likely to add variation in the data. To assess the validity of this statement, we compute a third variant of the novelty index that considers a backward time window ($t-2$ to t) to assess the frequency of pairs of Author Keywords.

All regressions are reproduced on those variants of novelty (reported in the Online Appendix). All the estimations reproduced using those variants of Pairwise Author Keywords Novelty lead to very similar results in magnitude to the ones obtained employing our main indicator.

Further, most results presented above rely on a certain definition of highly novel papers.

[22] See the ln(alpha) coefficients in Table 46 of the Appendix.

[23] See Online Appendix A.

Up to now, a paper is highly novel if it is in the top 10% most novel articles. To ensure this choice does not significantly affect the results, we use alternative thresholds to define highly novel papers: top 1% and top 5%. All the regressions using the top novelty dummy are replicated afresh using top 1% and top 5% dummies. All results are reported in Online Appendix G, Tables 86-109. Essentially, our results remain qualitatively unchanged when using alternative thresholds. However, the effect of Pairwise Author Keywords Novelty on forward citations is accentuated for most of the models when considering a more restrictive threshold (larger incidence ratio rates for top 1% and top 5% novelty dummies on impact).

4.7 Benchmarks

As discussed in Subsection 2.2, an important choice we made concerns the use of Author Keywords over KeyWords Plus or journal references. Relying on one or the other definition of novelty is not likely to lead to the same top novel articles. Naturally follow questions on the persistence of the positive relation between top novelty and scientific impact when relying on KeyWords Plus or journal references rather than Author Keywords. Table 4 provides a summary of the incidence rate ratios obtained when regressing citations and "big hits" on Paiwise KeyWords Plus Novelty and Pairwise Journal References Novelty (basically the analogue of Table 2).

A highly novel paper based on KeyWord Plus pairs frequencies receives, on average, 8% more citations. This is significantly less than the 38% rise when employing Author Keywords. Even more surprising is the negative relation with top cited papers, which is consistent across scientific fields and using different specifications of a "big hit" (top 10% and top 5%). Pairs of Author Keywords, which we believe are much better suited to capture the original angle of papers, turn out to be key for predicting articles future success, and in particular "big hits". Besides, nothing really differentiates the two forms of novelty in their relation with risk.

We next employ the indicator based on the novelty of journal references pairs (see Subsection 2.2) instead of the standard Pairwise Author Keywords Novelty in the regressions. Pairwise Journal References Novelty is a positive predictor of "big hits". However, incidence ratio rates are much smaller than the ones obtained for Pairwise Author Keywords Novelty.[24] For instance, high Pairwise Journal References Novelty rises the odds of a "big

[24]We study how the proportion of papers categorized as "big hits" varies across the centiles of novelty in Figure 19 of the Appendix. It turns out that the relationship between journal references novelty and academic impact is globally positive, but nonlinearities are observed for high levels of Pairwise Journal References Novelty. Moreover, the slope is much less pronounced as compared to the observed slope in Pairwise Author Keywords Novelty: in the top decile of Pairwise Journal References Novelty, the probability of being a top 5% cited article is "only" 25% higher than in the lowest decile (to be compared with the 200 to 300% for Pairwise Author Keywords Novelty).

hit" by only 12 to 16%, to be compared with 42 to 46% obtained for Pairwise Author Keywords Novelty. Interestingly, humanities and social sciences behave differently as high Pairwise Journal References Novelty increases the odds of a top 10% paper by 37%, while Pairwise Author Keywords Novelty does so by 25% only. The benchmark Pairwise Journal References Novelty indicator is thus a better predictor of impact for the sub-sample of humanities and social sciences only. This could reflect the fact that scientific innovation in those disciplines is often due to the importation of techniques and concepts from other fields that may be captured by atypical combinations of journal references.

4.8 Novelty and Longer-Term Impact

Recent empirical evidence (Wang, Veugelers and Stephan, 2016; Stephan, Veugelers and Wang, 2017) point to the idea that more novel papers may suffer from a delayed recognition. In other words, the greater impact of highly novel research takes time to materialize. To explore the extent to which this phenomenon occurs with our Pairwise Author Keywords Novelty indicator, we estimate the expected number of forward citations received by highly novel papers for citation windows ranging from 1 to 10 years. We restrict our analysis to the sub-sample of papers that were published within the period 1999 to 2001 to take advantage of a longer time period for citations to be recorded. The results presented in Table 5, clearly contradict the idea that highly novel papers suffer from a delayed recognition. Instead, the estimated number of forward citations are stable for alternative citation windows. For instance, the forward citations one year after publication are already greater by 42% for highly novel papers. The largest incidence ratio rates are obtained for three year citations (49%). They decrease slightly after that, down to 41% more citations over ten years after publication. This stability suggest that the form of novelty captured by our indicator leads to an immediate citation advantage which remains stable over time.[25] The results on the dispersion of citations (risk) are also invariant to the variation in the length of the citation window considered. Our main conclusion here is thus that the relation between novelty indicator based on pairs of Author Keywords and future citations is is quite invariant for citation windows of various sizes.

As additional robustness checks, we replicate the same analyzes with the Pairwise KeyWords Plus Novelty and with Pairwise Journal References Novelty (see Table 5). We find a very similar invariance of the relation between top novelty and the number of citations with respect to the depth of the citation window. The regressions on the two main benchmark indicators (top Pairwise KeyWords Plus Novelty and top Pairwise Journal References Novelty) do not evidence any delayed recognition effect either.

In Online Appendix G (Tables 122–123), we see that the results are mainly preserved

[25] Unreported regression results show that adding more controls (as we do for the regressions reported in Table 3) do not alter this result.

when focusing on more restrictive definitions of top novelty (top 5% and top 1%). The only significant exception is that the odds ratios are even larger when focusing on more selective definitions of top novelty according to Pairwise Author Keywords Novelty.

4.9 Concluding note

We have found so far that novelty and impact have a significant and positive relation, so that higher novelty is a good predictor of impact, even in the long run and in the short run, and that this does not come at a higher risk, conditioned on being published. Novelty actually reduces this form of risk, in part because it reduces the risk of addressing problems that are not active in the immediate future, not the reverse. In the next section, we specifically investigate and highlight the role played by journals and their stratification on the publication of highly novel papers, and this is likely to help us understand how a higher risk may be associated to novel research.

5 Journal Stratification, Novelty and Risk

Publishing in top journals has become one of the key signals as it enhances recruitment and promotion opportunities. Serious concerns have been raised in the literature on the peer review system being potentially biased against novelty due to some form of conservatism (Braben, 2004; Chubin and Hackett, 1990; Wesseley, 1998). Rather than being perceived as valuable, divergent contributions may instill confusion and even irritation among evaluators, leading to potential publication penalties (Leahey, Beckman and Stanko, 2017). Resch, Ernst and Garrow (2000) highlight that studies supporting unorthodox medical treatments receive lower ratings even though the supporting data are equally strong. Luukkonen (2012) argues that frontier research is less likely to appear as rigorous while Boudreau et al. (2016) show that high levels of novelty are associated with lower evaluations by experts.

To our knowledge, the literature analyzing peer review in science has only considered specific scientific journals or funding programs. We adopt a distributional approach to explore the potentially varying role played by academic journals on novelty with respect to their stratification. Subsection 5.1 documents that higher impact journals significantly publish more novel articles. Subsection 5.2 shows that high impact journals publish more highly novel articles than their average article quality suggests. This result will provide an interesting new angle to better understand the relation between novelty and risk.The last subsection discusses and rules out that top journals may be cherry picking trendy subjects and that this could confound the relation between novelty and impact factor.

5.1 Novelty and Journal Stratification

Are highly novel papers more likely to be published by large-audience academic journals, namely journals with higher impact factors, for instance because their editors pick path-breaking exploratory contributions? Or are highly novel articles more likely to be published in peripheral journals as alternative views would be more easily accepted in "niches"? To answer those questions, we arrange all articles according to the impact factor of the journals which publish them and calculate the average novelty for each centile of journal impact factor. Figure 8–Graph(a) shows that larger-audience journals are more likely to publish novel papers, at any point of the impact factor distribution and for all broad fields of science.

The positive relation between the centiles of journal impact factors and novelty is particularly strong when one focuses on the highly novel articles (Graphs b, c and d of Figure 8). Graph (b) shows in particular that the average article in the highest impact factor centile is *six times* more likely to be a highly novel paper (top 10%) than the average paper in the first centiles of impact factor. A similar multiplicative factor applies for top 5% and top 1% most novel articles (see graphs c and d), and for any broad field of science. What is really striking is that the relation between the centiles of impact factor and high novelty is convex and that convexity accentuates when we focus on the top 5% and top 1% most novel papers (see Graphs c and d). All the variation in the probability of a paper being in the top 1% most novel articles is observed above the 60th centile of impact factor. The average article published in a journal whose impact factor is median has about the same chance of being highly novel (top 1%) as the average paper published in the lowest impact factor journals. But at the same time, it is six times less likely to be highly novel than the average article published in top impact journals. This "publication premium" for highly novel papers in top journals is amplified for hard sciences and engineering, so that the average article published in a top impact journal in this broad field is *fourteen times* more likely to be highly novel than the average article in the lowest impact factor journals.

5.2 Disentangling The Quality From The Novelty Effect

The fact that top journals tend to publish more novel papers than lower tier journals seems to provide empirical grounds for the idea that they play a crucial role in selecting articles that uncover new research questions, or new dimensions of old questions. However, as we have seen previously, novelty and citations are positively correlated. Because it is likely that citations are also positively correlated with unobserved article quality, then novelty in turn likely positively correlates with article quality. As top journal editors have their own signals on the quality of submitted articles (e.g. their own reading, reviewers' comments,

authors' reputation or prestige of the host institutions...), they may end up selecting novel papers more often for other reasons than their novelty. They may pick novel papers more often, not because they are more novel, but because of their quality.

Disentangling precisely such 'quality effect' from the 'novelty effect' may be very difficult. However, we can appreciate the specific sorting effect of novelty by creating a benchmark built on the assumption that journals' selection is only guided by article quality. Under this assumption, the average article in a journal of a given centile of impact factor distribution should be as novel as the articles in the same centile of citations distribution. Therefore the average novelty according to positions in the citations distribution stands as a benchmark of the average novelty according to impact factor in the absence of any specific role played by novelty *per se* in the selection. Journal stratification would then be "neutral", that is the correlation between journal impact factor and novelty would only reflect the underlying relation between citations and novelty. The comparison of the observed novelty with respect to positions in the impact factor distribution with the benchmark reveals the specific role played by novelty along the impact factor distribution.

To test this, we have calculated the probability of an article being highly novel one all along the distributions of both journal impact factors and citations. Figure 9–Graph(a) shows that, up to the 80th centile of these distributions, the proportion of top 1% novel articles is lower in the impact factor distribution than in the citation distribution. Above the 80th percentile of these distributions, the likelihood of being top novel is greater in the impact factor distribution. This means that top-tier journals offer more slots to highly novel research than their quality standard suggests. In other words, their editors and reviewers allocate a *publication premium* to highly novel contributions. This finding suggests that top journals play a key role in publishing highly novel papers, to a significantly larger extent than their average quality would "naturally deliver". Similar results, though slightly less pronounced, are obtained with top 5% and top 10% novel papers. In other words, top journals match preferentially with highly novel papers, conditional on their quality standards.

An explanation could be that top academic journals play an important 'certifying' role in the science system whose value is magnified for highly novel papers. However this does not imply that more novel articles have an 'easy' access to top journals. To show this point we calculate, for each article, its number-of-citations-to-impact-factor ratio.[26] A ratio greater than the unity means that the article is cited more than the average article published the same year by the same journal. We plot the average number-of-citations-

[26]This is equivalent to the relative citation rate (RCR) introduced by Schubert and Braun (1986) because our Impact Factor is calculated according to exactly the same period length, citing sources and starting date as citations. Note this definition of Impact Factor is different from the traditional one. It is calculated in-house as an "expected citation score" which averages the citations received (in the future) by the articles published by a given journal in a given year.

to-impact-factor ratio for each centile of article novelty in Figure 10–Graph (b). An increasing relation between novelty centiles and the ratio is observed, indicating that the more novel the papers, the better they perform as compared to the average paper in their journal and year. Below a certain level of novelty (about the 40th centile), articles have a ratio below one. That is, those articles are published by journals that publish papers that are on average more cited. On the other side of the aisle, most novel papers have a 10% premium over the average paper in their journal. If we accept the idea that the number of citations is a reasonable proxy for the unobserved scientific quality of articles (Gottfredson, 1978; Campanario, 1995; Bornmann and Daniel, 2008, 2009; Bornmann et al., 2011), this graph in fact highlights that more novel articles face a negative bias as they should be published in higher impact scientific outlets.

This suggests an obvious conclusion in terms of risk. If highly novel papers are to match preferentially with top journals, and if they do not have a facilitated access to those journals relative to their quality, then their chances to get published are likely to be lower. Thus, even though highly novel papers face a lower citational risk (conditional on being published), their publication risk due to journal stratification being positive assortative with article novelty, is likely larger.

5.3 Are Top Journal Cherry Piking Trendy Subjects?

As novelty correlates with future commonness, we are worried that the apparent role of top journals in publishing novel papers could be partly explained by their capacity to pick articles on "hot topics", thereby anticipating rising interest from the scientific community. To assess this, we have computed, for each centile of journal impact factor, the proportion of highly-novel-and-common articles,[27] which we have divided by the overall proportion of such articles. Similar calculations have been performed for the other three categories of articles: not-highly-novel-and-common, highly-novel-and-not-common and not-high-novelty-and-not-common papers. A ratio greater than one for a given type of paper and a given centile of impact factor means that this type of article is more frequent in this centile relative to in all journals. Figure 9 presents the resulting ratios when high novelty is defined as the top 10% or as top 1% most novel articles. We find that the curves of high-novelty-and-common and of high-novelty-and-not-common articles have very similar patterns. They are both increasing from .5 up to 2 or 3 (depending on the definition of high novelty) with impact factor centiles. Those two types of papers tend to be significantly more relatively concentrated in the top journals and nothing really differentiates them. Top journals have a tendency to pick more novel papers, whenever their most "original angle" remains active in the following years or not. This rejects the hypothesis that cherry picking of hot topics may explain the propensity of top journals to publish highly novel

[27] See Section 4.3 for a detailed explanation of how those dummies are calculated.

papers. Surprisingly, we find that among highly novel articles, the not-common articles are even relatively more frequent than the common ones in the top journals when high novelty is defined as the top 1% most novel articles (Figure 9–Graph (b)). In fact, commonness instead makes a big difference among the not-highly-novel articles in gaining access to high impact factor journals. The not-highly-novel-and-not-common are less frequent in higher impact journals while the reverse holds for not-highly-novel-and-common papers. This is consistent with the idea that novelty and commonness are substitutes for getting published in top impact journals.

6 Conclusion and Discussion

Summary of the main results

In this article we propose a new measurement of scientific articles novelty based on pairwise Author Keywords frequencies. We compute this index as well as neighboring indexes on the set of all research articles in the WoS (about ten million articles) over fifteen years (from 1999 to 2013) and study their relation to a number of other variables, in particular citations and other indicators of scientific impact.

We find that novelty is not declining over the period and that more novel articles are more often produced in larger teams that span more institutions and geographical regions. Novelty turns out to be is a good predictor of citations, as it increases the probability of being a "big hit" by 42% to 50% and the number of citations by 38%. This holds both in the short and in the long run. Holding constant several co-variates whose coefficient estimates become closer to the citational returns of novelty at the individual level, the correlation between novelty on citations decreases but remains positive and significant (from 25% to 32%).

Besides, we show that science has common traits with finance and beauty contests as publishing a paper which has a new angle at the time of publication which itself remains active in the following years, has its probability of becoming a citational "big hit" increased by 67 to 82%. However, there is no complementarity between present novelty and future commonness in raising citations. As novelty has significant positive citational returns, we test whether this comes at the expense of a higher citational risk. The reverse rather holds: avoiding novelty increases citational risk, in particular the portion which is caused by others no longer being interested in the topic –and thus not citing the work. In a rapidly changing environment like science, avoiding novelty puts agents at citational risk, not novelty.

Further, we have uncovered the role played by journals in the publication of most novel articles along their impact factor stratification. Highly novel articles are six times more

likely to be published in top journals – up to fourteen times more likely in the hard sciences and engineering. The publication premium offered by top-journals to highly novel papers goes significantly beyond what their quality standards would deliver. This suggests that an assortative matching process is at play besides the pure two-sided quality sorting through which publication slots in highly reputed journals match more frequently with more novel articles. A natural explanation for this phenomenon would be that the signaling value of top journals peer review processes is larger for highly novel papers. This has some potential consequences in terms of risk. If highly novel articles are more likely to be published in top journals, those papers likely face higher publication risk as those journals are also more selective. Though it is difficult to assess whether the total risk implied by high novelty is positive or negative, novelty likely involves first a higher risk of publication due to a match with higher journal quality, but if this stage is passed, novelty brings more and more certain citations. This suggests some rationale for research management, in particular that publication records should not be considered too early in the career path as it may significantly impede risk taking.

Implications and discussion

Our findings may be of interest to other domains beyond a better understanding of science. First, we have shown that our main results on the relation between novelty and impact are reliant on using keyword combinations instead of journal reference combinations, and those keywords chosen by authors instead of assigned keywords. Externally assigned tokens frequencies are much less related to the outcomes. This finding speaks to the on-going discussion on subjective and objective assessment methods for creativity (Park, Chun and Lee, 2016). A tentative interpretation of our results is that individuals themselves are much more accurate in characterizing their work and therefore a frequentialist approach aiming at detecting originality will be more efficient if it uses people's tagging of their own work. While externally assigned keywords are less sensitive to actors' strategic choices of words, this advantage is overwhelmed by the gains in terms of accuracy brought forward thanks to using self-chosen tokens.

Another contribution concerns the connection between novelty and risk in highly innovative social systems characterized by a constant quest for both novelty and others' attention. A widely accepted premise in management research is that developing highly novel ideas inherently entails significant risk and uncertainty. Hence, it is often assumed organizations should balance their attention and resources between the "exploration of new possibilities" and the "exploitation of old certainties" (March, 1991). Supporting this view, prior work on ambidexterity and punctuated equilibrium (Gupta, Smith and Shalley, 2006; He and Wong, 2004; Burgelman, 2002) has offered a series of mechanisms through which organizations can manage this trade-off. For example, by oscillating be-

tween exploratory and exploitative projects over time (Gibson and Birkinshaw, 2004), by developing different organizational units either one specialized in exploration or exploitation (Benner and Tushman, 2003), or by carrying on projects that embrace moderate levels of novelty (Rosenkopf and McGrath, 2011). Our evidence offers a contrasting view concerning social contexts driven by the quest for innovation such as science. We find that publishing highly novel science brings more and less dispersed citation thus suggesting exploratory projects do not inherently entail higher risk, at least that part of the risk associated to the peer recognition of the idea. That risk rises in part because what is common today is more likely to not be even noticed tomorrow. In very creative social contexts such as science, but this could extend to music or fashion, highly exploratory projects bring more success than conventional projects *and* they also do not carry greater risk and uncertainty. In other words: 'playing it safe' is not an option.

That does not mean the quest for top novelty would be the only reliable strategy to generate impact in such social environments though. Indeed, we find that the impact of an idea is not only predicted by its current novelty but also by the anticipation of forward interest. This result speaks to prior research on "creative forecasting". Once ideas are generated, tying to predict their success remains a significant challenge (Berg, 2016). For instance, in creative industries (e.g.: video-games, music or cinema), or high-tech industries (e.g.: IT, software) where there is a constant "quest for originality" (Jung and Lee, 2016), creators and founders tend to choose their projects based on the intended success and recognition from the external community. Unconventionality may predict subsequent success as for Pixar Animation whose success is often explained by the fact that they developed "unconventional and mature plots and topics" (Mannucci and Yong, 2018). We find that this is also true in science as we have shown, but interestingly, we also find that dealing with themes attracting the community's attention in the future predicts success. This suggests being able to anticipate others' interest plays a very important role too. Even ideas that are conventional at the time they are created can be very successful if they address domains, themes or questions that gain interest in the future.

At a time good ideas may come from almost anywhere and anyone either inside or outside the organization (von Hippel, 2005), managers are constantly required to screen and select the most prolific ideas to incorporate them into the innovation process. External open-source communities or internal brainstorming sessions often supply firms with a continuous flow of potential ideas that should be correctly assessed (Christensen et al., 2017; Piezunka and Dahlander, 2015). According to our results, picking a winning idea among the pool of available ones can be viewed as a function of two independent vectors: its current originality and the likelihood that others will follow in the future.

References

Adams, James D., Grant C. Black, J. Roger Clemmons and Paula E. Stephan. 2005. "Scientific teams and institutional collaborations: Evidence from U.S. universities, 1981-1999." *Research Policy* 34(3):259 – 285.

Adams, Jonathan. 2013. "Collaborations: The fourth age of research." *Nature* 497(7451):557–560.

Amabile, Teresa M. 1996. *Creativity in context: Update to the social psychology of creativity.* Boulder, CO: Westview Press.

Arora, Ashish, Sharon Belenzon and Andrea Patacconi. 2018. "The decline of science in corporate R&D." *Strategic Management Journal* (39):3–32.

Azoulay, P., J. S. Graff Zivin and G. Manso. 2011. "Incentives and creativity: evidence from the academic life sciences." *The RAND Journal of Economics* 42(3):527–554.

Baer, M. 2012. "Putting Creativity to Work: The Implementation of Creative Ideas in Organizations." *Academy of Management Journal* 55(5):1102–1119.

Baker, Monya. 2016. "Is There a Reproductibility Crisis?" *Nature* 533(7604):452–454.

Benner, Mary and Michael Tushman. 2003. "Exploitation, Exploration, and Process Management: The Productivity Dilemma Revisited." *The Academy of Management Review* 28(2):238–256.

Berg, J. M. 2016. "Balancing on the Creative Highwire: Forecasting the Success of Novel Ideas in Organizations." *Administrative Science Quarterly* 61(3):433–468.

Bornmann, L. and H.-D. Daniel. 2008. "Selecting manuscripts for a high impact journal through peer review: A citation analysis of communications that were accepted by Angewandte Chemie International Edition, or rejected but published elsewhere." *Journal of the American Society for Information Science and Technology* 59(11):1841–1852.

Bornmann, L. and H.-D. Daniel. 2009. "Extent of type I and type II errors in editorial decisions: A case study on Angewandte Chemie International Edition." *Journal of Informetrics* 3(4):348–352.

Bornmann, L., R. Mutz, W. Marx, H. Schier and H.-D. Daniel. 2011. "A multilevel modeling approach to investigating the predictive validity of editorial decisions: Do the editors of a high profile journal select manuscripts that are highly cited after publication?" *Journal of the Royal Statistical Society* 174:857Ű879.

Boudreau, Kevin J., Eva C. Guinan, Karim R. Lakhani and Riedl Christoph. 2016. "Looking Across and Looking Beyond the Knowledge Frontier: Intellectual Distance, Novelty, and Resource Allocation in Science." *Management Science* 62(10):2765–2783.

Braben, D.W. 2004. *Pioneering research: A risk worth taking.* Hoboken, NJ: Wiley-Interscience.

Burgelman, Robert A. 2002. "Strategy as Vector and the Inertia of Coevolutionary Lock-in." *Administrative Science Quarterly* 47(2):325.

Campanario, J.M. 1995. "Commentary: On influential books and journal articles initially rejected because of negative refereesŠ evaluations." *Science Communication* 16:304Ű325.

Carayol, N. and J-M. Dalle. 2007. "Sequential problem choice and the reward system in the Open Science." *Structural Change and Economic Dynamics* 18:167–191.

Christensen, K., S. Nørskov, L. Frederiksen and J. Scholderer. 2017. "In Search of New Product Ideas: Identifying Ideas in Online Communities by Machine Learning and Text Mining: In Search of New Product Ideas." *Creativity and Innovation Management* 26(1):17–30.

Chubin, D.E. and E.J. Hackett. 1990. *Peerless science: Peer review and U.S. science policy.* Stony Brook, NY: State University of New York Press.

Dasgupta, Partha and Paul A David. 1994. "Toward a new economics of science." *Research Policy* 23(5):487–521.

Evans, James A. 2008. "Electronic publication and the narrowing of science and scholarship." *Science* 321(5887):395–399.

Fleming, Lee. 2001. "Recombinant uncertainty in technological search." *Management science* 47(1):117–132.

Garfield, Eugene. 1984. "The 100 most-cited papers ever and how we select citation classics." *Essays of an Information Scientist* 7:175–181.

Garfield, Eugene and Irving H. Sher. 1993. "KeyWords PlusŹ-algorithmic derivative indexing." *Journal of the Information Science Association* 44(5):298–299.

Gibson, Cristina B. and Julian Birkinshaw. 2004. "The Antecedents, Consequences, and Mediating Role of Organizational Ambidexterity." *Academy of Management Journal* 47(2):209–226.

Gomez-Mejia, L. R. and D. B. Balkin. 1992. "Determinants of Faculty Pay: An Agency Theory Perspective." *Academy of Management Journal* 35(5):921–955.

Gottfredson, S.D. 1978. "Evaluating psychological research reports: Dimensions, reliability, and correlates of quality judgments." *American Psychologist* 33:920–934.

Gupta, Anil K., Ken G. Smith and Christina E. Shalley. 2006. "The Interplay Between Exploration and Exploitation." *Academy of Management Journal* 49(4):693–706.

He, Zi-Lin and Poh-Kam Wong. 2004. "Exploration vs. Exploitation: An Empirical Test of the Ambidexterity Hypothesis." *Organization Science* 15(4):481–494.

Heinze, Thomas, Philip Shapira, Juan D. Rogers and Jacqueline M. Senker. 2009. "Organizational and institutional influences on creativity in scientific research." *Research Policy* 38(4):610–623.

Jacob, F. 1977. "Evolution and Tinkering." *Science* 196(4295):1161–1166.

Janssen, Onne. 2003. "Innovative behaviour and job involvement at the price of conflict and less satisfactory relations with co-workers." *Journal of Occupational and Organizational Psychology* 76(3):347–364.

Jones, Benjamin F. 2009. "The burden of knowledge and the 'death of the renaissance man': is innovation getting harder?" *Review of Economic Studies* 76(1):283–317.

Jones, Benjamin F, Stefan Wuchty and Brian Uzzi. 2008. "Multi-university research teams: shifting impact, geography, and stratification in science." *Science* 322(5905):1259–1262.

Jung, H. J. and J. J. Lee. 2016. "The Quest for Originality: A New Typology of Knowledge Search and Breakthrough Inventions." 59(5):1725–1753.

Kolata, G. 2009. "System leads cancer researchers to play it safe." *New York Times* June 27th.

Kuhn, Thomas S. 1962. *The Structure of Scientific Revolutions*. The University of Chicago Press.

Leahey, Erin, Christine M. Beckman and Taryn L. Stanko. 2017. "Prominent but Less Productive: The Impact of Interdisciplinarity on ScientistsŠ Research." *Administrative Science Quarterly* 62(1):105–139.

Lee, You-Na, John P Walsh and Jian Wang. 2015. "Creativity in scientific teams: Unpacking novelty and impact." *Research Policy* 44(3):684–697.

Li, Ling-Li, Guohua Ding, Nan Feng, Ming-Huang Wang and Yuh-Shan Ho. 2009. "Global stem cell research trend: Bibliometric analysis as a tool for mapping of trends from 1991 to 2006." *Scientometrics* 80(1):39–58.

Luukkonen, T. 2012. "Conservatism and risk-taking in peer review: Emerging ERC practices." *Research Evaluation* 21:48–60.

Lynn, Freda B. 2014. "Diffusing through disciplines: Insiders, outsiders, and socially influenced citation behavior." *Social Forces* p. sou069.

Mannucci, Pier Vittorio and Kevyn Yong. 2018. "The Differential Impact of Knowledge Depth and Knowledge Breadth on Creativity over Individual Careers." *Academy of Management Journal* 61(5):1741–1763.

March, James G. 1991. "Exploration and Exploitation in Organizational Learning." *Organization Science* 2(1):71–87.

Merton, R. K. 1942. "The Normative Structure of Science." *reprinted in Robert K., 1973, The Sociology of Science: Theoretical and Empirical Investigations, Chicago: University of Chicago Press* .

Merton, R. K. 1957. "Priorities in Scientific Discovery: A Chapter in the Sociology of Science." *American Sociological Review* 22(6):635–659.

Millar, Morgan M. and Don A. Dillman. 2012. Trends in interdisciplinary dissertation research: An analysis of the survey of earned doctorates. Technical report Working Paper NCSES 12-200.

Mueller, Jennifer S., Jack A. Goncalo and Dishan Kamdar. 2011. "Recognizing creative leadership: Can creative idea expression negatively relate to perceptions of leadership potential?" *Journal of Experimental Social Psychology* 47(2):494–498.

Mueller, Jennifer S., Shimul Melwani and Jack A. Goncalo. 2012. "The Bias Against Creativity: Why People Desire but Reject Creative Ideas." *Psychological Science* 23(1):13–17.

Olesen Larsen, Peder and Markus von Ins. 2010. "The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index." *Scientometrics* 84:575–603.

Park, Namgyoo K., Monica Youngshin Chun and Jinju Lee. 2016. "Revisiting Individual Creativity Assessment: Triangulation in Subjective and Objective Assessment Methods." *Creativity Research Journal* 28(1):1–10.

Piezunka, H. and L. Dahlander. 2015. "Distant Search, Narrow Attention: How Crowding Alters Organizations' Filtering of Suggestions in Crowdsourcing." *Academy of Management Journal* 58(3):856–880.

Planck, M. 1950. *Scientific Autobiography and Other Papers*. New York: Philosophical library.

Poincaré, Henri. 1910. "Mathematical Creation (Originally published in Science et Méthode, Paris: Flammarion, 1908)." *The Monist* 20(3):321–335.

Price, Derek John de Solla. 1961. *Science since Babylon.* New Haven, Connecticut: Yale University Press.

Resch, K.I., E. Ernst and J. Garrow. 2000. "A randomized controlled study of reviewer bias against an unconventional therapy." *Journal of the Royal Society of Medicine* 93:164–167.

Rosenkopf, L. and P. McGrath. 2011. "Advancing the Conceptualization and Operationalization of Novelty in Organizational Research." *Organization Science* 22(5):1297Ű1311.

Schubert, András and Tibor Braun. 1986. "Relative indicators and relational charts for comparative assessment of publication output and citation impact." *Scientometrics* 9(5-6):281–291.

Shadish, William R., Donna Tolliver, Maria Gray and Sunil K. Sen Gupta. 1995. "Author Judgements about Works They Cite: Three Studies from Psychology Journals." *Social Studies of Science* 25(3):477–498.

Staw, Barry M. 1995. "Why no one really wants creativity." *Creative action in organizations* pp. 161–66.

Stephan, Paula, Reinhilde Veugelers and Jian Wang. 2017. "Blinkered by bibliometrics." *Nature* 544:411–412.

Trapido, Denis. 2015. "How novelty in knowledge earns recognition: The role of consistent identities." *Research Policy* 44(8):1488–1500.

Uzzi, B., S. Mukherjee, M. Stringer and B. Jones. 2013. "Atypical Combinations and Scientific Impact." *Science* 342(6157):468–472.

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

von Hippel, E. 2005. *Democratizing innovation.* MIT press.

Waltman, Ludo and Michael Schreiber. 2013. "On the calculation of percentile-based bibliometric indicators." *Journal of the American Society for Information Science and Technology* 64(2):372–379.

Wang, Jian, Reinhilde Veugelers and Paula Stephan. 2016. "Bias against novelty in science: a cautionary tale for users of bibliometric indicators." *National Bureau of Economic Research* Working Paper 22180.

Weitzman, Martin L. 1998. "Recombinant growth." *Quarterly Journal of Economics* pp. 331–360.

Wesseley, S. 1998. "Peer review of grant applications: What do we know?" *Lancet* 352(9124):301–305.

Wuchty, Stefan, Benjamin F Jones and Brian Uzzi. 2007. "The increasing dominance of teams in production of knowledge." *Science* 316(5827):1036–1039.

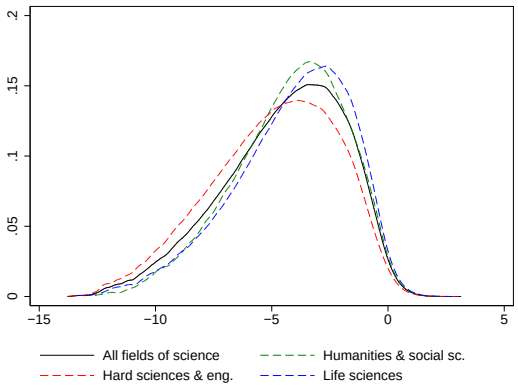
Yuan, Feirong and Richard W. Woodman. 2010. "Innovative behavior in the workplace: The role of performance and image outcome expectations." *Academy of Management Journal* 53(2):323–342.

Zhang, Juan, Qi Yu, Fashan Zheng, Chao Long, Zuxun Lu and Zhiguang Duan. 2016. "Comparing keywords plus of WOS and author keywords: A case study of patient adherence research: Comparing Keywords Plus of WOS and Author Keywords." *Journal of the Association for Information Science and Technology* 67(4):967–972.

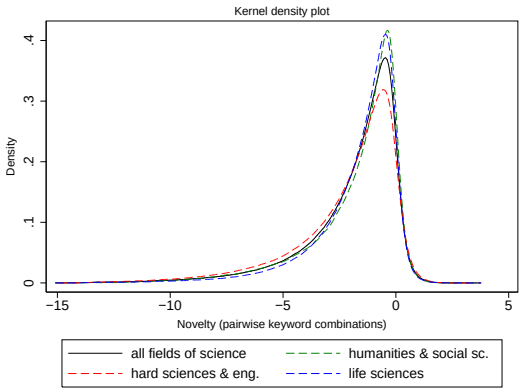
Tables and Figures

Figure 1 – Distributions of Pairwise Author Keywords Novelty, of Pairwise Keywords Plus Novelty, of Journal References Novelty, and of the number of Author Keywords Pairs and KeyWords Plus Pairs.

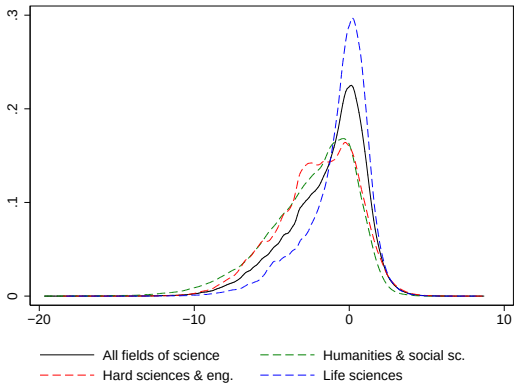
(a) Pairwise Author Keywords Novelty



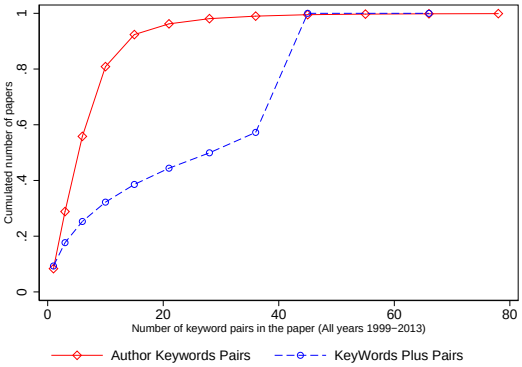
(b) Pairwise KeyWords Plus Novelty



(c) Pairwise Journal References Novelty



(d) Author Keyword and KeyWord Plus Pairs



Notes: Based on the set of all research articles published in journals indexed by the WoS over period 1999-2013.
Kernel density plots for graphs a, b and c. The precise computation of the benchmark indicator of novelty based on the atypicality of journal references combinations (graph c) is presented in Online Appendix C.
Graph d displays the distribution of documents with respect to their associated number of keywords pairs (considering Author Keywords or Keywords Plus).

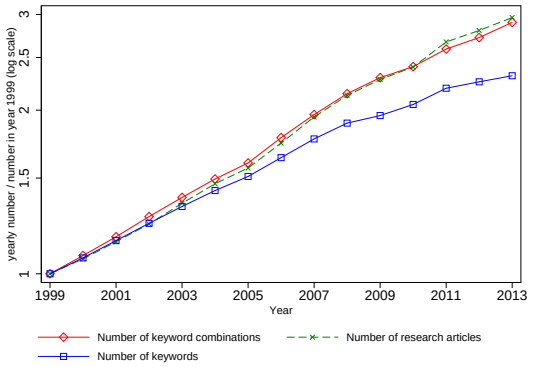
Table 1 – Correlation Table of Various Novelty Indicators and Other Variables.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1 Pairwise Author Keywords Novelty	1																
2 Pairwise Author Keywords Novelty (top-10%)	.49	1															
3 Pairwise Author Keywords Novelty (top-5%)	.37	.69	1														
4 Pairwise Author Keywords Novelty (top-1%)	.19	.30	.44	1													
5 Pairwise KeyWords Plus Novelty	.22	.13	.10	.05	1												
6 Pairwise KeyWords Plus Novelty (top-10%)	-.10	-.04	-.03	-.01	.35	1											
7 Pairwise KeyWords Plus Novelty (top-5%)	-.13	-.07	-.05	-.02	.26	.91	1										
8 Pairwise KeyWords Plus Novelty (top-1%)	-.16	-.09	-.07	-.03	.14	.84	.92	1									
9 Pairwise Journal References Novelty	.12	.09	.07	.03	.16	-.20	-.24	-.29	1								
10 Pairwise Journal References Novelty (top-10%)	-.10	-.06	-.05	-.03	-.08	.26	.29	.31	.44	1							
11 Pairwise Journal References Novelty (top-5%)	-.11	-.07	-.06	-.03	-.10	.29	.33	.36	.35	.92	1						
12 Pairwise Journal References Novelty (top-1%)	-.12	-.08	-.06	-.03	-.11	.33	.36	.40	.19	.85	.93	1					
13 Author Keywords Novelty	-.23	-.17	-.14	-.09	.01	.02	.01	-.00	.12	-.06	-.08	-.10	1				
14 KeyWords Plus Novelty	.01	.02	.01	.01	-.04	-.07	-.08	-.06	.30	-.05	-.10	-.14	.41	1			
15 Number of Author Keywords	.37	.09	.06	.01	.03	-.11	-.12	-.12	.10	-.06	-.08	-.09	.11	.08	1		
16 Field Size (Catcode)	.05	.07	.07	.04	.09	-.03	-.05	-.08	.14	-.03	-.05	-.06	.39	.35	.01	1	
17 Field Size (Journal)	-.02	-.01	-.00	.00	-.02	.10	.11	.11	-.00	.03	.04	.04	.09	.02	-.01	.08	1

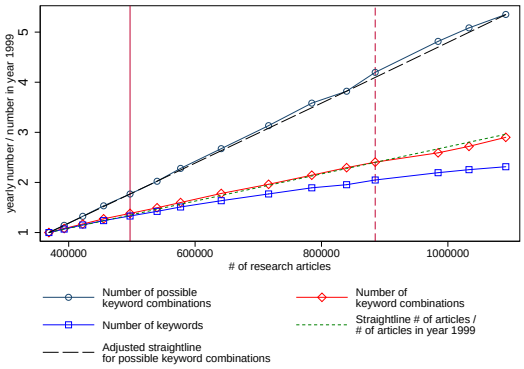
Notes: All correlation coefficients are significant at $p < 0.01$. Calculated on the sample of observations having non-missing values in any of variables (4,397,918 observations).

Figure 2 – The Expansion of Scientific Knowledge.

(a) Number of Distinct Keyword Combinations, Keywords, and Research Articles over the period 1999-2013.



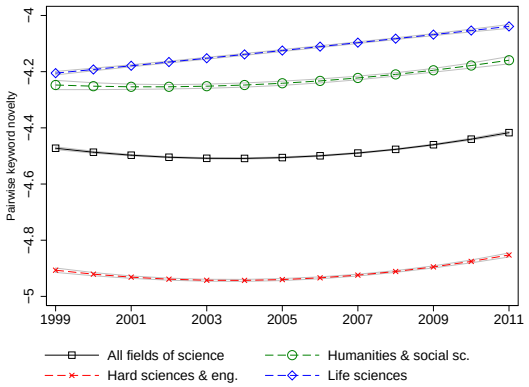
(b) Number of Distinct Possible Keyword Combinations, Keyword Combinations and Keywords with Respect to the Number of Research Articles.



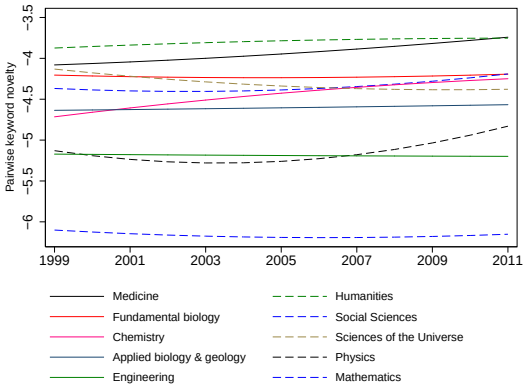
Note: Based on the set of all research articles published in journals indexed by the WoS over period 1999-2013.
The number of documents reached in year 2003 and year 2010 are given by the vertical red lines.
The year 1999 is taken as reference for both graphs.

Figure 3 – The Evolution of Pairwise Keyword Novelty.

(a) For All Fields and by Broad Field of Science



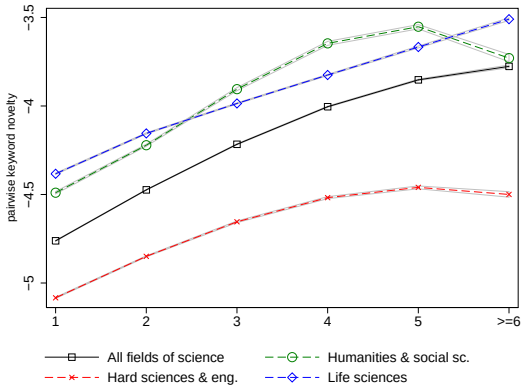
(b) For the Ten Large Scientific Disciplines



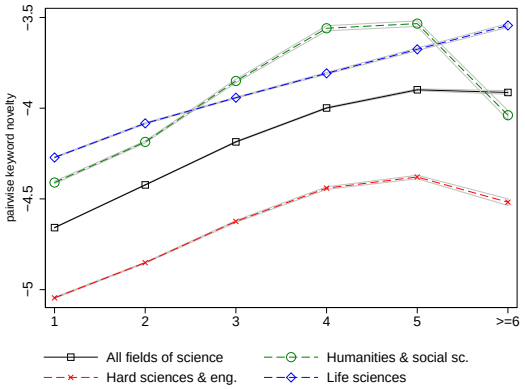
Notes: Based on the set of all research articles published in journals indexed by the WoS, over period 1999-2011. Fractional polynomial estimates and 90% confidence intervals.

Figure 4 – Contexts of Pairwise Keyword Novelty.

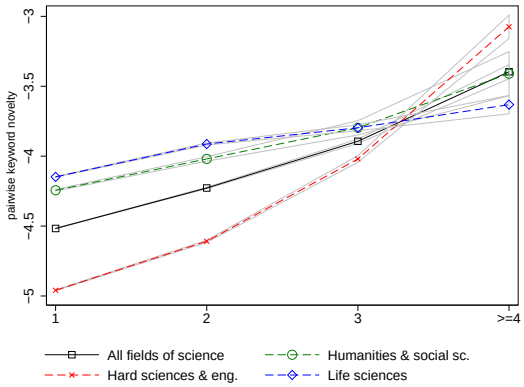
(a) Number of Authors



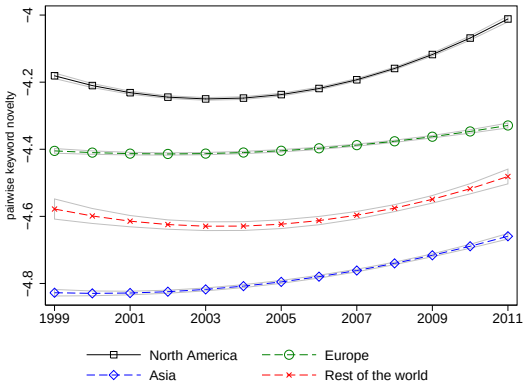
(b) Number of Institutions



(c) Number of World Regions



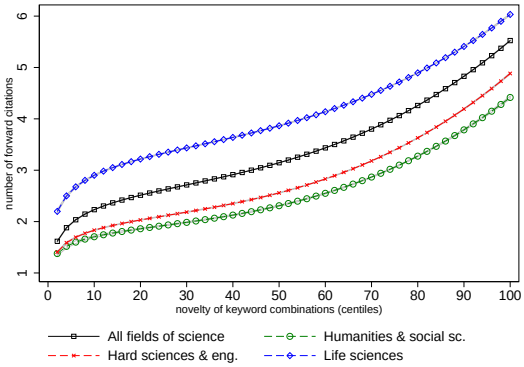
(d) By World Regions



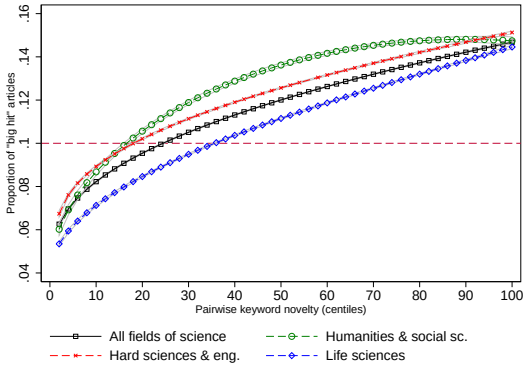
Notes: Based on the set of all research articles published in journals indexed by the WoS, over period 1999-2011. Fractional polynomial estimates and 90% confidence intervals.

Figure 5 – Pairwise Keyword Novelty and Academic Impact.

(a) Number of Forward Citations



(b) Proportion of "big hit" Articles (top 10%)



Notes: Based on the set of all research articles published in journals indexed by the WoS, over period 1999-2011.
Fractional polynomial estimates and 90% confidence intervals.
Citations are recorded in a 3-year window.
"Big hit" papers are defined as top 10% articles which received the most citations in their subject category.

Table 2 – Predicting Citations and "Big Hit" Probabilities With High Pairwise Keyword Novelty.

	"big hit" (top 10%)		"big hit" (top 5%)		Gen. Neg. Bin. (mean)		Gen. Neg. Bin. (ln(alpha))	
	3-year	5-year	3-year	5-year	3-year	5-year	3-year	5-year
Full sample	42%	45%	41%	44%	38%	37%	-15%	-4%
Hum and social sciences	25%	28%	25%	29%	30%	32%	-21%	-10%
Hard sciences and eng	45%	48%	42%	46%	44%	39%	-15%	-4%
Life sciences	45%	48%	46%	48%	33%	33%	-15%	-4%

Notes: Based on the set of all research articles published in journals indexed by the WoS, over period 1999-2011 (for 3-year citation window regressions) and 1999-2009 (for 5-year citation window regressions).

Obtained from exponentiated coefficients in generalized negative binomial estimations and logistic regressions.

Dependent variable for negative binomial regressions: number of forward citations (3-year and 5-year).

Dependent variable for logistic regressions: dummy taking the value 1 if the paper is a "big hit" in its field ("top 10%" or "top 5%").

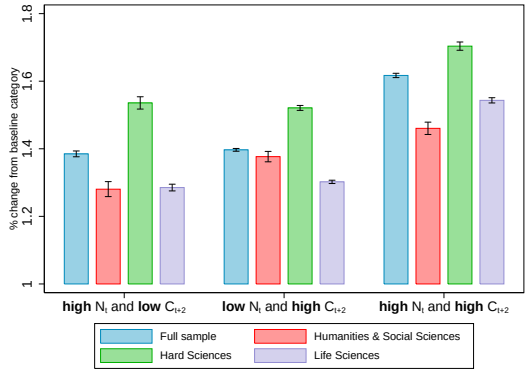
Control variables: number of keywords, publication year dummies and disciplines dummies. All results are significant at the 0.1% level. Detailed regression results can be found in the Online Appendix, see Tables 16–21.

All results have been replicated by employing three alternative variants of Pairwise Author Keywords Novelty. Overall, the obtained results remain consistent. A detailed explanation of the three variants can be found in Online Appendix A. Detailed regression results of the three variants are reported in Tables 22–39.

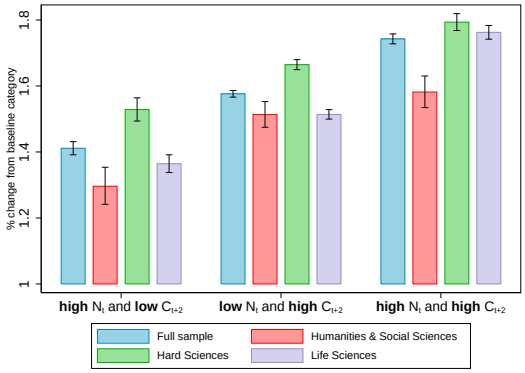
Highly novel papers means are defined here as being among the top 10% most novel papers. Regressions have been replicated by defining highly novel papers with different thresholds (top 5% and top 1% most novel), and can be found in the Online Appendix, Tables 86–97.

Figure 6 – Novel and Before the Crowd.

(a) Dependent Variable: Forward Citations



(b) Dependent Variable: "big hit" Dummy



Notes: Citations are recorded in a 3-year window. "Big hit" papers of Graph (b) are defined as top 10% most cited articles in their subject category.

Results are based on incidence rate ratios from generalized negative binomial regressions for the left graph. For the right graph, results are based on odds ratios from logistic regressions.

The sample is the set of all research articles published in journals indexed by the WoS, over period 1999-2011.

High novelty is a dummy equal to one if the paper is in the top 10% most novel papers. The High Commonness dummy is equal to one if the same keyword combination employed to assess novelty in period t is still used by papers published in periods $t+1$ and $t+2$. Otherwise, it takes the value 0.

High novelty is captured by the dummy equal to one if in the top 10% most novel articles. Commonness dummy equals one if pair of Author Keywords that characterizes the novelty of the article in t is still used in the two following years.

The baseline category for all regressions is the articles that are not-highly-novel in t and not-common in $(t+1)$–$(t+2)$.

Detailed regression results can be found in the Online Appendix: see Table 46 for Graph (a) and in Table 47 for Graph (b). We performed an additional robustness check employing the subset of papers published over the period 1999-2009: see Tables 48-49.

Table 3 – Predicting Citations And "Big Hit" Probabilities With More Controls.

	Basic regressions							
	"big hit" (top 10%)		"big hit" (top 5%)		Gen. Neg. Bin. (mean)		Gen. Neg. Bin. (ln(alpha))	
	3-year	5-year	3-year	5-year	3-year	5-year	3-year	5-year
Full Sample	25%	27%	24%	26%	32%	31%	-11%	-11%
Hum and social sciences	23%	28%	26%	30%	31%	32%	-18%	-18%
Hard sciences and eng	22%	25%	20%	23%	19%	18%	-12%	-10%
Life sciences	27%	29%	28%	30%	33%	32%	-10%	-8%

Notes: Based on the set of all research articles published in journals indexed by the WoS, over period 1999-2011 (for 3-year citation window regressions) and 1999-2009 (for 5-year citation window regressions).

Obtained from exponentiated coefficients in generalized negative binomial estimations and logistic regressions.

Dependent variable for negative binomial regressions: number of forward citations (3-year and 5-year). Dependent variable for logistic regressions: dummy taking the value 1 if the paper is a "big hit" in its field ("top 10%" "top 5%").

The explaining variable is high novelty, a dummy equal to one if the article is in the top 10% most novel articles. All results are significant at the 0.1% level.

Original tables are presented in the Online Appendix, see Tables 58–63.

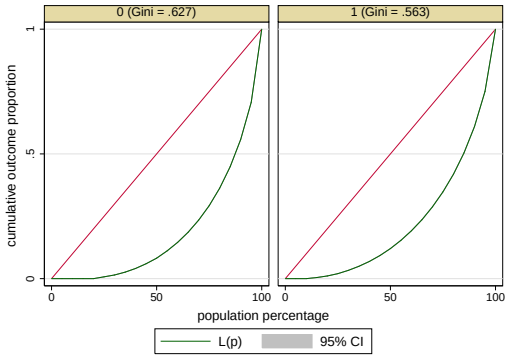
All results have been replicated by employing three alternative specifications of Pairwise Author Keywords Novelty. Overall, the obtained results remain consistent. A detailed explanation of the three variants can be found in Online Appendix A. And detailed regression results of the three variants are reported in Tables 64–81.

Control variables: number of keywords, keyword novelty, publication year, discipline dummies (10), geographical dummies (Europe, USA-Canada, America (other), Asia-Oceania, Other), number of institutions and number of authors.

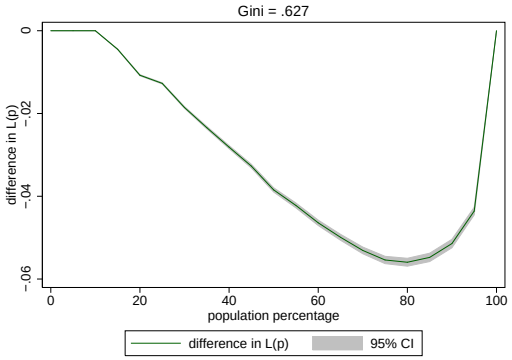
Highly novel papers means being among the top 10% most novel papers. Results have been replicated by defining highly novel papers with different thresholds (top 5% and top 1%), and can be found in Tables 98 - 109.

Figure 7 – Lorenz Curves for the Number of 5-year Citations, Contrasted by Top 10% Novelty (Graph a), and Difference Between Lorenz Curves (Top 10% Novel Papers Taken into Reference in Graph b).

(a) Lorenz curves



(b) Contrasts between Lorenz curves



Notes: (a) presents Lorenz curves (and Gini coefficients) in the group of not-highly novel papers (dummy = 0) and in that of highly-novel articles (dummy = 1). The right graph is the difference between the two Lorenz curves (left curve vertical coordinates minus right curve coordinates).

Table 4 – Predicting Citations and "Big Hit" Probabilities with Pairwise KeyWords Plus Novelty or Pairwise Journal References Novelty.

	Pairwise KeyWords Plus Novelty							
	"big hit" (top 10%)		"big hit" (top 5%)		Gen. Neg. Bin. (mean)		Gen. Neg. Bin. (ln(alpha))	
	3-year	5-year	3-year	5-year	3-year	5-year	3-year	5-year
Full sample	-8%	-7%	-10%	-8%	8%	8%	-3%	-3%
Hum and social sciences	-12%	-9%	-15%	-9%	0%	4%	-4%	-4%
Hard sciences and eng	0%	0%	-3%	-2%	5%	4%	-6%	-7%
Life sciences	-10%	-9%	-10%	-8%	9%	8%	-2%	0%
	Pairwise Journal References Novelty							
	"big hit" (top 10%)		"big hit" (top 5%)		Gen. Neg. Bin. (mean)		Gen. Neg. Bin. (ln(alpha))	
	3-year	5-year	3-year	5-year	3-year	5-year	3-year	5-year
Full Sample	12%	12%	14%	16%	13%	10%	-7%	-3%
Hum and social sciences	37%	31%	44%	40%	22%	20%	0%	4%
Hard sciences and eng	9%	10%	12%	15%	20%	14%	-6%	-1%
Life sciences	12%	12%	13%	14%	9%	6%	-8%	-4%

Notes: Obtained from exponentiated coefficients in generalized negative binomial estimations and logistic regressions. Dependent variable for negative binomial regressions: number of forward citations (3-year and 5-year).

Dependent variable for logistic regressions: dummy taking the value 1 if the paper is a "big hit" in its field ("top 10%" or "top 5%").

Control variables: number of keywords, publication year and disciplines dummies. All results are significant at the 0.1% level.

Detailed regression results using Pairwise KeyWords Plus Novelty can be found in Tables 40–45.

Detailed regression results using Pairwise Journal References Novelty can be found in Tables 110–115.

Table 5 – Novelty and Citation Impact In The Short To The Long Term.

	Citation window (years)									
Gen. Neg. Bin.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Mean										
Pairwise Author Keywords Novelty	42%	49%	49%	48%	46%	45%	44%	43%	42%	41%
Pairwise KeyWords Plus Novelty	5%	9%	10%	9%	8%	8%	7%	6%	6%	5%
Pairwise Journal References Novelty	4%	17%	20%	21%	21%	20%	20%	20%	19%	19%
Ln(alpha)										
Pairwise Author Keywords Novelty	-15%	-13%	-11%	-10%	-9%	-9%	-9%	-9%	-9%	-9%
Pairwise KeyWords Plus Novelty	1%	0%	-2%	-2%	-1%	-1%	-1%	-1%	-1%	-1%
Pairwise Journal References Novelty	ns	-12%	-9%	-7%	-5%	-4%	-3%	-2%	ns	ns

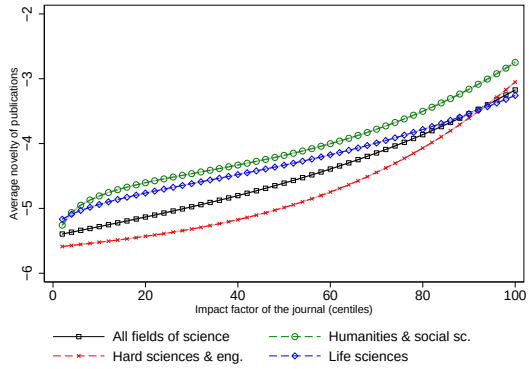
Estimates are obtained from exponentiated coefficients of generalized negative binomial models. Ln(alpha) are obtained from the dispersion estimates of generalized negative binomial models.

Dependent variable: number of forward citations after 1 to 10 years. Novelty variables are dummies indicating if the paper is among the top 10% most novel.

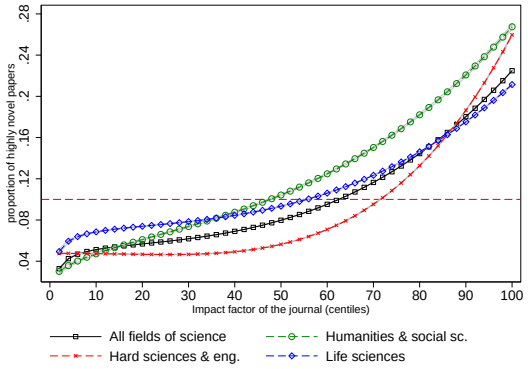
Control variables in all Pairwise Author Keywords Novelty models: number of Author Keywords, publication year dummies, discipline dummies. Control variables in all Pairwise Keywords Plus Novelty models: number of Keywords Plus, publication year dummies, discipline dummies. Control variables in all Pairwise Journal References Novelty models: number of references, publication year dummies, discipline dummies. All detailed regression results are available upon request.

Figure 8 – Pairwise Keyword Novelty and Journal Impact Factor.

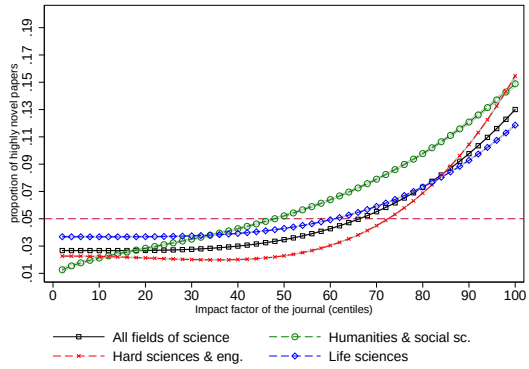
(a) Average Pairwise Keyword Novelty



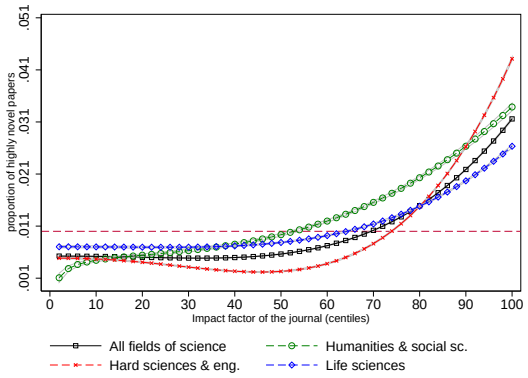
(b) Proportion of Highly Novel Papers (top 10%)



(c) Proportion of Highly Novel Papers (top 5%)



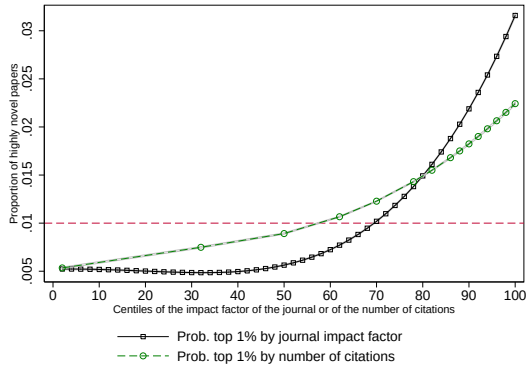
(d) Proportion of Highly Novel Papers (top 1%)



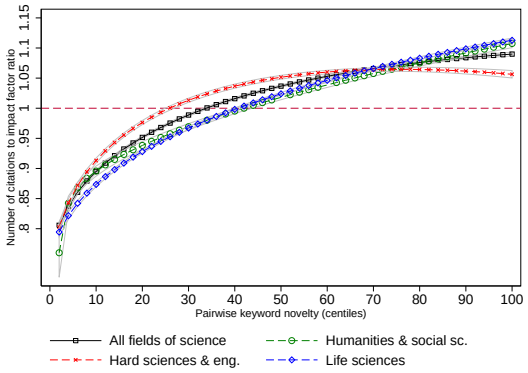
Notes: Based on the set of all research articles published in journals indexed by the WoS, over period 1999-2011.
Fractional polynomial estimates and 90% confidence intervals.
Citations are recorded in a 3-year window.

Figure 9 – Paiwise Keyword Novelty, Citations And Impact Factor distributions.

(a) Proportion of Highly Novel Papers (top 1%)



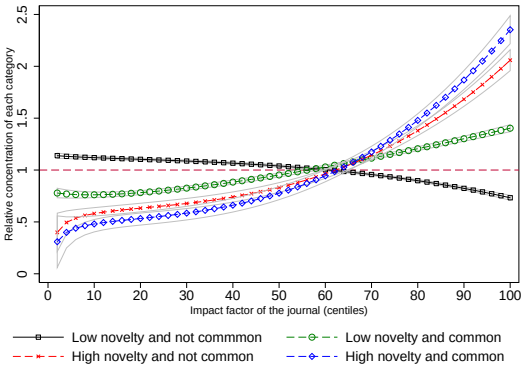
(b) Citations-to-Impact-Factor Ratio Across Novelty Centiles



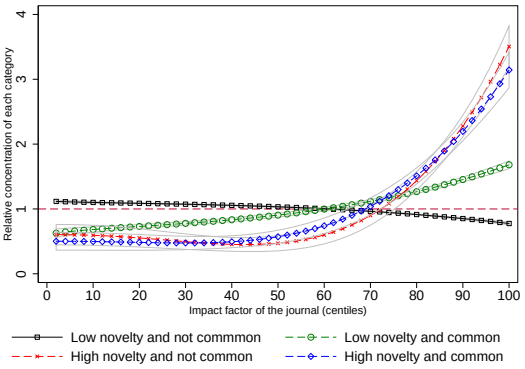
Notes: Based on the set of all research articles published in journals indexed by the WoS, over period 1999-2011. Fractional polynomial estimates and 90% confidence intervals. Citations are recorded in a 3-year window.

Figure 10 – Novelty, Commonness And Journal Impact Factor

(a) Relative Concentration Of Four Categories Of Articles (N-C, nN-C, N-nC, nN-nC) Along The Centiles Of Journals' Impact Factor (Top 10% Novelty)



(b) Relative Concentration Of Four Categories Of Articles (N-C, nN-C, N-nC, nN-nC) Along The Centiles Of Journals' Impact Factor (Top 1% Novelty)



Notes: Based on the set of all research articles published in journals indexed by the WoS, over period 1999-2011.
Fractional polynomial estimates and 90% confidence intervals. Citations are recorded in a 3-year window.
High novelty is a dummy equal to one if the paper is in the top 10% (for Graph(a)) or top 1% (Graph(b)) most novel papers. Commonness dummy is equal to one if the same keyword combination employed to assess novelty in period t is still used by papers published in periods $t+1$ and $t+2$. Otherwise, it takes the value 0.
Relative concentration is computed as the share of each considered category in the centile, divided by the share of that category over all centiles.

Online Appendix to the Article "The Right Job and the Job Right: Novelty, Impact and Journal Stratification in Science"

by N. Carayol, A. Lahatte and O. Llopis

March 18, 2019

Abstract

This appendix includes a set of robustness checks. Section 6 presents three alternative computations of Pairwise Author Keywords Novelty. Section 6 explains the computation of our benchmark novelty indicator based on journal references. Section 6 offers detailed information on all control variables employed in our models, as well as its relation with our Pairwise Author Keywords Novelty indicator. Section 6 offers additional graphs on the relationship between novelty and scientific impact. Section 6 shows the correlation table of the main variables, and Section 6 includes all detailed regressions.

Appendix A: Variants of Pairwise Author Keywords Novelty

In the preceding formulas, we made a series of choices that we believe are sound. However, we would like to check to what extent our results are sensitive to the choices we made. Therefore, we will explore a series of variants.

Variant 1: Not Considering Across Subject Categories Maximization

First, we will test if the across-subject categories maximization of Equation (4) makes a great difference. Therefore we collapse steps from Equation (2) to Equation (4) into one single equation, by taking the 10% percentile of commonness on the total distribution of keyword combination commonness (for all subject categories), as follows:

$$Com = 10thPercentile\left(Com_{ijct} \,|\forall ij \in K, \forall c \in C\right). \tag{5}$$

Detailed results corresponding to the robustness check of Table 2 employing this variant can be found in Tables 22-27. Detailed results corresponding to the robustness check of Table 3 employing this variant can be found in Tables 64-69.

Variant 2: Taking the Minimum Instead of 10th Percentile

Second, we investigate if taking the minimum commonness (maximum novelty) instead of the tenth percentile makes a big difference, that is:

$$Com = min\left(Com_{ijct} \,|\forall ij \in K, \forall c \in C\right). \tag{6}$$

Detailed results corresponding to the robustness check of Table 2 employing this variant can be found in Tables 28-33. Detailed results corresponding to the robustness check of Table 3 employing this variant can be found in Tables 70-75.

Variant 3: Considering Also Previous Years to Calculate Novelty

We could have taken the past into consideration so compute frequencies as the novelty of keyword combinations can be appreciated vis-a-vis predicting years as well. We consider moving windows to avoid introducing an artificial trend. As large moving window may not be efficient and as this would end out loosing more years for the analysis, we use, in

this variant, the past three years of data ($t-2$ to year t) to compute N_{ijct}, N_{ict}, and N_{ct} in Equation (1).

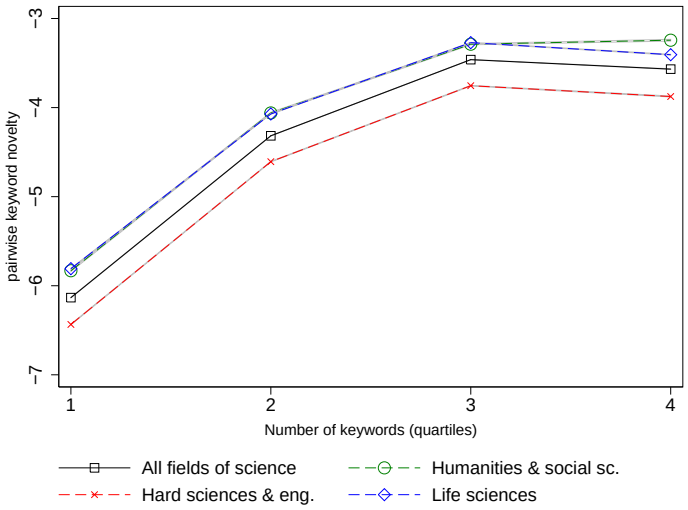
Detailed results corresponding to the robustness check of Table 2 employing this variant can be found in Tables 34-39. Detailed results corresponding to the robustness check of Table 3 employing this variant can be found in Tables 76-81.

Appendix B: Technical Controls

Number of keywords

Papers have different numbers of keywords and as this number grows, so do the number of possible pairwise combinations between them. As the novelty indicator uses the 10th percentile largest novelty among all pairwise combinations, it is expected that pairwise novelty increases with the number of keywords. The minimum number of keywords is two since we do not consider articles which have one or no keywords (because they then have no pairwise combination. The median is 4 and the mean 4.4, variance is 2.6. We see that indeed the number of keywords makes significant change between its first quartile to its third quartile (basically from 2 to 5 keywords), and then increasing the number of keywords makes no more change in novelty. This extends to within-broad-fields analyses.

Figure 11 – Novelty of keyword combinations and number of keywords (3-year window), for all broad fields of science and by domains.



Novelty of Keywords

We would like to investigate whether the novelty of the pair of keywords can be affected by the novelty of the keywords themselves. The idea is that when a paper is using a more novel pair of keywords, it may also be using keywords that are less novel. For this purpose, we need to introduce a measure of keyword novelty. The commonness of i:

$$comk_{ict} = \frac{M_{ict}}{M_{ct}} \tag{7}$$

with M_{ict} the number of papers that use i as a keyword, and M_{ct} the number of papers

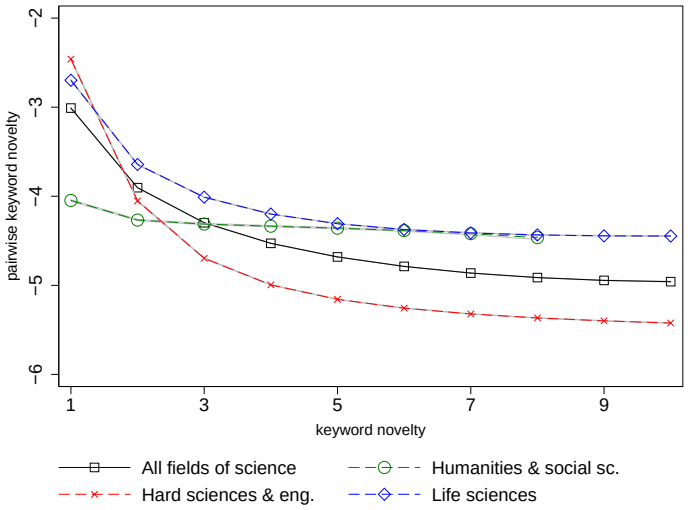
in specialty c in year t. As for the pairs, we use the 10th percentile of the distribution in each speciality c, and compute novelty as

$$comk_c = 10thPercentile\left(Com_{ict} \,|\forall ij \in k\right) \tag{8}$$

being k the set of keywords of the paper. Lastly, we compute the novelty for each keyword and take the max:

$$novk = max_{c \in C} - log(com_{ci}). \tag{9}$$

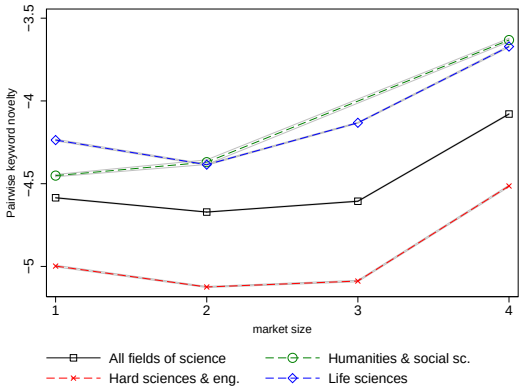
Figure 12 – Pairwise Keyword Novelty and Novelty of Keywords (3-year window).



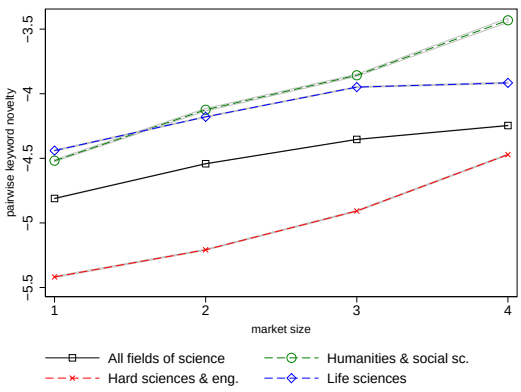
Market Size

Figure 13 – Pairwise Keyword Novelty and Market Size.

(a) Size (Subject Category)



(b) Size (Journal)

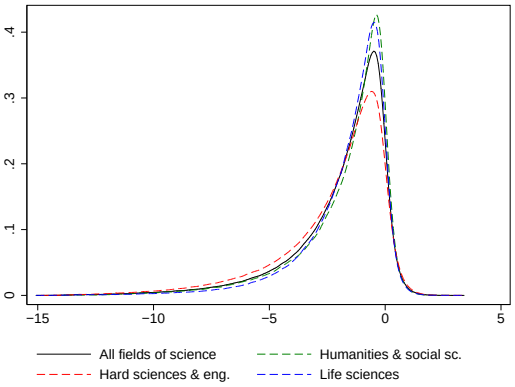


Appendix C: Benchmarks

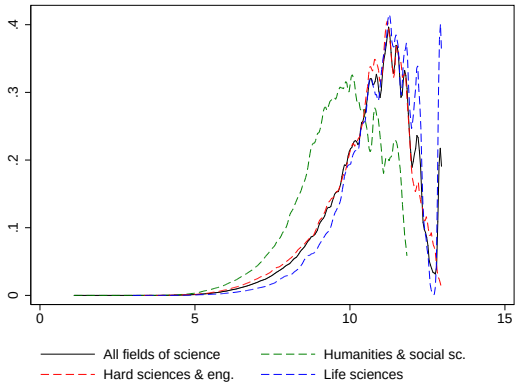
KeyWords Plus

Figure 14 – Distribution of Pairwise Keyword Novelty and Keyword Novelty (KeyWords Plus)

(a) Pairwise Keyword Novelty



(b) Keyword Novelty



Notes: Kernel density plots. Based on the set of all research articles published in journals indexed by the WoS over period 1999-2011.
Indicators based on KeyWords Plus.

Pairwise Journal References Novelty

As novelty of articles in science has been essentially measured using the frequency of journal references combinations, we compute a benchmark indicator using the basics of that approach. We build a journal references novelty indicator that is, as ours, time-variant. It is close in spirit to the one developed by Lee, Walsh and Wang (2015) based on the atypicality of journal reference combinations in year t. The reference commonness of the combination of journals i and j, and year t is given by:

$$RefCom_{ijt} = \frac{N'_{ijt} \times N'_t}{N'_{it} \times N'_{jt}}, \tag{10}$$

Each paper has a list of pairwise journal reference. Let R be the list of such pairwise unordered journals. We consider the tenth percentile of that distribution:

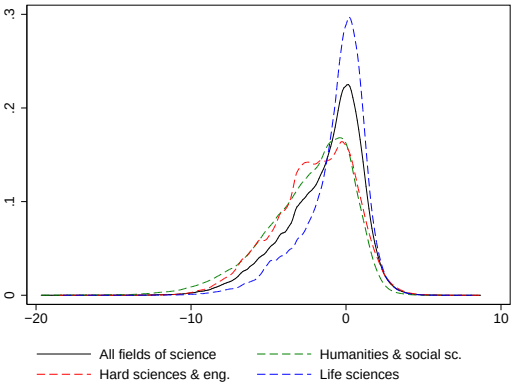
$$Refcom = 10thPercentile\left(RefCom_{ijt} \,|\forall i, j \in R\right) \tag{11}$$

with R the reference list set. The reference to the year is not there any-more since each article is published in a given year. Last, we use the inverse logarithmic transformation of commonness to have the novelty of a given paper:

$$RefNov = -log(RefCom_t) \tag{12}$$

Figure 15 – Distribution of the Pairwise Journal References Novelty. For all articles and for the three science domains.

(a) 3-year citations



(b) 5-year citations

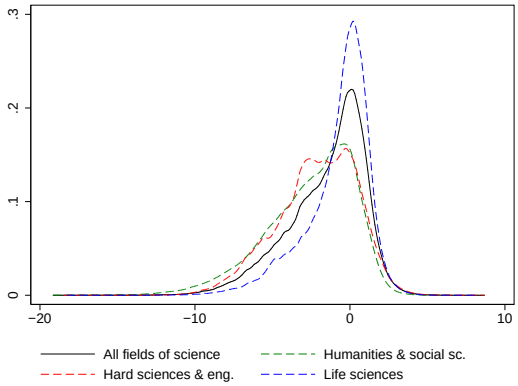
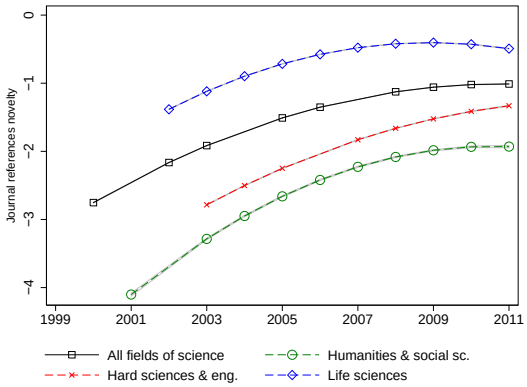
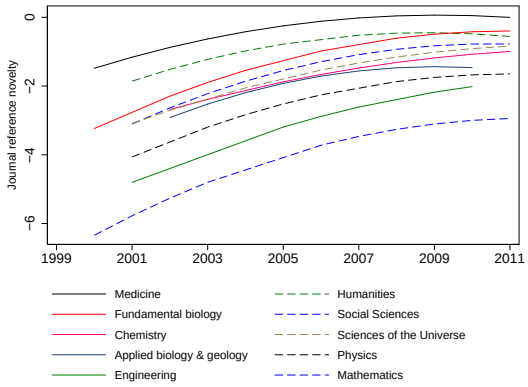


Figure 16 – Evolution of Pairwise Journal References Novelty

(a) For all Science and by Large Fields of Science

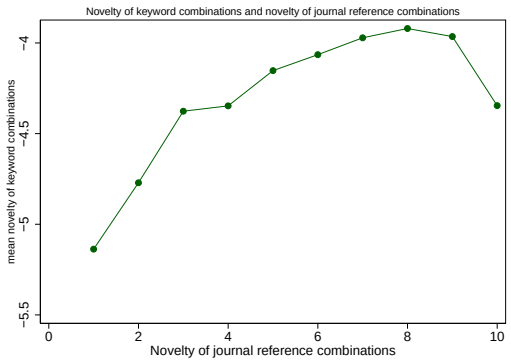


(b) For the Ten Large Scientific Disciplines



Notes: Based on the the set of all research articles published in journals indexed by the WoS, over period 1999-2011. Fractional polynomial estimates and 90% confidence intervals.

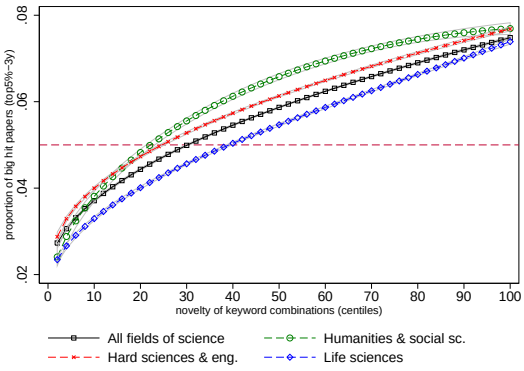
Figure 17 – The correlation between Pairwise Author Keywords Novelty and Pairwise Journal References Novelty.



Appendic D: Citations and Novelty

Figure 18 – Citations and Pairwise Keyword Novelty

(a) Proportion of "big hit" papers (top 5%)



(b) Number of forward citations (5y)

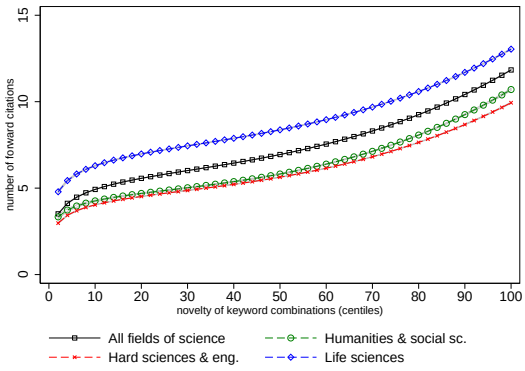
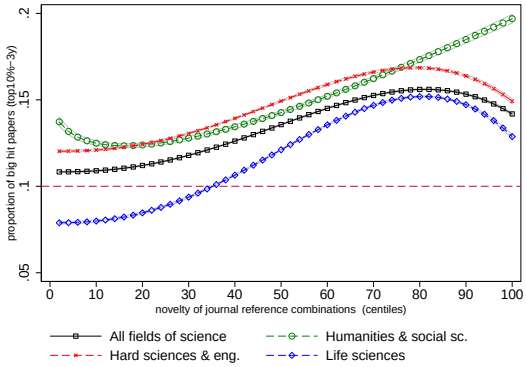
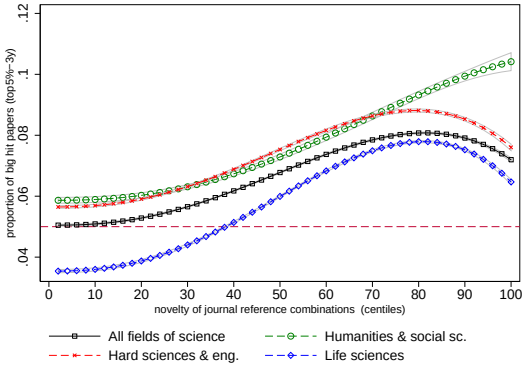


Figure 19 – Citations and Journal References Novelty

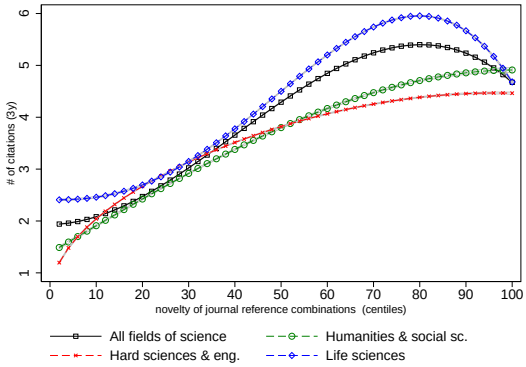
(a) Proportion of "big hit" papers (top 10%)



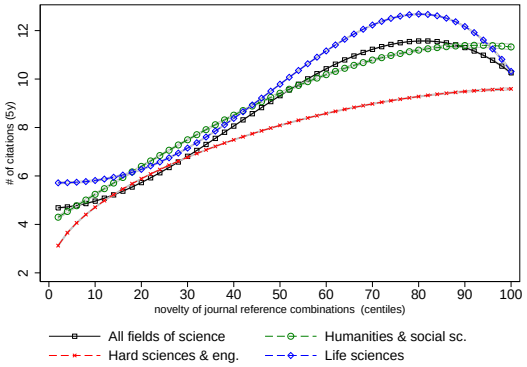
(b) Proportion of "big hit" papers (top 5%)



(c) Forward citations (3 years)



(d) Forward citations (5 years)



Appendix D: Correlations

Table 6 – Descriptive statistics and correlations

	Mean	SD	1	2	3	4	5	6	7	8	9	10	11	12	13
1 Number of citations (3y)	4.00	6.56	1												
2 Big hit papers (top-10%)	0.09	0.28	0.59*	1											
3 Big hit papers (top-5%)	0.04	0.20	0.55*	0.67*	1										
4 Pairwise keyword novelty	-4.30	2.59	0.16*	0.07*	0.05*	1									
5 Pairwise keyword novelty (top-10%)	0.11	0.32	0.11*	0.04*	0.03*	0.49*	1								
6 Keyword novelty	10.31	1.13	-0.02*	-0.04*	-0.03*	-0.23*	-0.17*	1							
7 Commonness(t+2)	0.26	0.44	0.13*	0.06*	0.04*	0.49*	0.38*	-0.10*	1						
8 Journal references novelty	-1.46	2.46	0.18*	0.03*	0.02*	0.12*	0.09*	0.12*	0.07*	1					
9 Journal references novelty (top-10%)	0.10	0.30	-0.12*	-0.07*	-0.05*	-0.10*	-0.06*	-0.06*	-0.08*	0.44*	1				
10 Journal impact factor (3y)	3.83	3.35	0.53*	0.24*	0.18*	0.23*	0.17*	-0.01*	0.16*	0.31*	-0.17*	1			
11 Number of keywords	4.51	1.62	0.10*	0.06*	0.05*	0.37*	0.09*	0.11*	0.10*	0.10*	-0.06*	0.12*	1		
12 Number of authors	2.23	1.72	0.20*	0.09*	0.07*	0.11*	0.08*	-0.02*	0.07*	0.16*	-0.09*	0.25*	0.05*	1	
13 Number of institutions	1.88	1.44	0.17*	0.09*	0.07*	0.08*	0.06*	-0.03*	0.06*	0.09*	-0.08*	0.19*	0.04*	0.90*	1

* denote significance at the 5% level.

Appendix F: Additional Regressions

Section 3.3. Which Teams Produce More Novel Papers?

Table 7 – Determinants of top 10% Pairwise Keyword Novelty. Logit models.

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
novkk_disc_catdom_top10				
#institutions	1.055***	1.085***	1.017***	1.053***
	(0.00)	(0.00)	(0.00)	(0.00)
USA-Can	1.153***	1.042*	1.186***	1.119***
	(0.01)	(0.02)	(0.01)	(0.01)
EUR	1.074***	0.933***	1.151***	1.031***
	(0.00)	(0.02)	(0.01)	(0.01)
Asia	1.038***	0.966	1.128***	0.965***
	(0.00)	(0.02)	(0.01)	(0.01)
Other regions	0.959***	0.778***	1.141***	0.842***
	(0.01)	(0.02)	(0.01)	(0.01)
#keywords	1.249***	1.197***	1.338***	1.214***
	(0.00)	(0.00)	(0.00)	(0.00)
keyword novelty	0.571***	0.610***	0.483***	0.601***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	7405451	639182	3283865	3482404
Log Likeli.	-1958724.5	-123308.4	-696519.0	-1086647.1

Exponentiated coefficients. Sample: Papers published between 1999-2011.
* p<0.05 ** p<0.01 *** p<0.001. Robust standard errors in parentheses.
Dependent variable: Pairwise keyword novelty (top 10%).

Table 8 – Determinants of top 5% Pairwise Keyword Novelty. Logit models.

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
novkk_disc_catdom_top5				
#institutions	1.047***	1.071***	1.010***	1.050***
	(0.00)	(0.00)	(0.00)	(0.00)
USA-Can	1.164***	1.006	1.212***	1.111***
	(0.01)	(0.02)	(0.01)	(0.01)
EUR	1.087***	0.926***	1.164***	1.028***
	(0.01)	(0.02)	(0.01)	(0.01)
Asia	1.069***	1.003	1.175***	0.968***
	(0.01)	(0.02)	(0.01)	(0.01)
Other regions	0.951***	0.791***	1.161***	0.813***
	(0.01)	(0.02)	(0.02)	(0.01)
#keywords	1.215***	1.156***	1.320***	1.170***
	(0.00)	(0.00)	(0.00)	(0.00)
keyword novelty	0.554***	0.580***	0.475***	0.588***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	7405451	639182	3283865	3482404
Log Likeli.	-1193865.9	-72484.7	-411738.6	-677325.4

Exponentiated coefficients. Sample: Papers published between 1999-2011.
* p<0.05 ** p<0.01 *** p<0.001. Robust standard errors in parentheses.
Dependent variable: Pairwise keyword novelty (top 5%).

Table 9 – Determinants of top 1% Pairwise Keyword Novelty. Logit models.

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
novkk_disc_catdom_top1				
#institutions	1.029***	1.053***	1.005	1.049***
	(0.00)	(0.01)	(0.00)	(0.00)
USA-Can	1.206***	1.009	1.212***	1.128***
	(0.01)	(0.04)	(0.02)	(0.02)
EUR	1.146***	0.982	1.160***	1.063***
	(0.01)	(0.04)	(0.02)	(0.02)
Asia	1.127***	1.028	1.199***	0.990
	(0.01)	(0.05)	(0.02)	(0.02)
Other regions	0.941***	0.825**	1.144***	0.774***
	(0.02)	(0.05)	(0.03)	(0.02)
#keywords	1.150***	1.095***	1.257***	1.103***
	(0.00)	(0.00)	(0.00)	(0.00)
keyword novelty	0.527***	0.537***	0.460***	0.569***
	(0.00)	(0.01)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	7405451	637482	3283865	3482404
Log Likeli.	-337773.4	-19225.6	-115294.6	-195284.9

Exponentiated coefficients. Sample: Papers published between 1999-2011.
* p<0.05 ** p<0.01 *** p<0.001. Robust standard errors in parentheses.
Dependent variable: Pairwise keyword novelty (top 1%).

Table 10 – Determinants of top 10% Pairwise Keyword Novelty. Logit models.

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
novkk_disc_catdom_top10				
#authors	1.101***	1.079***	1.050***	1.101***
	(0.00)	(0.00)	(0.00)	(0.00)
USA-Can	1.153***	1.242***	1.167***	1.081***
	(0.00)	(0.02)	(0.01)	(0.01)
EUR	1.087***	1.123***	1.134***	1.009
	(0.00)	(0.01)	(0.01)	(0.01)
Asia	1.050***	1.130***	1.115***	0.944***
	(0.00)	(0.02)	(0.01)	(0.01)
Other regions	0.966***	0.895***	1.125***	0.825***
	(0.01)	(0.02)	(0.01)	(0.01)
#keywords	1.247***	1.194***	1.337***	1.213***
	(0.00)	(0.00)	(0.00)	(0.00)
keyword novelty	0.572***	0.618***	0.484***	0.600***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	7447925	650357	3306793	3490775
Log Likeli.	-1962950.3	-124754.0	-699342.0	-1086577.1

Exponentiated coefficients. Sample: Papers published between 1999-2011.
* p<0.05 ** p<0.01 *** p<0.001. Robust standard errors in parentheses.
Dependent variable: Pairwise keyword novelty (top 10%).

Table 11 – Determinants of top 5% Pairwise Keyword Novelty. Logit models.

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
novkk_disc_catdom_top5				
#authors	1.107***	1.073***	1.051***	1.108***
	(0.00)	(0.00)	(0.00)	(0.00)
USA-Can	1.146***	1.174***	1.174***	1.064***
	(0.01)	(0.02)	(0.01)	(0.01)
EUR	1.083***	1.093***	1.129***	0.997
	(0.01)	(0.02)	(0.01)	(0.01)
Asia	1.066***	1.151***	1.144***	0.940***
	(0.01)	(0.02)	(0.01)	(0.01)
Other regions	0.946***	0.896***	1.129***	0.793***
	(0.01)	(0.02)	(0.01)	(0.01)
#keywords	1.213***	1.153***	1.319***	1.168***
	(0.00)	(0.00)	(0.00)	(0.00)
keyword novelty	0.556***	0.586***	0.477***	0.587***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	7447925	650357	3306793	3490775
Log Likeli.	-1195516.2	-73143.0	-413035.6	-677098.5

Exponentiated coefficients. Sample: Papers published between 1999-2011.
* p<0.05 ** p<0.01 *** p<0.001. Robust standard errors in parentheses.
Dependent variable: Pairwise keyword novelty (top 5%).

Table 12 – Determinants of top 1% Pairwise Keyword Novelty. Logit models.

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
novkk_disc_catdom_top1				
#authors	1.110***	1.058***	1.046***	1.115***
	(0.00)	(0.01)	(0.01)	(0.00)
USA-Can	1.148***	1.142***	1.160***	1.071***
	(0.01)	(0.04)	(0.02)	(0.02)
EUR	1.103***	1.124**	1.111***	1.024
	(0.01)	(0.04)	(0.02)	(0.02)
Asia	1.088***	1.149***	1.155***	0.955**
	(0.01)	(0.05)	(0.02)	(0.02)
Other regions	0.912***	0.912	1.102***	0.752***
	(0.01)	(0.05)	(0.03)	(0.02)
#keywords	1.147***	1.092***	1.256***	1.100***
	(0.00)	(0.00)	(0.00)	(0.00)
keyword novelty	0.529***	0.541***	0.461***	0.568***
	(0.00)	(0.01)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	7447925	648633	3306793	3490775
Log Likeli.	-337968.4	-19314.1	-115503.0	-195227.2

Exponentiated coefficients. Sample: Papers published between 1999-2011.
* p<0.05 ** p<0.01 *** p<0.001. Robust standard errors in parentheses.
Dependent variable: Pairwise keyword novelty (top 1%).

Table 13 – Determinants of top 10% Pairwise Keyword Novelty. Logit models.

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
novkk_disc_catdom_top10				
#locations	1.202***	1.167***	1.203***	1.111***
	(0.00)	(0.01)	(0.01)	(0.01)
#keywords	1.253***	1.197***	1.339***	1.220***
	(0.00)	(0.00)	(0.00)	(0.00)
keyword novelty	0.570***	0.617***	0.482***	0.600***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	7447925	650357	3306793	3490775
Log Likeli.	-1968123.7	-125095.7	-699637.2	-1090424.6

Exponentiated coefficients. Sample: Papers published between 1999-2011.
* p<0.05 ** p<0.01 *** p<0.001. Robust standard errors in parentheses.
Dependent variable: Pairwise keyword novelty (top 10%).

Table 14 – Determinants of top 5% Pairwise Keyword Novelty. Logit models.

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
novkk_disc_catdom_top5				
#locations	1.206***	1.147***	1.216***	1.103***
	(0.01)	(0.02)	(0.01)	(0.01)
#keywords	1.218***	1.156***	1.320***	1.175***
	(0.00)	(0.00)	(0.00)	(0.00)
keyword novelty	0.554***	0.585***	0.475***	0.587***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	7447925	650357	3306793	3490775
Log Likeli.	-1198705.8	-73293.8	-413202.7	-679525.7

Exponentiated coefficients. Sample: Papers published between 1999-2011.
* p<0.05 ** p<0.01 *** p<0.001. Robust standard errors in parentheses.
Dependent variable: Pairwise keyword novelty (top 5%).

Table 15 – Determinants of top 1% Pairwise Keyword Novelty. Logit models.

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
novkk_disc_catdom_top1				
#locations	1.222***	1.148***	1.201***	1.120***
	(0.01)	(0.04)	(0.02)	(0.01)
#keywords	1.151***	1.095***	1.256***	1.106***
	(0.00)	(0.00)	(0.00)	(0.00)
keyword novelty	0.527***	0.540***	0.459***	0.568***
	(0.00)	(0.01)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	7447925	648633	3306793	3490775
Log Likeli.	-338759.9	-19335.1	-115539.5	-195865.6

Exponentiated coefficients. Sample: Papers published between 1999-2011.
* p<0.05 ** p<0.01 *** p<0.001. Robust standard errors in parentheses.
Dependent variable: Pairwise keyword novelty (top 1%).

Section 4.2. Novelty as a Predictor of Impact and Excellence

Table 16 – Predicting Forward Citations (generalized negative binomial models).

	Full sample		Hum & SS		Hard Sc & Eng		Life Sc	
	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha
Pairwise keyword novelty	1.430***	0.884***	1.322***	0.804***	1.481***	0.885***	1.377***	0.902***
	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)
#keywords	1.088***	0.955***	1.070***	0.977***	1.073***	0.951***	1.092***	0.955***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)
Constant	1.731***	2.432***	1.073***	2.705***	1.266***	2.389***	2.395***	2.338***
	(0.01)	(0.04)	(0.01)	(0.06)	(0.01)	(0.03)	(0.01)	(0.07)
Year dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	7442453		650186		3302875		3489392	
Log Likeli.	-15238286.2		-954738.1		-6310191.5		-7937954.4	

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: Number of forward citations within 3 years.
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.

Table 17 – Predicting Forward Citations (generalized negative binomial models).

	Full sample		Hum & SS		Hard Sc & Eng		Life Sc	
	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha
Pairwise keyword novelty	1.411***	0.896***	1.334***	0.798***	1.428***	0.882***	1.379***	0.926***
	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)
#keywords	1.091***	0.960***	1.065***	0.983***	1.082***	0.960***	1.092***	0.958***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Constant	3.882***	2.065***	2.885***	2.441***	2.830***	2.088***	5.309***	1.999***
	(0.01)	(0.03)	(0.04)	(0.04)	(0.02)	(0.02)	(0.03)	(0.05)
Year dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	6615448		553206		2925656		3136586	
Log Likeli.	-17664002.6		-1107671.1		-7385412.2		-9136590.9	

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: Number of forward citations within 5 years.
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.

Table 18 – Predicting "Big Hit" Probabilities (logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top10_cl_kk_c_f3				
Pairwise keyword novelty	1.475***	1.380***	1.508***	1.490***
	(0.01)	(0.02)	(0.01)	(0.01)
#keywords	1.133***	1.116***	1.133***	1.140***
	(0.00)	(0.00)	(0.00)	(0.00)
Constant	0.038***	0.046***	0.039***	0.034***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	7447925	650357	3306793	3490775
Log Likeli.	-1752847.8	-133864.4	-834254.7	-782060.3

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05 ** p<0.01 *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 10%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.

Table 19 – Predicting "Big Hit" Probabilities (logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top10_cl_kk_c_f5				
Pairwise keyword novelty	1.490***	1.434***	1.509***	1.511***
	(0.01)	(0.02)	(0.01)	(0.01)
#keywords	1.134***	1.115***	1.134***	1.141***
	(0.00)	(0.00)	(0.00)	(0.00)
Constant	0.041***	0.051***	0.043***	0.035***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	6620920	553377	2929574	3137969
Log Likeli.	-1625044.6	-117681.0	-781481.3	-723181.7

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05 ** p<0.01 *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 10%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.

Table 20 – Predicting "Big Hit" Probabilities (logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top5_cl_kk_c_f3				
Pairwise keyword novelty	1.503***	1.448***	1.510***	1.547***
	(0.01)	(0.03)	(0.01)	(0.01)
#keywords	1.138***	1.124***	1.140***	1.140***
	(0.00)	(0.00)	(0.00)	(0.00)
Constant	0.016***	0.022***	0.016***	0.015***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	7447925	650357	3306793	3490775
Log Likeli.	-997108.9	-79205.2	-484818.6	-431224.9

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05 ** p<0.01 *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 5%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.

Table 21 – Predicting "Big Hit" Probabilities (logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top5_cl_kk_c_f5				
Pairwise keyword novelty	1.514***	1.460***	1.509***	1.566***
	(0.01)	(0.03)	(0.02)	(0.01)
#keywords	1.140***	1.122***	1.145***	1.141***
	(0.00)	(0.00)	(0.00)	(0.00)
Constant	0.017***	0.023***	0.019***	0.015***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	6620920	553377	2929574	3137969
Log Likeli.	-920512.9	-68708.5	-453093.8	-396959.0

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05 ** p<0.01 *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 5%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.

Table 22 – Predicting Forward Citations (Variant 1, generalized negative binomial models).

	Full sample		Hum & SS		Hard Sc & Eng		Life Sc	
	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha
Pairwise keyword novelty	1.455***	0.886***	1.340***	0.803***	1.560***	0.890***	1.369***	0.902***
	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)
#keywords	1.087***	0.955***	1.070***	0.977***	1.073***	0.950***	1.092***	0.955***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)
Constant	1.701***	2.437***	1.054***	2.738***	1.235***	2.395***	2.341***	2.350***
	(0.01)	(0.04)	(0.01)	(0.06)	(0.01)	(0.03)	(0.01)	(0.07)
Year dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	7442453		650186		3302875		3489392	
Log Likeli.	-15235950.0		-954591.1		-6307393.1		-7938594.4	

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: Number of forward citations within 3 years.
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.
Pairwise keyword novelty computed as Variant 1.

Table 23 – Predicting Forward Citations (Variant 1, generalized negative binomial models).

	Full sample		Hum & SS		Hard Sc & Eng		Life Sc	
	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha
Pairwise keyword novelty	1.431***	0.898***	1.349***	0.801***	1.490***	0.883***	1.371***	0.929***
	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)
#keywords	1.091***	0.960***	1.065***	0.983***	1.082***	0.960***	1.092***	0.958***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Constant	3.818***	2.069***	2.837***	2.469***	2.767***	2.095***	5.190***	2.004***
	(0.01)	(0.03)	(0.04)	(0.04)	(0.02)	(0.02)	(0.03)	(0.05)
Year dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	6615448		553206		2925656		3136586	
Log Likeli.	-17662204.4		-1107547.0		-7383246.3		-9137256.5	

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: Number of forward citations within 5 years.
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.
Pairwise keyword novelty computed as Variant 1.

Table 24 – Predicting "Big Hit" Probabilities (Variant 1, logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top10_cl_kk_c_f3				
Pairwise keyword novelty	1.499***	1.396***	1.568***	1.488***
	(0.01)	(0.02)	(0.01)	(0.01)
#keywords	1.133***	1.116***	1.133***	1.139***
	(0.00)	(0.00)	(0.00)	(0.00)
Constant	0.037***	0.045***	0.038***	0.033***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	7447925	650357	3306793	3490775
Log Likeli.	-1752564.2	-133842.9	-833961.9	-782078.7

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 10%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.
Pairwise keyword novelty computed as Variant 1.

Table 25 – Predicting "Big Hit" Probabilities (Variant 1, logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top10_cl_kk_c_f5				
Pairwise keyword novelty	1.510***	1.435***	1.563***	1.511***
	(0.01)	(0.02)	(0.01)	(0.01)
#keywords	1.134***	1.115***	1.134***	1.141***
	(0.00)	(0.00)	(0.00)	(0.00)
Constant	0.040***	0.050***	0.042***	0.034***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	6620920	553377	2929574	3137969
Log Likeli.	-1624818.1	-117671.1	-781237.6	-723190.1

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 10%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.
Pairwise keyword novelty computed as Variant 1.

Table 26 – Predicting "Big Hit" Probabilities (Variant 1, logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top5_cl_kk_c_f3				
Pairwise keyword novelty	1.519***	1.442***	1.559***	1.537***
	(0.01)	(0.03)	(0.02)	(0.01)
#keywords	1.138***	1.124***	1.141***	1.140***
	(0.00)	(0.00)	(0.00)	(0.00)
Constant	0.016***	0.021***	0.016***	0.014***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	7447925	650357	3306793	3490775
Log Likeli.	-997020.7	-79202.5	-484705.3	-431262.7

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 5%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.
Pairwise keyword novelty computed as Variant 1.

Table 27 – Predicting "Big Hit" Probabilities (Variant 1, logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top5_cl_kk_c_f5				
Pairwise keyword novelty	1.529***	1.454***	1.552***	1.561***
	(0.01)	(0.03)	(0.02)	(0.01)
#keywords	1.140***	1.122***	1.145***	1.140***
	(0.00)	(0.00)	(0.00)	(0.00)
Constant	0.017***	0.022***	0.019***	0.015***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	6620920	553377	2929574	3137969
Log Likeli.	-920436.4	-68705.8	-453001.8	-396980.9

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 5%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.
Pairwise keyword novelty computed as Variant 1.

Table 28 – Predicting Forward Citations (Variant 2, generalized negative binomial models).

	Full sample		Hum & SS		Hard Sc & Eng		Life Sc	
	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha
Pairwise keyword novelty	1.371***	0.888***	1.301***	0.796***	1.393***	0.899***	1.329***	0.902***
	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)	(0.01)
#keywords	1.079***	0.955***	1.063***	0.982***	1.068***	0.953***	1.082***	0.954***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)
Constant	1.813***	2.425***	1.104***	2.653***	1.302***	2.373***	2.510***	2.348***
	(0.01)	(0.04)	(0.02)	(0.06)	(0.01)	(0.03)	(0.01)	(0.06)
Year dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	7442453		650186		3302875		3489392	
Log Likeli.	-15244739.4		-954792.9		-6313595.4		-7940783.0	

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: Number of forward citations within 3 years.
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.
Pairwise keyword novelty computed as Variant 2.

Table 29 – Predicting Forward Citations (Variant 2, generalized negative binomial models).

	Full sample		Hum & SS		Hard Sc & Eng		Life Sc	
	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha
Pairwise keyword novelty	1.353***	0.900***	1.309***	0.796***	1.347***	0.894***	1.330***	0.926***
	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)	(0.01)
#keywords	1.083***	0.961***	1.058***	0.987***	1.078***	0.963***	1.083***	0.958***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Constant	4.051***	2.056***	2.966***	2.401***	2.900***	2.073***	5.555***	2.005***
	(0.02)	(0.02)	(0.04)	(0.04)	(0.02)	(0.02)	(0.03)	(0.05)
Year dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	6615448		553206		2925656		3136586	
Log Likeli.	-17670361.3		-1107757.3		-7388406.4		-9139608.1	

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: Number of forward citations within 5 years.
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.
Pairwise keyword novelty computed as Variant 2.

Table 30 – Predicting "Big Hit" Probabilities (Variant 2, logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top10_cl_kk_c_f3				
Pairwise keyword novelty	1.401***	1.334***	1.407***	1.416***
	(0.01)	(0.02)	(0.01)	(0.01)
#keywords	1.124***	1.109***	1.126***	1.127***
	(0.00)	(0.00)	(0.00)	(0.00)
Constant	0.039***	0.047***	0.041***	0.036***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	7447925	650357	3306793	3490775
Log Likeli.	-1753827.7	-133892.4	-834798.1	-782563.9

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05 ** p<0.01 *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 10%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.
Pairwise keyword novelty computed as Variant 2.

Table 31 – Predicting "Big Hit" Probabilities (Variant 2, logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top10_cl_kk_c_f5				
Pairwise keyword novelty	1.417***	1.378***	1.411***	1.440***
	(0.01)	(0.02)	(0.01)	(0.01)
#keywords	1.125***	1.107***	1.127***	1.128***
	(0.00)	(0.00)	(0.00)	(0.00)
Constant	0.043***	0.052***	0.045***	0.038***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	6620920	553377	2929574	3137969
Log Likeli.	-1625973.3	-117712.9	-781987.1	-723656.4

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05 ** p<0.01 *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 10%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.
Pairwise keyword novelty computed as Variant 2.

Table 32 – Predicting "Big Hit" Probabilities (Variant 2, logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top5_cl_kk_c_f3				
Pairwise keyword novelty	1.415***	1.363***	1.386***	1.468***
	(0.01)	(0.03)	(0.01)	(0.01)
#keywords	1.128***	1.116***	1.134***	1.126***
	(0.00)	(0.00)	(0.00)	(0.00)
Constant	0.017***	0.022***	0.017***	0.016***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	7447925	650357	3306793	3490775
Log Likeli.	-997682.9	-79237.1	-485135.6	-431504.1

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05 ** p<0.01 *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 5%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.
Pairwise keyword novelty computed as Variant 2.

Table 33 – Predicting "Big Hit" Probabilities (Variant 2, logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top5_cl_kk_c_f5				
Pairwise keyword novelty	1.425***	1.374***	1.390***	1.483***
	(0.01)	(0.03)	(0.01)	(0.01)
#keywords	1.130***	1.114***	1.138***	1.126***
	(0.00)	(0.00)	(0.00)	(0.00)
Constant	0.018***	0.023***	0.020***	0.016***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	6620920	553377	2929574	3137969
Log Likeli.	-921068.3	-68735.8	-453382.6	-397242.0

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05 ** p<0.01 *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 5%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.
Pairwise keyword novelty computed as Variant 2.

Table 34 – Predicting Forward Citations (Variant 3, generalized negative binomial models).

	Full sample		Hum & SS		Hard Sc & Eng		Life Sc	
	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha
Pairwise keyword novelty	1.425***	0.885***	1.327***	0.794***	1.471***	0.888***	1.373***	0.903***
	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)
#keywords	1.088***	0.955***	1.070***	0.978***	1.073***	0.950***	1.092***	0.955***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)
Constant	1.739***	2.428***	1.073***	2.704***	1.266***	2.389***	2.395***	2.338***
	(0.01)	(0.04)	(0.01)	(0.06)	(0.01)	(0.03)	(0.01)	(0.07)
Year dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	7442453		650186		3302875		3489392	
Log Likeli.	-15239168.6		-954617.7		-6310638.2		-7938450.4	

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: Number of forward citations within 3 years.
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.
Pairwise keyword novelty computed as Variant 3.

Table 35 – Predicting Forward Citations (Variant 3, generalized negative binomial models).

	Full sample		Hum & SS		Hard Sc & Eng		Life Sc	
	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha
Pairwise keyword novelty	1.407***	0.897***	1.337***	0.793***	1.422***	0.885***	1.375***	0.927***
	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)
#keywords	1.091***	0.960***	1.064***	0.983***	1.082***	0.960***	1.092***	0.958***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Constant	3.899***	2.062***	2.887***	2.444***	2.831***	2.089***	5.307***	1.999***
	(0.01)	(0.03)	(0.04)	(0.04)	(0.02)	(0.02)	(0.03)	(0.05)
Year dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	6615448		553206		2925656		3136586	
Log Likeli.	-17664760.2		-1107564.3		-7385729.2		-9137051.3	

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: Number of forward citations within 5 years.
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.
Pairwise keyword novelty computed as Variant 3.

Table 36 – Predicting "Big Hit" Probabilities (Variant 3, logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top10_cl_kk_c_f3				
Pairwise keyword novelty	1.476***	1.382***	1.504***	1.485***
	(0.01)	(0.02)	(0.01)	(0.01)
#keywords	1.133***	1.116***	1.133***	1.140***
	(0.00)	(0.00)	(0.00)	(0.00)
Constant	0.038***	0.046***	0.039***	0.034***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	7447925	650357	3306793	3490775
Log Likeli.	-1752888.5	-133853.0	-834297.1	-782130.2

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05 ** p<0.01 *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 10%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.
Pairwise keyword novelty computed as Variant 3.

Table 37 – Predicting "Big Hit" Probabilities (Variant 3, logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top10_cl_kk_c_f5				
Pairwise keyword novelty	1.495***	1.434***	1.515***	1.508***
	(0.01)	(0.02)	(0.01)	(0.01)
#keywords	1.135***	1.115***	1.134***	1.142***
	(0.00)	(0.00)	(0.00)	(0.00)
Constant	0.041***	0.051***	0.043***	0.035***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	6620920	553377	2929574	3137969
Log Likeli.	-1625034.8	-117669.9	-781473.2	-723240.3

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05 ** p<0.01 *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 10%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.
Pairwise keyword novelty computed as Variant 3.

Table 38 – Predicting "Big Hit" Probabilities (Variant 3, logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top5_cl_kk_c_f3				
Pairwise keyword novelty	1.505***	1.439***	1.510***	1.539***
	(0.01)	(0.03)	(0.01)	(0.01)
#keywords	1.138***	1.124***	1.141***	1.141***
	(0.00)	(0.00)	(0.00)	(0.00)
Constant	0.016***	0.022***	0.016***	0.015***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	7447925	650357	3306793	3490775
Log Likeli.	-997119.7	-79202.8	-484826.0	-431274.9

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05 ** p<0.01 *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 5%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.
Pairwise keyword novelty computed as Variant 3.

Table 39 – Predicting "Big Hit" Probabilities (Variant 3, logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top5_cl_kk_c_f5				
Pairwise keyword novelty	1.520***	1.447***	1.518***	1.560***
	(0.01)	(0.03)	(0.02)	(0.01)
#keywords	1.140***	1.122***	1.145***	1.141***
	(0.00)	(0.00)	(0.00)	(0.00)
Constant	0.018***	0.023***	0.019***	0.015***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	6620920	553377	2929574	3137969
Log Likeli.	-920505.5	-68707.2	-453080.8	-397002.4

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05 ** p<0.01 *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 5%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.
Pairwise keyword novelty computed as Variant 3.

Table 40 – Predicting Forward Citations (generalized negative binomial models).

	Full sample		Hum & SS		Hard Sc & Eng		Life Sc	
	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha
Journal reference novelty	1.131***	0.928***	1.219***	1.000	1.204***	0.938***	1.088***	0.921***
	(0.00)	(0.00)	(0.01)	(0.02)	(0.01)	(0.01)	(0.00)	(0.00)
#references	1.001***	0.995***	1.011***	0.999**	1.001***	0.995***	1.002***	0.994***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Constant	5.275***	1.859***	1.636***	1.354***	3.995***	1.792***	6.566***	1.917***
	(0.07)	(0.03)	(0.07)	(0.10)	(0.08)	(0.05)	(0.11)	(0.04)
Year dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	5583731		462672		2362111		2758948	
Log Likeli.	-12366366.8		-740331.1		-4978827.1		-6614843.4	

Exponentiated coefficients
Sample: Papers published between .
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: Number of forward citations within 3 years.
Journal reference novelty is a dummy equal to one if the paper is in the top 10% most novel papers.

Table 41 – Predicting Forward Citations (generalized negative binomial models).

	Full sample		Hum & SS		Hard Sc & Eng		Life Sc	
	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha
Journal reference novelty	1.098***	0.973***	1.193***	1.041**	1.143***	0.992	1.063***	0.959***
	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)
#references	1.018***	0.997***	1.012***	0.998***	1.022***	0.997***	1.017***	0.997***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Constant	5.812***	1.508***	3.985***	1.374***	4.160***	1.632***	7.860***	1.486***
	(0.06)	(0.03)	(0.15)	(0.08)	(0.07)	(0.06)	(0.11)	(0.03)
Year dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	4835869		382401		2031210		2422258	
Log Likeli.	-13635540.1		-813850.2		-5487388.6		-7309691.7	

Exponentiated coefficients
Sample: Papers published between .
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: Number of forward citations within 5 years.
Journal reference novelty is a dummy equal to one if the paper is in the top 10% most novel papers.

Table 42 – Predicting "Big Hit" Probabilities (logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top10_cl_kk_c_f3				
Journal reference novelty	1.118***	1.373***	1.086***	1.117***
	(0.01)	(0.03)	(0.01)	(0.01)
#references	1.023***	1.019***	1.027***	1.021***
	(0.00)	(0.00)	(0.00)	(0.00)
Constant	0.066***	0.074***	0.075***	0.060***
	(0.00)	(0.01)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	4835869	382401	2031210	2422258
Log Likeli.	-1241972.6	-84344.0	-569728.6	-584401.7

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05 ** p<0.01 *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 10%).
Journal reference novelty is a dummy equal to one if the paper is in the top 10% most novel papers.

Table 43 – Predicting "Big Hit" Probabilities (logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top10_cl_kk_c_f5				
Journal reference novelty	1.123***	1.310***	1.102***	1.120***
	(0.01)	(0.03)	(0.01)	(0.01)
#references	1.024***	1.021***	1.028***	1.022***
	(0.00)	(0.00)	(0.00)	(0.00)
Constant	0.074***	0.079***	0.083***	0.067***
	(0.00)	(0.01)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	4835869	382401	2031210	2422258
Log Likeli.	-1292626.2	-89639.5	-598485.2	-600472.9

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05 ** p<0.01 *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 10%).
Journal reference novelty is a dummy equal to one if the paper is in the top 10% most novel papers.

Table 44 – Predicting "Big Hit" Probabilities (logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top5_cl_kk_c_f3				
Journal reference novelty	1.144***	1.437***	1.118***	1.135***
	(0.01)	(0.04)	(0.01)	(0.01)
#references	1.023***	1.020***	1.027***	1.020***
	(0.00)	(0.00)	(0.00)	(0.00)
Constant	0.030***	0.041***	0.035***	0.028***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	4835869	382401	2031210	2422258
Log Likeli.	-718995.3	-50788.5	-339686.9	-326150.2

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05 ** p<0.01 *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 5%).
Journal reference novelty is a dummy equal to one if the paper is in the top 10% most novel papers.

Table 45 – Predicting "Big Hit" Probabilities (logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top5_cl_kk_c_f5				
Journal reference novelty	1.158***	1.398***	1.150***	1.142***
	(0.01)	(0.04)	(0.01)	(0.01)
#references	1.024***	1.022***	1.028***	1.021***
	(0.00)	(0.00)	(0.00)	(0.00)
Constant	0.034***	0.035***	0.040***	0.030***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	4835869	382401	2031210	2422258
Log Likeli.	-746527.6	-53552.5	-355909.6	-334366.1

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05 ** p<0.01 *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 5%).
Journal reference novelty is a dummy equal to one if the paper is in the top 10% most novel papers.

Section 4.3. Novelty, and the Crowd

Table 46 – Predicting Forward Citations (generalized negative binomial models: novelty and commoness.

	Full sample		Hum & SS		Hard Sc & Eng		Life Sc	
	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha
novelty=1/commonness=0	1.385***	0.897***	1.280***	0.822***	1.536***	0.887***	1.285***	0.917***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
novelty=0/commonness=1	1.397***	0.925***	1.377***	0.833***	1.521***	0.966***	1.302***	0.905***
	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
novelty=1/commonness=1	1.617***	0.876***	1.461***	0.755***	1.704***	0.913***	1.543***	0.878***
	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)
#keywords	1.081***	0.956***	1.066***	0.980***	1.067***	0.957***	1.085***	0.954***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)
Year dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	6615448		553206		2925656		3136586	
Log Likeli.	-13391561.6		-778377.7		-5501972.8		-7078950.2	

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: Number of forward citations within 3 years.

Table 47 – Predicting "Big Hit" Probabilities (logit models: novelty and commoness.

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top10_cl_kk_c_f3				
novelty=1/commonness=0	1.411***	1.296***	1.528***	1.364***
	(0.01)	(0.03)	(0.02)	(0.02)
novelty=0/commonness=1	1.576***	1.513***	1.665***	1.514***
	(0.01)	(0.02)	(0.01)	(0.01)
novelty=1/commonness=1	1.743***	1.582***	1.794***	1.762***
	(0.01)	(0.03)	(0.02)	(0.01)
#keywords	1.125***	1.113***	1.122***	1.132***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	6620920	553377	2929574	3137969
Log Likeli.	-1544940.9	-109768.7	-733280.4	-699383.6

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05 **, p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 10%).

Table 48 – Predicting Forward Citations (generalized negative binomial models: novelty and commoness.

	Full sample		Hum & SS		Hard Sc & Eng		Life Sc	
	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha
novelty=1/commonness=0	1.357***	0.901***	1.296***	0.828***	1.457***	0.879***	1.285***	0.933***
	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
novelty=0/commonness=1	1.369***	0.934***	1.374***	0.843***	1.471***	0.963***	1.290***	0.925***
	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
novelty=1/commonness=1	1.584***	0.888***	1.477***	0.757***	1.626***	0.898***	1.537***	0.908***
	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)
#keywords	1.086***	0.960***	1.062***	0.983***	1.078***	0.963***	1.087***	0.957***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	6615448		553206		2925656		3136586	
Log Likeli.	-17635152.9		-1106245.9		-7369328.3		-9124918.7	

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: Number of forward citations within 5 years.

Table 49 – Predicting "Big Hit" Probabilities (logit models: novelty and commoness.

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top10_cl_kk_c_f5				
novelty=1/commonness=0	1.408***	1.325***	1.488***	1.386***
	(0.01)	(0.03)	(0.02)	(0.02)
novelty=0/commonness=1	1.566***	1.507***	1.644***	1.514***
	(0.01)	(0.02)	(0.01)	(0.01)
novelty=1/commonness=1	1.760***	1.668***	1.797***	1.791***
	(0.01)	(0.03)	(0.02)	(0.01)
#keywords	1.128***	1.113***	1.126***	1.135***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	6620920	553377	2929574	3137969
Log Likeli.	-1618421.6	-117405.6	-777517.2	-720564.4

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05 **, p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 10%).

Table 50 – Predicting "Big Hit" Probabilities (logit models: novelty and commoness.

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top5_cl_kk_c_f3				
novelty=1/commonness=0	1.409***	1.297***	1.511***	1.381***
	(0.02)	(0.05)	(0.03)	(0.02)
novelty=0/commonness=1	1.643***	1.521***	1.746***	1.579***
	(0.01)	(0.03)	(0.01)	(0.01)
novelty=1/commonness=1	1.817***	1.656***	1.849***	1.875***
	(0.01)	(0.04)	(0.02)	(0.02)
#keywords	1.130***	1.122***	1.131***	1.132***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	6620920	553377	2929574	3137969
Log Likeli.	-876869.8	-64633.5	-425521.2	-384982.6

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05 **, p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 5%).

Table 51 – Predicting "Big Hit" Probabilities (logit models: novelty and commoness.

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top5_cl_kk_c_f5				
novelty=1/commonness=0	1.417***	1.369***	1.473***	1.415***
	(0.02)	(0.05)	(0.03)	(0.02)
novelty=0/commonness=1	1.626***	1.560***	1.713***	1.570***
	(0.01)	(0.03)	(0.01)	(0.01)
novelty=1/commonness=1	1.821***	1.707***	1.834***	1.894***
	(0.01)	(0.04)	(0.02)	(0.02)
#keywords	1.134***	1.120***	1.137***	1.135***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	6620920	553377	2929574	3137969
Log Likeli.	-916655.6	-68543.4	-450709.6	-395483.9

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05 **, p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 5%).

Table 52 – Predicting Forward Citations (generalized negative binomial models: interaction of novelty and commoness).

	Full sample		Hum & SS		Hard Sc & Eng		Life Sc	
	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha
Pairwise keyword novelty	1.385***	0.897***	1.280***	0.822***	1.536***	0.887***	1.285***	0.917***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Commonness	1.397***	0.925***	1.377***	0.833***	1.521***	0.966***	1.302***	0.905***
	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
Novelty*Commonnness	0.836***	1.055***	0.828***	1.103***	0.729***	1.065***	0.922***	1.057***
	(0.00)	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
#keywords	1.081***	0.956***	1.066***	0.980***	1.067***	0.957***	1.085***	0.954***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)
Year dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	6615448		553206		2925656		3136586	
Log Likeli.	-13391561.6		-778377.7		-5501972.8		-7078950.2	

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: Number of forward citations within 3 years.
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.
Commonness is a dummy that takes the value 1 if >0.

Table 53 – Predicting "Big Hit" Probabilities (logit models: interaction of novelty and commoness).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top10_cl_kk_c_f3				
Pairwise keyword novelty	1.411***	1.296***	1.528***	1.364***
	(0.01)	(0.03)	(0.02)	(0.02)
Commonness	1.576***	1.513***	1.665***	1.514***
	(0.01)	(0.02)	(0.01)	(0.01)
Novelty*Commonnness	0.783***	0.806***	0.705***	0.853***
	(0.01)	(0.03)	(0.01)	(0.01)
#keywords	1.125***	1.113***	1.122***	1.132***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	6620920	553377	2929574	3137969
Log Likeli.	-1544940.9	-109768.7	-733280.4	-699383.6

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05 ** p<0.01 *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 10%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.
Commonness is a dummy that takes the value 1 if >0.

Table 54 – Predicting "Big Hit" Probabilities (logit models: interaction of novelty and commoness).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top5_cl_kk_c_f3				
Pairwise keyword novelty	1.409***	1.297***	1.511***	1.381***
	(0.02)	(0.05)	(0.03)	(0.02)
Commonness	1.643***	1.521***	1.746***	1.579***
	(0.01)	(0.03)	(0.01)	(0.01)
Novelty*Commonnness	0.785***	0.840***	0.701***	0.860***
	(0.01)	(0.04)	(0.02)	(0.02)
#keywords	1.130***	1.122***	1.131***	1.132***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	6620920	553377	2929574	3137969
Log Likeli.	-876869.8	-64633.5	-425521.2	-384982.6

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05 ** p<0.01 *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 5%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.
Commonness is a dummy that takes the value 1 if >0.

Table 55 – Predicting Forward Citations (generalized negative binomial models: interaction of novelty and commoness).

	Full sample		Hum & SS		Hard Sc & Eng		Life Sc	
	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha
Pairwise keyword novelty	1.357***	0.901***	1.296***	0.828***	1.457***	0.879***	1.285***	0.933***
	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Commonness	1.369***	0.934***	1.374***	0.843***	1.471***	0.963***	1.290***	0.925***
	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
Novelty*Commonnness	0.853***	1.055***	0.829***	1.085***	0.759***	1.061***	0.927***	1.053***
	(0.00)	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
#keywords	1.086***	0.960***	1.062***	0.983***	1.078***	0.963***	1.087***	0.957***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	6615448		553206		2925656		3136586	
Log Likeli.	-17635152.9		-1106245.9		-7369328.3		-9124918.7	

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: Number of forward citations within 5 years.
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.
Commonness is a dummy that takes the value 1 if >0.

Table 56 – Predicting "Big Hit" Probabilities (logit models: interaction of novelty and commoness).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top10_cl_kk_c_f5				
Pairwise keyword novelty	1.408***	1.325***	1.488***	1.386***
	(0.01)	(0.03)	(0.02)	(0.02)
Commonness	1.566***	1.507***	1.644***	1.514***
	(0.01)	(0.02)	(0.01)	(0.01)
Novelty*Commonnness	0.798***	0.835***	0.735***	0.853***
	(0.01)	(0.03)	(0.01)	(0.01)
#keywords	1.128***	1.113***	1.126***	1.135***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	6620920	553377	2929574	3137969
Log Likeli.	-1618421.6	-117405.6	-777517.2	-720564.4

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05 ** p<0.01 *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 10%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.
Commonness is a dummy that takes the value 1 if >0.

Table 57 – Predicting "Big Hit" Probabilities (logit models: interaction of novelty and commoness).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top5_cl_kk_c_f5				
Pairwise keyword novelty	1.417***	1.369***	1.473***	1.415***
	(0.02)	(0.05)	(0.03)	(0.02)
Commonness	1.626***	1.560***	1.713***	1.570***
	(0.01)	(0.03)	(0.01)	(0.01)
Novelty*Commonnness	0.791***	0.799***	0.727***	0.852***
	(0.01)	(0.04)	(0.02)	(0.02)
#keywords	1.134***	1.120***	1.137***	1.135***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	6620920	553377	2929574	3137969
Log Likeli.	-916655.6	-68543.4	-450709.6	-395483.9

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05 ** p<0.01 *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 5%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.
Commonness is a dummy that takes the value 1 if >0.

Section 4.4. Does it Pay to Address more Novel Research Questions?

Table 58 – Predicting Forward Citations With More Controls (generalized negative binomial models).

	Full sample		Hum & SS		Hard Sc & Eng		Life Sc	
	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha
Pairwise keyword novelty	1.324***	0.876***	1.306***	0.821***	1.188***	0.884***	1.325***	0.904***
	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
Keyword novelty	0.963***	0.975***	1.089***	0.918***	0.867***	1.010***	1.008***	0.965***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
#keywords	1.082***	0.963***	1.059***	0.988***	1.086***	0.969***	1.077***	0.955***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
#co-authors	1.113***	0.948***	1.116***	0.954***	1.153***	0.921***	1.091***	0.957***
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
#institutions	1.002*	1.043***	1.022***	1.034***	0.967***	1.083***	1.014***	1.030***
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)
Constant	1.633***	3.199***	0.358***	5.101***	2.961***	2.378***	1.468***	3.314***
	(0.01)	(0.04)	(0.01)	(0.23)	(0.03)	(0.04)	(0.02)	(0.06)
Year dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	6579761		543702		2906544		3129515	
Log Likeli.	-13218164.6		-757628.5		-5434151.6		-6981262.3	

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: Number of forward citations within 3 years.
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.

Table 59 – Predicting Forward Citations With More Controls (generalized negative binomial models).

	Full sample		Hum & SS		Hard Sc & Eng		Life Sc	
	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha
Pairwise keyword novelty	1.314***	0.892***	1.322***	0.818***	1.183***	0.895***	1.321***	0.924***
	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
Keyword novelty	0.971***	0.983***	1.083***	0.921***	0.892***	1.020***	1.003**	0.969***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
#keywords	1.085***	0.966***	1.056***	0.990***	1.092***	0.970***	1.080***	0.960***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
#co-authors	1.112***	0.946***	1.108***	0.957***	1.148***	0.921***	1.092***	0.956***
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
#institutions	0.999	1.040***	1.027***	1.024**	0.962***	1.076***	1.011***	1.028***
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)
Constant	3.394***	2.621***	0.989	4.618***	5.218***	1.991***	3.378***	2.735***
	(0.02)	(0.03)	(0.02)	(0.16)	(0.06)	(0.03)	(0.04)	(0.05)
Year dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	6579761		543702		2906544		3129515	
Log Likeli.	-17427718.6		-1077185.8		-7290555.6		-9014690.0	

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: Number of forward citations within 5 years.
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.

Table 60 – Predicting "Big Hit" Probabilities With More Controls (logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top10_cl_kk_c_f3				
Pairwise keyword novelty	1.247***	1.234***	1.216***	1.270***
	(0.01)	(0.02)	(0.01)	(0.01)
Keyword novelty	0.871***	0.932***	0.881***	0.836***
	(0.00)	(0.01)	(0.00)	(0.00)
#keywords	1.123***	1.111***	1.127***	1.125***
	(0.00)	(0.00)	(0.00)	(0.00)
#co-authors	1.147***	1.152***	1.215***	1.137***
	(0.00)	(0.01)	(0.01)	(0.00)
#institutions	1.005*	1.023**	0.938***	1.020***
	(0.00)	(0.01)	(0.01)	(0.00)
Constant	0.075***	0.042***	0.072***	0.095***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes
N	6579761	543702	2906544	3129515
Log Likeli.	-1502950.5	-106573.1	-718726.6	-673412.0

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 10%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.

Table 61 – Predicting "Big Hit" Probabilities With More Controls (logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top10_cl_kk_c_f5				
Pairwise keyword novelty	1.267***	1.279***	1.247***	1.287***
	(0.01)	(0.02)	(0.01)	(0.01)
Keyword novelty	0.877***	0.916***	0.899***	0.831***
	(0.00)	(0.01)	(0.00)	(0.00)
#keywords	1.125***	1.112***	1.128***	1.129***
	(0.00)	(0.00)	(0.00)	(0.00)
#co-authors	1.148***	1.151***	1.206***	1.144***
	(0.00)	(0.01)	(0.01)	(0.00)
#institutions	1.005	1.024**	0.944***	1.017***
	(0.00)	(0.01)	(0.01)	(0.00)
Constant	0.076***	0.053***	0.066***	0.103***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes
N	6579761	543702	2906544	3129515
Log Likeli.	-1574309.5	-113756.1	-762071.1	-693581.4

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 10%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.

Table 62 – Predicting "Big Hit" Probabilities With More Controls (logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top5_cl_kk_c_f3				
Pairwise keyword novelty	1.241***	1.260***	1.205***	1.275***
	(0.01)	(0.03)	(0.01)	(0.01)
Keyword novelty	0.858***	0.920***	0.879***	0.805***
	(0.00)	(0.01)	(0.00)	(0.00)
#keywords	1.127***	1.120***	1.132***	1.125***
	(0.00)	(0.00)	(0.00)	(0.00)
#co-authors	1.148***	1.150***	1.211***	1.143***
	(0.00)	(0.01)	(0.01)	(0.00)
#institutions	1.001	1.020	0.927***	1.021***
	(0.00)	(0.01)	(0.01)	(0.00)
Constant	0.035***	0.021***	0.029***	0.053***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes
N	6579761	543702	2906544	3129515
Log Likeli.	-851690.6	-62821.6	-416691.6	-369355.7

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 5%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.

Table 63 – Predicting "Big Hit" Probabilities With More Controls (logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top5_cl_kk_c_f5				
Pairwise keyword novelty	1.260***	1.298***	1.226***	1.297***
	(0.01)	(0.03)	(0.01)	(0.01)
Keyword novelty	0.866***	0.916***	0.894***	0.806***
	(0.00)	(0.01)	(0.00)	(0.00)
#keywords	1.129***	1.118***	1.136***	1.128***
	(0.00)	(0.00)	(0.00)	(0.00)
#co-authors	1.148***	1.143***	1.209***	1.145***
	(0.00)	(0.01)	(0.01)	(0.00)
#institutions	1.000	1.025*	0.926***	1.018***
	(0.00)	(0.01)	(0.01)	(0.00)
Constant	0.035***	0.023***	0.030***	0.054***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes
N	6579761	543702	2906544	3129515
Log Likeli.	-890539.0	-66510.5	-441319.3	-379578.4

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 5%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.

Table 64 – Predicting Forward Citations With More Controls. (Variant 1, generalized negative binomial models).

	Full sample		Hum & SS		Hard Sc & Eng		Life Sc	
	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha
Pairwise keyword novelty	1.332***	0.874***	1.348***	0.815***	1.220***	0.898***	1.310***	0.898***
	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
Keyword novelty								
#keywords	1.082***	0.962***	1.060***	0.986***	1.083***	0.966***	1.079***	0.956***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
novk_disc_alldis_c	0.960***	0.980***	1.092***	0.916***	0.873***	1.014***	0.998*	0.968***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
#co-authors	1.115***	0.949***	1.110***	0.953***	1.165***	0.929***	1.092***	0.956***
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
#institutions	1.002	1.058***	1.022***	1.033***	0.961***	1.091***	1.012***	1.034***
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
Constant	1.626***	3.136***	0.329***	5.586***	2.777***	2.323***	1.556***	3.296***
	(0.01)	(0.03)	(0.01)	(0.23)	(0.03)	(0.04)	(0.02)	(0.05)
Year dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	7402656		639126		3282006		3481524	
Log Likeli.	-15007200.4		-927718.3		-6213938.1		-7815137.7	

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: Number of forward citations within 3 years.
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.
Pairwise keyword novelty computed as Variant 1.

Table 65 – Predicting Forward Citations With More Controls. (Variant 1, generalized negative binomial models).

	Full sample		Hum & SS		Hard Sc & Eng		Life Sc	
	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha
Pairwise keyword novelty	1.321***	0.890***	1.355***	0.810***	1.208***	0.894***	1.311***	0.922***
	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
Keyword novelty								
#keywords	1.084***	0.965***	1.057***	0.990***	1.089***	0.970***	1.081***	0.959***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
novk_disc_alldis_c	0.969***	0.983***	1.090***	0.915***	0.896***	1.015***	0.995***	0.972***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
#co-authors	1.112***	0.946***	1.106***	0.957***	1.147***	0.920***	1.093***	0.955***
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
#institutions	0.999	1.040***	1.028***	1.024**	0.962***	1.077***	1.011***	1.028***
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)
Constant	3.435***	2.620***	0.901***	5.034***	5.104***	2.084***	3.580***	2.691***
	(0.02)	(0.03)	(0.02)	(0.18)	(0.06)	(0.03)	(0.04)	(0.04)
Year dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	6579761		543702		2906544		3129515	
Log Likeli.	-17426026.7		-1076847.9		-7290171.3		-9015064.6	

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: Number of forward citations within 5 years.
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.
Pairwise keyword novelty computed as Variant 1.

Table 66 – Predicting "Big Hit" Probabilities With More Controls. (Variant 1, logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top10_cl_kk_c_f3				
Pairwise keyword novelty	1.232***	1.273***	1.218***	1.238***
	(0.01)	(0.02)	(0.01)	(0.01)
#keywords	1.122***	1.110***	1.126***	1.123***
	(0.00)	(0.00)	(0.00)	(0.00)
novk_disc_alldis_c	0.872***	0.947***	0.883***	0.830***
	(0.00)	(0.01)	(0.00)	(0.00)
#co-authors	1.147***	1.146***	1.219***	1.137***
	(0.00)	(0.01)	(0.01)	(0.00)
#institutions	1.001	1.022**	0.928***	1.017***
	(0.00)	(0.01)	(0.00)	(0.00)
Constant	0.073***	0.037***	0.070***	0.102***
	(0.00)	(0.00)	(0.00)	(0.00)
$\overline{\text{K}}$eyword novelty				
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes
N	7402656	639126	3282006	3481524
Log Likeli.	-1697967.6	-129646.2	-813218.6	-749978.8

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 10%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.
Pairwise keyword novelty computed as Variant 1.

Table 67 – Predicting "Big Hit" Probabilities With More Controls. (Variant 1, logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top10_cl_kk_c_f5				
Pairwise keyword novelty	1.252***	1.294***	1.250***	1.255***
	(0.01)	(0.02)	(0.01)	(0.01)
#keywords	1.123***	1.110***	1.126***	1.126***
	(0.00)	(0.00)	(0.00)	(0.00)
novk_disc_alldis_c	0.876***	0.935***	0.898***	0.825***
	(0.00)	(0.01)	(0.00)	(0.00)
#co-authors	1.149***	1.151***	1.207***	1.144***
	(0.00)	(0.01)	(0.01)	(0.00)
#institutions	1.004	1.024**	0.943***	1.016***
	(0.00)	(0.01)	(0.01)	(0.00)
Constant	0.076***	0.045***	0.067***	0.112***
	(0.00)	(0.00)	(0.00)	(0.00)
$\overline{\text{K}}$eyword novelty				
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes
N	6579761	543702	2906544	3129515
Log Likeli.	-1574276.0	-113782.7	-761958.6	-693459.1

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 10%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.
Pairwise keyword novelty computed as Variant 1.

Table 68 – Predicting "Big Hit" Probabilities With More Controls. (Variant 1, logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top5_cl_kk_c_f3				
Pairwise keyword novelty	1.216***	1.306***	1.188***	1.228***
	(0.01)	(0.03)	(0.01)	(0.01)
#keywords	1.125***	1.117***	1.132***	1.123***
	(0.00)	(0.00)	(0.00)	(0.00)
novk_disc_alldis_c	0.858***	0.939***	0.878***	0.797***
	(0.00)	(0.01)	(0.00)	(0.00)
#co-authors	1.146***	1.147***	1.207***	1.140***
	(0.00)	(0.01)	(0.01)	(0.00)
#institutions	0.996	1.018	0.916***	1.019***
	(0.00)	(0.01)	(0.01)	(0.00)
Constant	0.034***	0.018***	0.030***	0.059***
	(0.00)	(0.00)	(0.00)	(0.00)
$\overline{\text{K}}$eyword novelty				
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes
N	7402656	639126	3282006	3481524
Log Likeli.	-964352.4	-76760.1	-472227.5	-411861.6

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 5%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.
Pairwise keyword novelty computed as Variant 1.

Table 69 – Predicting "Big Hit" Probabilities With More Controls. (Variant 1, logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top5_cl_kk_c_f5				
Pairwise keyword novelty	1.237***	1.309***	1.212***	1.254***
	(0.01)	(0.03)	(0.01)	(0.01)
#keywords	1.127***	1.116***	1.134***	1.124***
	(0.00)	(0.00)	(0.00)	(0.00)
novk_disc_alldis_c	0.864***	0.936***	0.890***	0.799***
	(0.00)	(0.01)	(0.00)	(0.00)
#co-authors	1.148***	1.143***	1.210***	1.145***
	(0.00)	(0.01)	(0.01)	(0.00)
#institutions	1.000	1.025*	0.926***	1.017***
	(0.00)	(0.01)	(0.01)	(0.00)
Constant	0.036***	0.019***	0.032***	0.060***
	(0.00)	(0.00)	(0.00)	(0.00)
$\overline{\text{K}}$eyword novelty				
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes
N	6579761	543702	2906544	3129515
Log Likeli.	-890550.2	-66526.4	-441256.2	-379503.0

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 5%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.
Pairwise keyword novelty computed as Variant 1.

Table 70 – Predicting Forward Citations With More Controls. (Variant 2, generalized negative binomial models).

	Full sample		Hum & SS		Hard Sc & Eng		Life Sc	
	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha
Pairwise keyword novelty	1.269***	0.885***	1.278***	0.830***	1.137***	0.902***	1.272***	0.912***
	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
Keyword novelty								
#keywords	1.077***	0.965***	1.052***	0.991***	1.087***	0.968***	1.071***	0.958***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
novk_min_c	0.955***	0.983***	1.093***	0.909***	0.864***	1.020***	0.999	0.968***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
#co-authors	1.116***	0.949***	1.110***	0.955***	1.167***	0.930***	1.092***	0.956***
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
#institutions	1.002	1.057***	1.023***	1.032***	0.960***	1.090***	1.012***	1.034***
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
Constant	1.767***	3.012***	0.345***	5.711***	3.004***	2.182***	1.623***	3.230***
	(0.01)	(0.03)	(0.01)	(0.23)	(0.03)	(0.03)	(0.02)	(0.05)
Year dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	7402656		639126		3282006		3481524	
Log Likeli.	-15013714.0		-927852.4		-6215542.6		-7817296.0	

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: Number of forward citations within 3 years.
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.
Pairwise keyword novelty computed as Variant 2.

Table 71 – Predicting Forward Citations With More Controls. (Variant 2, generalized negative binomial models).

	Full sample		Hum & SS		Hard Sc & Eng		Life Sc	
	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha
Pairwise keyword novelty	1.261***	0.899***	1.287***	0.826***	1.131***	0.902***	1.272***	0.931***
	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
Keyword novelty								
#keywords	1.080***	0.968***	1.049***	0.994**	1.092***	0.972***	1.073***	0.961***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
novk_min_c	0.964***	0.985***	1.092***	0.912***	0.887***	1.020***	0.996***	0.972***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
#co-authors	1.113***	0.946***	1.106***	0.958***	1.148***	0.921***	1.093***	0.956***
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
#institutions	0.998	1.040***	1.028***	1.023**	0.961***	1.076***	1.011***	1.028***
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)
Constant	3.726***	2.529***	0.938**	5.004***	5.532***	1.965***	3.743***	2.650***
	(0.03)	(0.03)	(0.02)	(0.17)	(0.06)	(0.03)	(0.04)	(0.05)
Year dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	6579761		543702		2906544		3129515	
Log Likeli.	-17432147.4		-1076994.2		-7291607.1		-9017386.1	

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: Number of forward citations within 5 years.
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.
Pairwise keyword novelty computed as Variant 2.

Table 72 – Predicting "Big Hit" Probabilities With More Controls. (Variant 2, logit models).

	Full sample	Domain 1	Domain 2	Domain 3
top10_cl_kk_c_f3				
Pairwise keyword novelty	1.206***	1.231***	1.157***	1.229***
	(0.01)	(0.02)	(0.01)	(0.01)
#keywords	1.120***	1.105***	1.127***	1.119***
	(0.00)	(0.00)	(0.00)	(0.00)
novk_min_c	0.874***	0.954***	0.882***	0.832***
	(0.00)	(0.01)	(0.00)	(0.00)
#co-authors	1.147***	1.147***	1.220***	1.137***
	(0.00)	(0.01)	(0.01)	(0.00)
#institutions	1.001	1.021**	0.929***	1.017***
	(0.00)	(0.01)	(0.00)	(0.00)
Constant	0.073***	0.036***	0.071***	0.101***
	(0.00)	(0.00)	(0.00)	(0.00)
$\overline{\text{K}}$eyword novelty				
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes
N	7402656	639126	3282006	3481524
Log Likeli.	-1698844.4	-129696.7	-813640.4	-750360.4

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 10%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.
Pairwise keyword novelty computed as Variant 2.

Table 73 – Predicting "Big Hit" Probabilities With More Controls. (Variant 2, logit models).

	Full sample	Domain 1	Domain 2	Domain 3
top10_cl_kk_c_f5				
Pairwise keyword novelty	1.227***	1.266***	1.186***	1.249***
	(0.01)	(0.02)	(0.01)	(0.01)
#keywords	1.120***	1.105***	1.126***	1.122***
	(0.00)	(0.00)	(0.00)	(0.00)
novk_min_c	0.878***	0.943***	0.896***	0.828***
	(0.00)	(0.01)	(0.00)	(0.00)
#co-authors	1.149***	1.152***	1.207***	1.144***
	(0.00)	(0.01)	(0.01)	(0.00)
#institutions	1.004	1.024**	0.944***	1.016***
	(0.00)	(0.01)	(0.01)	(0.00)
Constant	0.076***	0.044***	0.069***	0.111***
	(0.00)	(0.00)	(0.00)	(0.00)
$\overline{\text{K}}$eyword novelty				
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes
N	6579761	543702	2906544	3129515
Log Likeli.	-1575105.2	-113827.8	-762348.4	-693855.1

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 10%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.
Pairwise keyword novelty computed as Variant 2.

Table 74 – Predicting "Big Hit" Probabilities With More Controls. (Variant 2, logit models).

	Full sample	Domain 1	Domain 2	Domain 3
top5_cl_kk_c_f3				
Pairwise keyword novelty	1.197***	1.250***	1.130***	1.238***
	(0.01)	(0.02)	(0.01)	(0.01)
#keywords	1.123***	1.112***	1.133***	1.120***
	(0.00)	(0.00)	(0.00)	(0.00)
novk_min_c	0.863***	0.944***	0.880***	0.802***
	(0.00)	(0.01)	(0.00)	(0.00)
#co-authors	1.146***	1.148***	1.207***	1.141***
	(0.00)	(0.01)	(0.01)	(0.00)
#institutions	0.996	1.018	0.917***	1.019***
	(0.00)	(0.01)	(0.01)	(0.00)
Constant	0.033***	0.018***	0.029***	0.056***
	(0.00)	(0.00)	(0.00)	(0.00)
$\overline{\text{K}}$eyword novelty				
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes
N	7402656	639126	3282006	3481524
Log Likeli.	-964892.1	-76793.7	-472479.6	-412104.9

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 5%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.
Pairwise keyword novelty computed as Variant 2.

Table 75 – Predicting "Big Hit" Probabilities With More Controls. (Variant 2, logit models).

	Full sample	Domain 1	Domain 2	Domain 3
top5_cl_kk_c_f5				
Pairwise keyword novelty	1.213***	1.259***	1.154***	1.254***
	(0.01)	(0.03)	(0.01)	(0.01)
#keywords	1.124***	1.111***	1.134***	1.121***
	(0.00)	(0.00)	(0.00)	(0.00)
novk_min_c	0.867***	0.945***	0.890***	0.802***
	(0.00)	(0.01)	(0.00)	(0.00)
#co-authors	1.148***	1.144***	1.210***	1.146***
	(0.00)	(0.01)	(0.01)	(0.00)
#institutions	1.000	1.024*	0.926***	1.017***
	(0.00)	(0.01)	(0.01)	(0.00)
Constant	0.035***	0.018***	0.031***	0.058***
	(0.00)	(0.00)	(0.00)	(0.00)
$\overline{\text{K}}$eyword novelty				
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes
N	6579761	543702	2906544	3129515
Log Likeli.	-891045.2	-66557.1	-441470.0	-379746.5

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 5%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.
Pairwise keyword novelty computed as Variant 2.

Table 76 – Predicting Forward Citations With More Controls. (Variant 3, generalized negative binomial models).

	Full sample		Hum & SS		Hard Sc & Eng		Life Sc	
	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha
Pairwise keyword novelty	1.316***	0.878***	1.314***	0.822***	1.187***	0.895***	1.315***	0.904***
	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
#keywords	1.083***	0.962***	1.058***	0.987***	1.087***	0.966***	1.079***	0.956***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Keyword novelty	0.960***	0.981***	1.098***	0.908***	0.869***	1.018***	1.004***	0.967***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
#co-authors	1.115***	0.949***	1.110***	0.955***	1.166***	0.930***	1.091***	0.956***
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
#institutions	1.002	1.058***	1.023***	1.032***	0.961***	1.090***	1.012***	1.034***
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
Constant	1.638***	3.092***	0.326***	5.863***	2.842***	2.244***	1.496***	3.289***
	(0.01)	(0.03)	(0.01)	(0.24)	(0.03)	(0.04)	(0.02)	(0.06)
Year dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	7402656		639126		3282006		3481524	
Log Likeli.	-15009750.8		-927626.5		-6214747.7		-7815147.1	

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: Number of forward citations within 3 years.
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.
Pairwise keyword novelty computed as Variant 3.

Table 77 – Predicting Forward Citations With More Controls. (Variant 3, generalized negative binomial models).

	Full sample		Hum & SS		Hard Sc & Eng		Life Sc	
	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha
Pairwise keyword novelty	1.310***	0.893***	1.324***	0.820***	1.182***	0.895***	1.318***	0.926***
	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
#keywords	1.085***	0.966***	1.055***	0.990***	1.093***	0.970***	1.081***	0.960***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Keyword novelty	0.969***	0.984***	1.097***	0.911***	0.892***	1.018***	1.000	0.971***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
#co-authors	1.112***	0.946***	1.105***	0.959***	1.148***	0.920***	1.093***	0.956***
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
#institutions	0.999	1.040***	1.028***	1.023**	0.962***	1.076***	1.011***	1.028***
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)
Constant	3.457***	2.581***	0.884***	5.115***	5.233***	2.015***	3.449***	2.680***
	(0.03)	(0.03)	(0.02)	(0.17)	(0.06)	(0.03)	(0.04)	(0.05)
Year dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	6579761		543702		2906544		3129515	
Log Likeli.	-17428080.1		-1076752.8		-7290801.5		-9015090.0	

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: Number of forward citations within 5 years.
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.
Pairwise keyword novelty computed as Variant 3.

Table 78 – Predicting "Big Hit" Probabilities With More Controls. (Variant 3, logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top10_cl_kk_c_f3				
Pairwise keyword novelty	1.253***	1.273***	1.222***	1.268***
	(0.01)	(0.02)	(0.01)	(0.01)
Keyword novelty	0.878***	0.957***	0.888***	0.836***
	(0.00)	(0.01)	(0.00)	(0.00)
#keywords	1.123***	1.110***	1.128***	1.126***
	(0.00)	(0.00)	(0.00)	(0.00)
#co-authors	1.146***	1.147***	1.219***	1.137***
	(0.00)	(0.01)	(0.01)	(0.00)
#institutions	1.001	1.021**	0.929***	1.017***
	(0.00)	(0.01)	(0.00)	(0.00)
Constant	0.068***	0.034***	0.067***	0.093***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes
N	7402656	639126	3282006	3481524
Log Likeli.	-1698503.0	-129671.2	-813491.4	-750218.3

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 10%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.
Pairwise keyword novelty computed as Variant 3.

Table 79 – Predicting "Big Hit" Probabilities With More Controls. (Variant 3, logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top10_cl_kk_c_f5				
Pairwise keyword novelty	1.278***	1.311***	1.261***	1.287***
	(0.01)	(0.02)	(0.01)	(0.01)
Keyword novelty	0.883***	0.947***	0.903***	0.832***
	(0.00)	(0.01)	(0.00)	(0.00)
#keywords	1.124***	1.111***	1.127***	1.129***
	(0.00)	(0.00)	(0.00)	(0.00)
#co-authors	1.148***	1.151***	1.206***	1.144***
	(0.00)	(0.01)	(0.01)	(0.00)
#institutions	1.005*	1.024**	0.944***	1.017***
	(0.00)	(0.01)	(0.01)	(0.00)
Constant	0.071***	0.041***	0.064***	0.102***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes
N	6579761	543702	2906544	3129515
Log Likeli.	-1574738.7	-113801.2	-762160.2	-693720.3

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 10%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.
Pairwise keyword novelty computed as Variant 3.

Table 80 – Predicting "Big Hit" Probabilities With More Controls. (Variant 3, logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top5_cl_kk_c_f3				
Pairwise keyword novelty	1.253***	1.317***	1.215***	1.272***
	(0.01)	(0.03)	(0.01)	(0.01)
Keyword novelty	0.868***	0.948***	0.887***	0.806***
	(0.00)	(0.01)	(0.00)	(0.00)
#keywords	1.127***	1.117***	1.133***	1.126***
	(0.00)	(0.00)	(0.00)	(0.00)
#co-authors	1.145***	1.147***	1.206***	1.140***
	(0.00)	(0.01)	(0.01)	(0.00)
#institutions	0.997	1.018	0.917***	1.020***
	(0.00)	(0.01)	(0.01)	(0.00)
Constant	0.031***	0.017***	0.027***	0.052***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes
N	7402656	639126	3282006	3481524
Log Likeli.	-964693.5	-76769.8	-472390.4	-412041.3

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 5%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.
Pairwise keyword novelty computed as Variant 3.

Table 81 – Predicting "Big Hit" Probabilities With More Controls. (Variant 3, logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top5_cl_kk_c_f5				
Pairwise keyword novelty	1.275***	1.322***	1.245***	1.294***
	(0.01)	(0.03)	(0.01)	(0.01)
Keyword novelty	0.873***	0.949***	0.898***	0.807***
	(0.00)	(0.01)	(0.00)	(0.00)
#keywords	1.128***	1.116***	1.135***	1.128***
	(0.00)	(0.00)	(0.00)	(0.00)
#co-authors	1.147***	1.143***	1.209***	1.145***
	(0.00)	(0.01)	(0.01)	(0.00)
#institutions	1.001	1.025*	0.927***	1.018***
	(0.00)	(0.01)	(0.01)	(0.00)
Constant	0.032***	0.017***	0.029***	0.054***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes
N	6579761	543702	2906544	3129515
Log Likeli.	-890833.0	-66538.2	-441369.4	-379672.6

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 5%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.
Pairwise keyword novelty computed as Variant 3.

Table 82 – Predicting Forward Citations With More Controls (generalized negative binomial models).

Table 83 – Predicting Forward Citations With More Controls (generalized negative binomial models).

Table 84 – Predicting Forward Citations With More Controls. (Variant 3, generalized negative binomial models).

	Full sample		Hum & SS		Hard Sc & Eng		Life Sc	
	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha
Pairwise keyword novelty	1.222***	0.920***	1.192***	0.915***	1.130***	0.929***	1.239***	0.932***
	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)
Journal reference novelty	1.115***	0.991***	1.082***	0.981***	1.100***	0.987***	1.116***	0.992***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Keyword novelty	0.935***	0.990***	1.028***	1.004	0.873***	1.009***	0.974***	0.976***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
#keywords	1.061***	0.970***	1.035***	0.979***	1.069***	0.973***	1.055***	0.969***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
#references	1.001***	0.995***	1.009***	0.999*	1.001***	0.995***	1.002***	0.995***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
#co-authors	1.069***	0.963***	1.065***	0.969***	1.098***	0.947***	1.057***	0.969***
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
#institutions	1.020***	1.042***	1.022***	1.024***	0.991***	1.068***	1.027***	1.023***
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
Constant	7.377***	2.079***	1.558***	1.263**	10.154***	1.803***	5.803***	2.361***
	(0.11)	(0.04)	(0.07)	(0.11)	(0.22)	(0.06)	(0.13)	(0.07)
Year dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	5568912		459319		2354868		2754725	
Log Likeli.	-12096148.0		-723651.2		-4876090.6		-6470773.3	

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: Number of forward citations within 3 years.
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.
Pairwise keyword novelty computed as Variant 3.

Table 85 – Predicting Forward Citations With More Controls. (Variant 3, generalized negative binomial models).

	Full sample		Hum & SS		Hard Sc & Eng		Life Sc	
	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha
Pairwise keyword novelty	1.206***	0.941***	1.198***	0.915***	1.130***	0.925***	1.223***	0.961***
	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
Journal reference novelty	1.081***	0.991***	1.079***	0.989***	1.068***	0.990***	1.078***	0.992***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Keyword novelty	0.954***	0.996***	1.028***	1.013**	0.914***	1.017***	0.973***	0.979***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
#keywords	1.043***	0.973***	1.030***	0.983***	1.053***	0.976***	1.037***	0.971***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
#references	1.013***	0.997***	1.010***	0.998***	1.017***	0.997***	1.012***	0.997***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
#co-authors	1.070***	0.962***	1.058***	0.973***	1.085***	0.934***	1.062***	0.972***
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
#institutions	1.015***	1.028***	1.030***	1.017*	0.987***	1.061***	1.025***	1.018***
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
Constant	8.257***	1.651***	3.848***	1.213**	8.467***	1.564***	8.386***	1.833***
	(0.10)	(0.04)	(0.17)	(0.09)	(0.16)	(0.06)	(0.14)	(0.06)
Year dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	4823039		379567		2024866		2418606	
Log Likeli.	-13441998.3		-796591.2		-5421947.8		-7203415.5	

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: Number of forward citations within 5 years.
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.
Pairwise keyword novelty computed as Variant 3.

Appendix G: Additional checks

Alternative thresholds to define novel papers

Prediction

Table 86 – Predicting Forward Citations (generalized negative binomial models).

	Full sample		Hum & SS		Hard Sc & Eng		Life Sc	
	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha
Pairwise keyword novelty	1.453***	0.871***	1.288***	0.795***	1.512***	0.862***	1.398***	0.898***
	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)
#keywords	1.091***	0.954***	1.073***	0.976***	1.076***	0.951***	1.095***	0.954***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)
Constant	1.735***	2.432***	1.063***	2.717***	1.265***	2.391***	2.393***	2.339***
	(0.01)	(0.04)	(0.01)	(0.06)	(0.01)	(0.03)	(0.01)	(0.06)
Year dummies(10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	7442453		650186		3302875		3489392	
Log Likeli.	-15248632.1		-955313.8		-6314063.7		-7943007.4	

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: Number of forward citations within 3 years.
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 5% most novel papers.

Table 87 – Predicting Forward Citations (generalized negative binomial models).

	Full sample		Hum & SS		Hard Sc & Eng		Life Sc	
	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha
Pairwise keyword novelty	1.431***	0.884***	1.303***	0.781***	1.449***	0.858***	1.399***	0.924***
	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)
#keywords	1.095***	0.960***	1.068***	0.981***	1.085***	0.960***	1.096***	0.958***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Constant	3.892***	2.065***	2.856***	2.454***	2.829***	2.091***	5.305***	1.999***
	(0.01)	(0.03)	(0.04)	(0.04)	(0.02)	(0.02)	(0.03)	(0.05)
Year dummies(10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	6615448		553206		2925656		3136586	
Log Likeli.	-17674018.8		-1108268.6		-7388838.2		-9141836.2	

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: Number of forward citations within 5 years.
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 5% most novel papers.

Table 88 – Predicting "Big Hit" Probabilities (logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top10_cl_kk_c_f3				
Pairwise keyword novelty	1.481***	1.320***	1.530***	1.488***
	(0.01)	(0.03)	(0.01)	(0.01)
#keywords	1.136***	1.119***	1.135***	1.143***
	(0.00)	(0.00)	(0.00)	(0.00)
Constant	0.038***	0.046***	0.039***	0.034***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies(10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	7447925	650357	3306793	3490775
Log Likeli.	-1754433.4	-133965.2	-834909.0	-782950.0

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05 ** p<0.01 *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 10%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 5% most novel papers.

Table 89 – Predicting "Big Hit" Probabilities (logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top10_cl_kk_c_f5				
Pairwise keyword novelty	1.500***	1.394***	1.537***	1.510***
	(0.01)	(0.03)	(0.02)	(0.01)
#keywords	1.138***	1.118***	1.136***	1.145***
	(0.00)	(0.00)	(0.00)	(0.00)
Constant	0.041***	0.050***	0.044***	0.036***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies(10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	6620920	553377	2929574	3137969
Log Likeli.	-1626579.5	-117779.6	-782089.2	-724074.0

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05 ** p<0.01 *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 10%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 5% most novel papers.

Table 90 – Predicting "Big Hit" Probabilities (logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top5_cl_kk_c_f3				
Pairwise keyword novelty	1.503***	1.391***	1.524***	1.538***
	(0.01)	(0.04)	(0.02)	(0.02)
#keywords	1.141***	1.127***	1.143***	1.144***
	(0.00)	(0.00)	(0.00)	(0.00)
Constant	0.016***	0.021***	0.016***	0.015***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies(10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	7447925	650357	3306793	3490775
Log Likeli.	-997991.3	-79272.1	-485160.1	-431753.3

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05 ** p<0.01 *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 5%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 5% most novel papers.

Table 91 – Predicting "Big Hit" Probabilities (logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top5_cl_kk_c_f5				
Pairwise keyword novelty	1.502***	1.379***	1.509***	1.546***
	(0.01)	(0.04)	(0.02)	(0.02)
#keywords	1.143***	1.125***	1.147***	1.145***
	(0.00)	(0.00)	(0.00)	(0.00)
Constant	0.018***	0.023***	0.019***	0.015***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies(10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	6620920	553377	2929574	3137969
Log Likeli.	-921410.0	-68772.3	-453436.8	-397503.5

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05 ** p<0.01 *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 5%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 5% most novel papers.

Table 92 – Predicting Forward Citations (generalized negative binomial models).

	Full sample		Hum & SS		Hard Sc & Eng		Life Sc	
	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha
Pairwise keyword novelty	1.513***	0.860***	1.269***	0.796***	1.620***	0.838***	1.440***	0.898***
	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)
#keywords	1.094***	0.955***	1.075***	0.975***	1.080***	0.951***	1.098***	0.955***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)
Constant	1.739***	2.432***	1.054***	2.738***	1.258***	2.393***	2.400***	2.338***
	(0.01)	(0.04)	(0.01)	(0.06)	(0.01)	(0.03)	(0.01)	(0.06)
Year dummies(10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	7442453		650186		3302875		3489392	
Log Likeli.	-15260375.3		-955756.2		-6318342.1		-7948960.3	

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: Number of forward citations within 3 years.
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 1% most novel papers.

Table 93 – Predicting Forward Citations (generalized negative binomial models).

	Full sample		Hum & SS		Hard Sc & Eng		Life Sc	
	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha
Pairwise keyword novelty	1.486***	0.875***	1.253***	0.788***	1.536***	0.841***	1.447***	0.922***
	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)
#keywords	1.098***	0.960***	1.069***	0.980***	1.088***	0.960***	1.098***	0.958***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Constant	3.901***	2.066***	2.834***	2.473***	2.820***	2.094***	5.322***	2.001***
	(0.01)	(0.03)	(0.04)	(0.04)	(0.02)	(0.02)	(0.03)	(0.05)
Year dummies(10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	6615448		553206		2925656		3136586	
Log Likeli.	-17685070.4		-1108787.1		-7392476.8		-9147838.9	

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: Number of forward citations within 5 years.
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 1% most novel papers.

Table 94 – Predicting "Big Hit" Probabilities (logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top10_cl_kk_c_f3				
Pairwise keyword novelty	1.503***	1.197***	1.575***	1.509***
	(0.02)	(0.05)	(0.03)	(0.03)
#keywords	1.139***	1.120***	1.138***	1.145***
	(0.00)	(0.00)	(0.00)	(0.00)
Constant	0.038***	0.045***	0.039***	0.034***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies(10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	7447925	650357	3306793	3490775
Log Likeli.	-1755976.8	-134026.2	-835577.9	-783814.2

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05 ** p<0.01 *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 10%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 1% most novel papers.

Table 95 – Predicting "Big Hit" Probabilities (logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top10_cl_kk_c_f5				
Pairwise keyword novelty	1.526***	1.283***	1.590***	1.530***
	(0.02)	(0.06)	(0.03)	(0.03)
#keywords	1.140***	1.119***	1.139***	1.147***
	(0.00)	(0.00)	(0.00)	(0.00)
Constant	0.041***	0.050***	0.044***	0.036***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies(10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	6620920	553377	2929574	3137969
Log Likeli.	-1628112.9	-117853.5	-782726.7	-724946.2

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05 ** p<0.01 *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 10%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 1% most novel papers.

Predicting Forward Citations With More Controls

Table 96 – Predicting "Big Hit" Probabilities (logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top5_cl_kk_c_f3				
Pairwise keyword novelty	1.509***	1.273***	1.568***	1.529***
	(0.03)	(0.07)	(0.04)	(0.04)
#keywords	1.143***	1.128***	1.146***	1.146***
	(0.00)	(0.00)	(0.00)	(0.00)
Constant	0.016***	0.021***	0.016***	0.015***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies(10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	7447925	650357	3306793	3490775
Log Likeli.	-998829.0	-79317.1	-485486.1	-432263.0

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05 ** p<0.01 *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 5%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 1% most novel papers.

Table 97 – Predicting "Big Hit" Probabilities (logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top5_cl_kk_c_f5				
Pairwise keyword novelty	1.514***	1.276***	1.542***	1.551***
	(0.03)	(0.08)	(0.05)	(0.04)
#keywords	1.145***	1.126***	1.149***	1.147***
	(0.00)	(0.00)	(0.00)	(0.00)
Constant	0.018***	0.022***	0.019***	0.015***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies(10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	6620920	553377	2929574	3137969
Log Likeli.	-922177.9	-68807.6	-453730.4	-397982.3

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05 ** p<0.01 *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 5%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 1% most novel papers.

Table 98 – Predicting Forward Citations With More Controls (generalized negative binomial models).

	Full sample		Hum & SS		Hard Sc & Eng		Life Sc	
	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha
Pairwise keyword novelty	1.328***	0.867***	1.284***	0.810***	1.175***	0.888***	1.333***	0.899***
	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)
Keyword novelty	0.956***	0.981***	1.075***	0.924***	0.865***	1.022***	1.000	0.964***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
#keywords	1.086***	0.962***	1.062***	0.986***	1.089***	0.966***	1.082***	0.956***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
#co-authors	1.117***	0.949***	1.115***	0.952***	1.167***	0.930***	1.093***	0.956***
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
#institutions	1.001	1.058***	1.021***	1.034***	0.961***	1.090***	1.012***	1.034***
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
Constant	1.704***	3.090***	0.385***	5.059***	2.961***	2.165***	1.551***	3.364***
	(0.01)	(0.03)	(0.01)	(0.21)	(0.03)	(0.03)	(0.02)	(0.06)
Year dummies(10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	7402656		639126		3282006		3481524	
Log Likeli.	-15016318.8		-928595.8		-6215602.7		-7818700.4	

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: Number of forward citations within 3 years.
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 5% most novel papers.

Table 99 – Predicting Forward Citations With More Controls (generalized negative binomial models).

	Full sample		Hum & SS		Hard Sc & Eng		Life Sc	
	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha
Pairwise keyword novelty	1.319***	0.882***	1.298***	0.793***	1.162***	0.881***	1.333***	0.924***
	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)
Keyword novelty	0.964***	0.983***	1.076***	0.920***	0.887***	1.022***	0.996***	0.969***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
#keywords	1.089***	0.966***	1.059***	0.990***	1.095***	0.970***	1.084***	0.960***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
#co-authors	1.114***	0.946***	1.110***	0.956***	1.148***	0.921***	1.094***	0.956***
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
#institutions	0.998	1.040***	1.026***	1.026**	0.962***	1.076***	1.010***	1.028***
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)
Constant	3.609***	2.595***	1.032	4.669***	5.491***	1.945***	3.591***	2.736***
	(0.03)	(0.03)	(0.03)	(0.16)	(0.06)	(0.03)	(0.04)	(0.05)
Year dummies(10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	6579761		543702		2906544		3129515	
Log Likeli.	-17434590.9		-1077712.3		-7291613.8		-9018836.8	

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: Number of forward citations within 5 years.
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 5% most novel papers.

Table 100 – Predicting "Big Hit" Probabilities With More Controls (logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top10_cl_kk_c_f3				
Pairwise keyword novelty	1.216***	1.184***	1.174***	1.245***
	(0.01)	(0.02)	(0.01)	(0.01)
Keyword novelty	0.867***	0.922***	0.878***	0.830***
	(0.00)	(0.01)	(0.00)	(0.00)
#keywords	1.127***	1.114***	1.131***	1.129***
	(0.00)	(0.00)	(0.00)	(0.00)
#co-authors	1.148***	1.149***	1.220***	1.138***
	(0.00)	(0.01)	(0.01)	(0.00)
#institutions	1.000	1.021**	0.929***	1.016***
	(0.00)	(0.01)	(0.00)	(0.00)
Constant	0.078***	0.045***	0.073***	0.100***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies(10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes
N	7402656	639126	3282006	3481524
Log Likeli.	-1698583.1	-129681.6	-813560.7	-750439.7

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 10%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 5% most novel papers.

Table 101 – Predicting "Big Hit" Probabilities With More Controls (logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top10_cl_kk_c_f5				
Pairwise keyword novelty	1.240***	1.236***	1.210***	1.261***
	(0.01)	(0.03)	(0.01)	(0.01)
Keyword novelty	0.871***	0.911***	0.892***	0.825***
	(0.00)	(0.01)	(0.00)	(0.00)
#keywords	1.128***	1.114***	1.130***	1.132***
	(0.00)	(0.00)	(0.00)	(0.00)
#co-authors	1.150***	1.154***	1.207***	1.146***
	(0.00)	(0.01)	(0.01)	(0.00)
#institutions	1.004	1.024**	0.944***	1.016***
	(0.00)	(0.01)	(0.01)	(0.00)
Constant	0.081***	0.055***	0.071***	0.110***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies(10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes
N	6579761	543702	2906544	3129515
Log Likeli.	-1574939.1	-113805.2	-762292.3	-693955.5

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 10%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 5% most novel papers.

Table 102 – Predicting "Big Hit" Probabilities With More Controls (logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top10_cl_kk_c_f3				
Pairwise keyword novelty	1.216***	1.184***	1.174***	1.245***
	(0.01)	(0.02)	(0.01)	(0.01)
Keyword novelty	0.867***	0.922***	0.878***	0.830***
	(0.00)	(0.01)	(0.00)	(0.00)
#keywords	1.127***	1.114***	1.131***	1.129***
	(0.00)	(0.00)	(0.00)	(0.00)
#co-authors	1.148***	1.149***	1.220***	1.138***
	(0.00)	(0.01)	(0.01)	(0.00)
#institutions	1.000	1.021**	0.929***	1.016***
	(0.00)	(0.01)	(0.00)	(0.00)
Constant	0.078***	0.045***	0.073***	0.100***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies(10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes
N	7402656	639126	3282006	3481524
Log Likeli.	-1698583.1	-129681.6	-813560.7	-750439.7

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 10%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 5% most novel papers.

Table 103 – Predicting "Big Hit" Probabilities With More Controls (logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top10_cl_kk_c_f5				
Pairwise keyword novelty	1.240***	1.236***	1.210***	1.261***
	(0.01)	(0.03)	(0.01)	(0.01)
Keyword novelty	0.871***	0.911***	0.892***	0.825***
	(0.00)	(0.01)	(0.00)	(0.00)
#keywords	1.128***	1.114***	1.130***	1.132***
	(0.00)	(0.00)	(0.00)	(0.00)
#co-authors	1.150***	1.154***	1.207***	1.146***
	(0.00)	(0.01)	(0.01)	(0.00)
#institutions	1.004	1.024**	0.944***	1.016***
	(0.00)	(0.01)	(0.01)	(0.00)
Constant	0.081***	0.055***	0.071***	0.110***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies(10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes
N	6579761	543702	2906544	3129515
Log Likeli.	-1574939.1	-113805.2	-762292.3	-693955.5

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 10%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 5% most novel papers.

Table 104 – Predicting Forward Citations With More Controls (generalized negative binomial models).

	Full sample		Hum & SS		Hard Sc & Eng		Life Sc	
	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha
Pairwise keyword novelty	1.352***	0.864***	1.278***	0.817***	1.181***	0.904***	1.360***	0.902***
	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
Keyword novelty	0.948***	0.983***	1.068***	0.926***	0.859***	1.027***	0.991***	0.964***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
#keywords	1.089***	0.962***	1.064***	0.986***	1.091***	0.965***	1.085***	0.957***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
#co-authors	1.119***	0.948***	1.116***	0.952***	1.167***	0.930***	1.095***	0.956***
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
#institutions	1.000	1.058***	1.020***	1.035***	0.960***	1.091***	1.010***	1.034***
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
Constant	1.839***	3.017***	0.402***	5.007***	3.110***	2.079***	1.683***	3.359***
	(0.01)	(0.03)	(0.01)	(0.20)	(0.03)	(0.03)	(0.02)	(0.06)
Year dummies(10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	7402656		639126		3282006		3481524	
Log Likeli.	-15023667.8		-929012.1		-6216437.8		-7823276.6	

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: Number of forward citations within 3 years.
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 1% most novel papers.

Table 105 – Predicting Forward Citations With More Controls (generalized negative binomial models).

	Full sample		Hum & SS		Hard Sc & Eng		Life Sc	
	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha
Pairwise keyword novelty	1.340***	0.879***	1.262***	0.809***	1.161***	0.895***	1.365***	0.922***
	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
Keyword novelty	0.956***	0.985***	1.068***	0.922***	0.882***	1.027***	0.987***	0.968***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
#keywords	1.092***	0.966***	1.061***	0.989***	1.097***	0.969***	1.087***	0.960***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
#co-authors	1.116***	0.946***	1.112***	0.956***	1.149***	0.920***	1.097***	0.956***
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
#institutions	0.997**	1.040***	1.025***	1.026**	0.961***	1.077***	1.009***	1.028***
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)
Constant	3.889***	2.549***	1.082**	4.629***	5.754***	1.869***	3.900***	2.751***
	(0.03)	(0.03)	(0.03)	(0.15)	(0.06)	(0.03)	(0.04)	(0.05)
Year dummies(10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	6579761		543702		2906544		3129515	
Log Likeli.	-17441695.3		-1078207.9		-7292439.6		-9023435.1	

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: Number of forward citations within 5 years.
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 1% most novel papers.

Table 106 – Predicting "Big Hit" Probabilities With More Controls (logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top10_cl_kk_c_f3				
Pairwise keyword novelty	1.162***	1.057	1.099***	1.218***
	(0.01)	(0.05)	(0.02)	(0.02)
Keyword novelty	0.861***	0.918***	0.872***	0.823***
	(0.00)	(0.00)	(0.00)	(0.00)
#keywords	1.128***	1.114***	1.132***	1.130***
	(0.00)	(0.00)	(0.00)	(0.00)
#co-authors	1.150***	1.150***	1.220***	1.140***
	(0.00)	(0.01)	(0.01)	(0.00)
#institutions	0.999	1.020*	0.928***	1.015***
	(0.00)	(0.01)	(0.00)	(0.00)
Constant	0.083***	0.046***	0.078***	0.108***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies(10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes
N	7402656	639126	3282006	3481524
Log Likeli.	-1699000.2	-129705.2	-813670.5	-750713.0

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 10%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 1% most novel papers.

Table 107 – Predicting "Big Hit" Probabilities With More Controls (logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top10_cl_kk_c_f5				
Pairwise keyword novelty	1.192***	1.119*	1.145***	1.233***
	(0.02)	(0.05)	(0.03)	(0.02)
Keyword novelty	0.865***	0.905***	0.885***	0.818***
	(0.00)	(0.01)	(0.00)	(0.00)
#keywords	1.129***	1.115***	1.132***	1.133***
	(0.00)	(0.00)	(0.00)	(0.00)
#co-authors	1.152***	1.155***	1.207***	1.148***
	(0.00)	(0.01)	(0.01)	(0.00)
#institutions	1.003	1.023**	0.943***	1.014***
	(0.00)	(0.01)	(0.01)	(0.00)
Constant	0.086***	0.057***	0.076***	0.120***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies(10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes
N	6579761	543702	2906544	3129515
Log Likeli.	-1575402.8	-113836.2	-762432.2	-694241.9

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 10%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 1% most novel papers.

Table 108 – Predicting "Big Hit" Probabilities With More Controls (logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top10_cl_kk_c_f3				
Pairwise keyword novelty	1.162***	1.057	1.099***	1.218***
	(0.01)	(0.05)	(0.02)	(0.02)
Keyword novelty	0.861***	0.918***	0.872***	0.823***
	(0.00)	(0.00)	(0.00)	(0.00)
#keywords	1.128***	1.114***	1.132***	1.130***
	(0.00)	(0.00)	(0.00)	(0.00)
#co-authors	1.150***	1.150***	1.220***	1.140***
	(0.00)	(0.01)	(0.01)	(0.00)
#institutions	0.999	1.020*	0.928***	1.015***
	(0.00)	(0.01)	(0.00)	(0.00)
Constant	0.083***	0.046***	0.078***	0.108***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies(10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes
N	7402656	639126	3282006	3481524
Log Likeli.	-1699000.2	-129705.2	-813670.5	-750713.0

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 10%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 1% most novel papers.

Table 109 – Predicting "Big Hit" Probabilities With More Controls (logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top10_cl_kk_c_f5				
Pairwise keyword novelty	1.192***	1.119*	1.145***	1.233***
	(0.02)	(0.05)	(0.03)	(0.02)
Keyword novelty	0.865***	0.905***	0.885***	0.818***
	(0.00)	(0.01)	(0.00)	(0.00)
#keywords	1.129***	1.115***	1.132***	1.133***
	(0.00)	(0.00)	(0.00)	(0.00)
#co-authors	1.152***	1.155***	1.207***	1.148***
	(0.00)	(0.01)	(0.01)	(0.00)
#institutions	1.003	1.023**	0.943***	1.014***
	(0.00)	(0.01)	(0.01)	(0.00)
Constant	0.086***	0.057***	0.076***	0.120***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies(10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes
N	6579761	543702	2906544	3129515
Log Likeli.	-1575402.8	-113836.2	-762432.2	-694241.9

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 10%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 1% most novel papers.

Replicate results with KeyWords Plus

Table 110 – Predicting Forward Citations (generalized negative binomial models).

	Full sample		Hum & SS		Hard Sc & Eng		Life Sc	
	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha
Pairwise keyword novelty	1.079***	0.966***	1.001	0.965**	1.050***	0.939***	1.094***	0.983***
	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)
#keywords	1.126***	0.945***	1.101***	0.956***	1.113***	0.945***	1.134***	0.942***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Constant	1.395***	2.328***	1.161***	1.871***	1.274***	2.214***	1.549***	2.430***
	(0.01)	(0.02)	(0.02)	(0.05)	(0.01)	(0.02)	(0.01)	(0.03)
Year dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	5795630		476260		2227645		3091725	
Log Likeli.	-12607657.5		-715697.3		-4682657.7		-7196205.6	

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: Number of forward citations within 3 years.
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.

Table 111 – Predicting Forward Citations (generalized negative binomial models).

	Full sample		Hum & SS		Hard Sc & Eng		Life Sc	
	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha
Pairwise keyword novelty	1.075***	0.969***	1.038***	0.961***	1.044***	0.934***	1.085***	0.996
	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
#keywords	1.120***	0.952***	1.103***	0.961***	1.107***	0.950***	1.126***	0.952***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Constant	3.285***	1.936***	3.001***	1.681***	2.993***	1.948***	3.650***	1.963***
	(0.01)	(0.01)	(0.04)	(0.03)	(0.02)	(0.02)	(0.03)	(0.03)
Year dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	5073369		389754		1941718		2741897	
Log Likeli.	-14360434.1		-807454.6		-5322181.1		-8219650.9	

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: Number of forward citations within 5 years.
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.

Table 112 – Predicting "Big Hit" Probabilities (logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top10_cl_kk_c_f3				
Pairwise keyword novelty	0.922***	0.880***	1.001	0.905***
	(0.00)	(0.01)	(0.01)	(0.01)
#keywords	1.148***	1.139***	1.138***	1.185***
	(0.00)	(0.00)	(0.00)	(0.00)
Constant	0.034***	0.052***	0.048***	0.021***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	5797407	476318	2228516	3092573
Log Likeli.	-1458821.2	-103180.1	-628675.1	-722202.6

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05 ** p<0.01 *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 10%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.

Table 113 – Predicting "Big Hit" Probabilities (logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top10_cl_kk_c_f5				
Pairwise keyword novelty	0.926***	0.911***	1.005	0.907***
	(0.01)	(0.02)	(0.01)	(0.01)
#keywords	1.145***	1.147***	1.135***	1.182***
	(0.00)	(0.00)	(0.00)	(0.00)
Constant	0.038***	0.057***	0.053***	0.023***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	5075146	389812	1942589	2742745
Log Likeli.	-1341307.7	-89869.1	-579501.4	-667342.5

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05 ** p<0.01 *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 10%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.

Table 114 – Predicting "Big Hit" Probabilities (logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top5_cl_kk_c_f3				
Pairwise keyword novelty	0.904***	0.851***	0.973*	0.903***
	(0.01)	(0.02)	(0.01)	(0.01)
#keywords	1.151***	1.144***	1.145***	1.190***
	(0.00)	(0.00)	(0.00)	(0.00)
Constant	0.015***	0.024***	0.020***	0.009***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	5797407	476318	2228516	3092573
Log Likeli.	-839658.2	-62418.8	-372907.2	-401352.6

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05 ** p<0.01 *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 5%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.

Table 115 – Predicting "Big Hit" Probabilities (logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top5_cl_kk_c_f5				
Pairwise keyword novelty	0.919***	0.914***	0.984	0.915***
	(0.01)	(0.02)	(0.01)	(0.01)
#keywords	1.149***	1.157***	1.142***	1.187***
	(0.00)	(0.00)	(0.00)	(0.00)
Constant	0.016***	0.024***	0.023***	0.009***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	5075146	389812	1942589	2742745
Log Likeli.	-768739.5	-53716.7	-343132.7	-369085.1

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05 ** p<0.01 *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 5%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.

Table 116 – Predicting Forward Citations With More Controls (generalized negative binomial models).

	Full sample		Hum & SS		Hard Sc & Eng		Life Sc	
	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha
Pairwise keyword novelty	1.070***	0.958***	1.006	0.976*	1.019***	0.922***	1.088***	0.979***
	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)
Keyword novelty	0.988***	0.982***	1.021***	1.016***	0.952***	0.982***	0.994***	0.981***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
#keywords	1.117***	0.952***	1.085***	0.960***	1.115***	0.951***	1.121***	0.950***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
#co-authors	1.077***	0.972***	1.080***	0.966***	1.098***	0.964***	1.067***	0.975***
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
#institutions	1.013***	1.031***	1.001	1.027***	0.993**	1.045***	1.022***	1.022***
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
Constant	1.137***	2.660***	0.727***	1.651***	1.419***	2.657***	1.176***	2.672***
	(0.01)	(0.03)	(0.02)	(0.07)	(0.02)	(0.05)	(0.01)	(0.04)
Year dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	5784594		473936		2223031		3087627	
Log Likeli.	-12473250.8		-707092.3		-4640105.5		-7108736.2	

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: Number of forward citations within 3 years.
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.

Table 117 – Predicting Forward Citations With More Controls (generalized negative binomial models).

	Full sample		Hum & SS		Hard Sc & Eng		Life Sc	
	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha
Pairwise keyword novelty	1.064***	0.964***	1.037***	0.973**	1.014***	0.923***	1.079***	0.993
	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
Keyword novelty	0.982***	0.987***	1.010***	1.022***	0.950***	0.984***	0.989***	0.987***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
#keywords	1.112***	0.956***	1.089***	0.963***	1.109***	0.955***	1.115***	0.957***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
#co-authors	1.077***	0.970***	1.072***	0.970***	1.092***	0.954***	1.069***	0.975***
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
#institutions	1.010***	1.022***	1.009**	1.021***	0.991***	1.042***	1.020***	1.017***
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
Constant	2.856***	2.159***	2.065***	1.423***	3.459***	2.316***	2.934***	2.074***
	(0.02)	(0.03)	(0.04)	(0.05)	(0.04)	(0.04)	(0.03)	(0.04)
Year dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	5063697		387761		1937601		2738335	
Log Likeli.	-14227068.9		-798068.8		-5280414.1		-8133888.8	

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: Number of forward citations within 5 years.
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.

Table 118 – Predicting "Big Hit" Probabilities With More Controls (logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top10_cl_kk_c_f3				
Pairwise keyword novelty	0.855***	0.834***	0.916***	0.849***
	(0.00)	(0.01)	(0.01)	(0.01)
Keyword novelty	0.833***	0.897***	0.857***	0.813***
	(0.00)	(0.00)	(0.00)	(0.00)
#keywords	1.167***	1.137***	1.158***	1.201***
	(0.00)	(0.00)	(0.00)	(0.00)
#co-authors	1.107***	1.116***	1.147***	1.108***
	(0.00)	(0.01)	(0.01)	(0.00)
#institutions	1.019***	1.006	0.966***	1.032***
	(0.00)	(0.01)	(0.01)	(0.00)
Constant	0.102***	0.071***	0.110***	0.077***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes
N	5784594	473936	2223031	3087627
Log Likeli.	-1417744.5	-101263.2	-615554.9	-695475.4

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 10%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.

Papers published in 2001

Table 119 – Predicting "Big Hit" Probabilities With More Controls (logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top10_cl_kk_c_f5				
Pairwise keyword novelty	0.857***	0.853***	0.915***	0.853***
	(0.00)	(0.02)	(0.01)	(0.01)
Keyword novelty	0.826***	0.871***	0.852***	0.809***
	(0.00)	(0.00)	(0.00)	(0.00)
#keywords	1.166***	1.150***	1.155***	1.199***
	(0.00)	(0.00)	(0.00)	(0.00)
#co-authors	1.111***	1.117***	1.140***	1.115***
	(0.00)	(0.01)	(0.01)	(0.00)
#institutions	1.020***	1.005	0.975***	1.031***
	(0.00)	(0.01)	(0.01)	(0.00)
Constant	0.120***	0.098***	0.130***	0.087***
	(0.00)	(0.01)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes
N	5063697	387761	1937601	2738335
Log Likeli.	-1302869.9	-88057.6	-567210.9	-642429.6

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 10%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.

Prediction

Table 120 – Predicting "Big Hit" Probabilities With More Controls (logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top5_cl_kk_c_f3				
Pairwise keyword novelty	0.827***	0.798***	0.878***	0.836***
	(0.01)	(0.02)	(0.01)	(0.01)
Keyword novelty	0.809***	0.874***	0.836***	0.786***
	(0.00)	(0.01)	(0.00)	(0.00)
#keywords	1.175***	1.144***	1.169***	1.209***
	(0.00)	(0.00)	(0.00)	(0.00)
#co-authors	1.112***	1.117***	1.139***	1.117***
	(0.00)	(0.01)	(0.01)	(0.00)
#institutions	1.016***	1.011	0.959***	1.033***
	(0.00)	(0.01)	(0.01)	(0.00)
Constant	0.053***	0.038***	0.055***	0.040***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes
N	5784594	473936	2223031	3087627
Log Likeli.	-813590.1	-61159.7	-364296.1	-384681.9

Exponentiated coefficients
Sample: Papers published between 1999-2011.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 5%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.

Table 121 – Predicting "Big Hit" Probabilities With More Controls (logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top5_cl_kk_c_f5				
Pairwise keyword novelty	0.841***	0.854***	0.882***	0.852***
	(0.01)	(0.02)	(0.01)	(0.01)
Keyword novelty	0.803***	0.862***	0.829***	0.782***
	(0.00)	(0.01)	(0.00)	(0.00)
#keywords	1.175***	1.159***	1.167***	1.208***
	(0.00)	(0.00)	(0.00)	(0.00)
#co-authors	1.113***	1.105***	1.146***	1.119***
	(0.00)	(0.01)	(0.01)	(0.00)
#institutions	1.019***	1.019	0.964***	1.033***
	(0.00)	(0.01)	(0.01)	(0.00)
Constant	0.063***	0.042***	0.069***	0.045***
	(0.00)	(0.00)	(0.00)	(0.00)
Year dummies (10)	Yes	Yes	Yes	Yes
Subdomain dummies (10)	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes
N	5063697	387761	1937601	2738335
Log Likeli.	-744788.7	-52606.7	-334968.0	-354067.1

Exponentiated coefficients
Sample: Papers published between 1999-2009.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 5%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.

Table 122 – Novelty and Citation Impact In The Short To The Long Term.

	Citation window (years)									
Gen. Neg. Bin.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Mean										
Pairwise Author Keywords Novelty	50%	57%	55%	54%	52%	51%	49%	48%	47%	46%
Pairwise KeyWords Plus Novelty	3	8%	8%	8%	7%	7%	7%	6%	5%	5%
Pairwise Journal References Novelty	ns	14%	16%	17%	17%	17%	17%	16%	16%	16%
Ln(alpha)										
Pairwise Author Keywords Novelty	-17%	-15%	-12%	-10%	-9%	-9%	-9%	-8%	-8%	-8%
Pairwise KeyWords Plus Novelty	ns	-3%	-3%	-2%	ns	ns	ns	ns	ns	ns
Pairwise Journal References Novelty	ns	-9%	-9%	-9%	-9%	-10%	-10%	ns	ns	ns

Estimates are obtained from exponentiated coefficients of generalized negative binomial models. Ln(alpha) are obtained from the dispersion estimates of generalized negative binomial models.
Dependent variable: number of forward citations after 1 to 10 years. Novelty variables are dummies indicating if the paper is among the top 5% most novel.
Control variables in all Pairwise Author Keywords Novelty models: number of Author Keywords, publication year dummies, discipline dummies. Control variables in all Pairwise Keywords Plus Novelty models: number of Keywords Plus, publication year dummies, discipline dummies. Control variables in all Pairwise Journal References Novelty models: number of references, publication year dummies, discipline dummies. All detailed regression results are available upon request.

Table 123 – Novelty and Citation Impact In The Short To The Long Term.

	Citation window (years)									
Gen. Neg. Bin.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Mean										
Pairwise Author Keywords Novelty	64%	73%	71%	68%	65%	63%	60%	58%	56%	54%
Pairwise KeyWords Plus Novelty	ns	7%	10%	9%	9%	9%	8%	8%	7%	6%
Pairwise Journal References Novelty	ns	10%	13%	14%	14%	14%	14%	14%	14%	14%
Ln(alpha)										
Pairwise Author Keywords Novelty	-27%	-21%	-16%	-13%	-12%	-11%	-11%	-11%	-10%	-10%
Pairwise KeyWords Plus Novelty	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
Pairwise Journal References Novelty	ns	-10%	-11%	-10%	-8%	ns	ns	ns	ns	ns

Estimates are obtained from exponentiated coefficients of generalized negative binomial models. Ln(alpha) are obtained from the dispersion estimates of generalized negative binomial models.
Dependent variable: number of forward citations after 1 to 10 years. Novelty variables are dummies indicating if the paper is among the top 1% most novel.
Control variables in all Pairwise Author Keywords Novelty models: number of Author Keywords, publication year dummies, discipline dummies. Control variables in all Pairwise Keywords Plus Novelty models: number of Keywords Plus, publication year dummies, discipline dummies. Control variables in all Pairwise Journal References Novelty models: number of references, publication year dummies, discipline dummies. All detailed regression results are available upon request.

Table 124 – Predicting Forward Citations (generalized negative binomial models).

	Full sample		Hum & SS		Hard Sc & Eng		Life Sc	
	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha
Pairwise keyword novelty	1.510***	0.897***	1.355***	0.852***	1.621***	0.917***	1.434***	0.900***
	(0.01)	(0.01)	(0.04)	(0.04)	(0.02)	(0.02)	(0.01)	(0.01)
#keywords	1.080***	0.949***	1.082***	0.983	1.051***	0.958***	1.089***	0.944***
	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
Constant	2.112***	2.162***	1.239***	2.458***	1.701***	2.116***	2.667***	2.148***
	(0.02)	(0.03)	(0.05)	(0.21)	(0.03)	(0.06)	(0.04)	(0.05)
Subdomain dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	449710		31323		185211		233176	
Log Likeli.	-899423.2		-39607.9		-337750.0		-520008.0	

Exponentiated coefficients
Sample: Papers published in 2001
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: Number of forward citations within 3 years.
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.

Table 125 – Predicting Forward Citations (generalized negative binomial models).

	Full sample		Hum & SS		Hard Sc & Eng		Life Sc	
	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha
Pairwise keyword novelty	1.481***	0.918***	1.381***	0.838***	1.524***	0.918***	1.432***	0.935***
	(0.01)	(0.01)	(0.04)	(0.03)	(0.02)	(0.02)	(0.01)	(0.01)
#keywords	1.085***	0.957***	1.073***	0.989	1.062***	0.966***	1.092***	0.951***
	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
Constant	4.459***	1.865***	3.364***	2.158***	3.534***	1.928***	5.706***	1.850***
	(0.04)	(0.02)	(0.13)	(0.13)	(0.06)	(0.04)	(0.08)	(0.04)
Subdomain dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	449710		31323		185211		233176	
Log Likeli.	-1190982.9		-56797.4		-456322.2		-675446.7	

Exponentiated coefficients
Sample: Papers published in 2001
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: Number of forward citations within 5 years.
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.

Table 126 – Predicting "Big Hit" Probabilities (logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top10_cl_kk_c_f3				
Pairwise keyword novelty	1.555***	1.513***	1.727***	1.470***
	(0.03)	(0.10)	(0.05)	(0.03)
#keywords	1.123***	1.145***	1.108***	1.138***
	(0.00)	(0.02)	(0.01)	(0.00)
Constant	0.040***	0.041***	0.050***	0.033***
	(0.00)	(0.00)	(0.00)	(0.00)
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	450148	31332	185543	233273
Log Likeli.	-106914.5	-5634.0	-48172.1	-52962.6

Exponentiated coefficients
Sample: Papers published in 2001
Significance levels: * p<0.05 ** p<0.01 *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 10%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.

Table 127 – Predicting "Big Hit" Probabilities (logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top10_cl_kk_c_f5				
Pairwise keyword novelty	1.570***	1.554***	1.712***	1.504***
	(0.03)	(0.10)	(0.05)	(0.03)
#keywords	1.129***	1.158***	1.117***	1.141***
	(0.00)	(0.02)	(0.01)	(0.00)
Constant	0.042***	0.047***	0.052***	0.034***
	(0.00)	(0.00)	(0.00)	(0.00)
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	450148	31332	185543	233273
Log Likeli.	-111750.2	-6327.5	-50718.2	-54530.4

Exponentiated coefficients
Sample: Papers published in 2001
Significance levels: * p<0.05 ** p<0.01 *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 10%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.

Individual decision making

Table 128 – Predicting "Big Hit" Probabilities (logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top5_cl_kk_c_f3				
Pairwise keyword novelty	1.719***	1.569***	1.887***	1.649***
	(0.04)	(0.14)	(0.07)	(0.05)
#keywords	1.131***	1.158***	1.119***	1.141***
	(0.00)	(0.02)	(0.01)	(0.01)
Constant	0.017***	0.019***	0.019***	0.014***
	(0.00)	(0.00)	(0.00)	(0.00)
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	450148	31332	185543	233273
Log Likeli.	-59515.7	-3381.4	-27299.3	-28756.9

Exponentiated coefficients
Sample: Papers published in 2001
Significance levels: * p<0.05 ** p<0.01 *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 5%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.

Table 129 – Predicting "Big Hit" Probabilities (logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top5_cl_kk_c_f5				
Pairwise keyword novelty	1.673***	1.586***	1.762***	1.660***
	(0.04)	(0.14)	(0.07)	(0.05)
#keywords	1.135***	1.151***	1.133***	1.140***
	(0.00)	(0.02)	(0.01)	(0.01)
Constant	0.018***	0.022***	0.023***	0.014***
	(0.00)	(0.00)	(0.00)	(0.00)
Subdomain dummies (10)	Yes	Yes	Yes	Yes
N	450148	31332	185543	233273
Log Likeli.	-63482.6	-3679.6	-29623.9	-30075.8

Exponentiated coefficients
Sample: Papers published in 2001
Significance levels: * p<0.05 ** p<0.01 *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 5%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.

Table 130 – Predicting Forward Citations With More Controls (generalized negative binomial models).

	Full sample		Hum & SS		Hard Sc & Eng		Life Sc	
	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha	Citations (3y)	Lnalpha
Pairwise keyword novelty	1.404***	0.873***	1.352***	0.902*	1.288***	0.872***	1.386***	0.898***
	(0.01)	(0.01)	(0.04)	(0.04)	(0.02)	(0.02)	(0.01)	(0.01)
Keyword novelty	0.980***	0.963***	1.101***	0.948**	0.882***	0.979***	1.017***	0.971***
	(0.00)	(0.00)	(0.01)	(0.02)	(0.00)	(0.01)	(0.00)	(0.01)
#keywords	1.074***	0.962***	1.075***	0.983	1.065***	0.978***	1.076***	0.954***
	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
#co-authors	1.111***	0.941***	1.148***	0.941**	1.098***	0.931***	1.098***	0.948***
	(0.00)	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)	(0.00)	(0.01)
#institutions	1.018***	1.044***	0.991	1.040	1.018	1.063***	1.026***	1.032***
	(0.00)	(0.01)	(0.01)	(0.03)	(0.01)	(0.02)	(0.01)	(0.01)
Constant	1.643***	3.046***	0.366***	3.851***	3.601***	2.581***	1.425***	2.779***
	(0.04)	(0.12)	(0.04)	(0.69)	(0.15)	(0.16)	(0.05)	(0.15)
Subdomain dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	448414		30988		184582		232844	
Log Likeli.	-886591.8		-38728.3		-332794.8		-512499.2	

Exponentiated coefficients
Sample: Papers published in 2001
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: Number of forward citations within 3 years.
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.

Table 131 – Predicting Forward Citations With More Controls (generalized negative binomial models).

	Full sample		Hum & SS		Hard Sc & Eng		Life Sc	
	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha	Citations (5y)	Lnalpha
Pairwise keyword novelty	1.381***	0.897***	1.379***	0.882**	1.241***	0.876***	1.380***	0.933***
	(0.01)	(0.01)	(0.04)	(0.03)	(0.02)	(0.02)	(0.01)	(0.01)
Keyword novelty	0.986***	0.966***	1.099***	0.942***	0.906***	0.983***	1.011***	0.974***
	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
#keywords	1.078***	0.968***	1.067***	0.989	1.073***	0.980***	1.079***	0.960***
	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
#co-authors	1.111***	0.940***	1.139***	0.942***	1.096***	0.928***	1.099***	0.948***
	(0.00)	(0.00)	(0.01)	(0.02)	(0.01)	(0.01)	(0.00)	(0.01)
#institutions	1.011*	1.038***	0.993	1.023	1.008	1.055***	1.022***	1.028***
	(0.00)	(0.01)	(0.01)	(0.02)	(0.01)	(0.02)	(0.01)	(0.01)
Constant	3.348***	2.614***	1.007	3.533***	6.031***	2.288***	3.298***	2.381***
	(0.08)	(0.09)	(0.09)	(0.48)	(0.24)	(0.12)	(0.11)	(0.11)
Subdomain dummies (10)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	448414		30988		184582		232844	
Log Likeli.	-1175955.0		-55561.8		-450820.8		-666848.4	

Exponentiated coefficients
Sample: Papers published in 2001
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: Number of forward citations within 5 years.
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.

Table 132 – Predicting "Big Hit" Probabilities With More Controls (logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top10_cl_kk_c_f3				
Pairwise keyword novelty	1.285***	1.371***	1.342***	1.257***
	(0.02)	(0.09)	(0.04)	(0.03)
Keyword novelty	0.864***	0.936*	0.886***	0.834***
	(0.01)	(0.03)	(0.01)	(0.01)
#keywords	1.119***	1.144***	1.109***	1.126***
	(0.00)	(0.02)	(0.01)	(0.01)
#co-authors	1.132***	1.191***	1.118***	1.137***
	(0.01)	(0.03)	(0.02)	(0.01)
#institutions	1.037***	1.011	1.026	1.044***
	(0.01)	(0.04)	(0.02)	(0.01)
Constant	0.086***	0.040***	0.095***	0.089***
	(0.01)	(0.01)	(0.01)	(0.01)
Subdomain dummies (10)	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes
N	448414	30988	184582	232844
Log Likeli.	-103401.4	-5476.2	-46862.3	-50854.7

Exponentiated coefficients
Sample: Papers published in 2001
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 10%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.

Correlation Table

Table 133 – Predicting "Big Hit" Probabilities With More Controls (logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top10_cl_kk_c_f5				
Pairwise keyword novelty	1.300***	1.390***	1.350***	1.280***
	(0.02)	(0.09)	(0.04)	(0.03)
Keyword novelty	0.864***	0.913***	0.894***	0.827***
	(0.00)	(0.02)	(0.01)	(0.01)
#keywords	1.124***	1.159***	1.116***	1.130***
	(0.00)	(0.02)	(0.01)	(0.00)
#co-authors	1.134***	1.178***	1.116***	1.143***
	(0.01)	(0.03)	(0.02)	(0.01)
#institutions	1.033**	1.024	1.018	1.041***
	(0.01)	(0.03)	(0.02)	(0.01)
Constant	0.091***	0.054***	0.090***	0.104***
	(0.01)	(0.01)	(0.01)	(0.01)
Subdomain dummies (10)	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes
N	448414	30988	184582	232844
Log Likeli.	-107932.6	-6134.4	-49251.0	-52298.6

Exponentiated coefficients
Sample: Papers published in 2001
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 10%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.

Table 134 – Predicting "Big Hit" Probabilities With More Controls (logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top10_cl_kk_c_f3				
Pairwise keyword novelty	1.285***	1.371***	1.342***	1.257***
	(0.02)	(0.09)	(0.04)	(0.03)
Keyword novelty	0.864***	0.936*	0.886***	0.834***
	(0.01)	(0.03)	(0.01)	(0.01)
#keywords	1.119***	1.144***	1.109***	1.126***
	(0.00)	(0.02)	(0.01)	(0.01)
#co-authors	1.132***	1.191***	1.118***	1.137***
	(0.01)	(0.03)	(0.02)	(0.01)
#institutions	1.037***	1.011	1.026	1.044***
	(0.01)	(0.04)	(0.02)	(0.01)
Constant	0.086***	0.040***	0.095***	0.089***
	(0.01)	(0.01)	(0.01)	(0.01)
Subdomain dummies (10)	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes
N	448414	30988	184582	232844
Log Likeli.	-103401.4	-5476.2	-46862.3	-50854.7

Exponentiated coefficients
Sample: Papers published in 2001
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 10%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.

Table 135 – Predicting "Big Hit" Probabilities With More Controls (logit models).

	Full sample	Hum & SS	Hard Sc & Eng	Life Sc
top10_cl_kk_c_f5				
Pairwise keyword novelty	1.300***	1.390***	1.350***	1.280***
	(0.02)	(0.09)	(0.04)	(0.03)
Keyword novelty	0.864***	0.913***	0.894***	0.827***
	(0.00)	(0.02)	(0.01)	(0.01)
#keywords	1.124***	1.159***	1.116***	1.130***
	(0.00)	(0.02)	(0.01)	(0.00)
#co-authors	1.134***	1.178***	1.116***	1.143***
	(0.01)	(0.03)	(0.02)	(0.01)
#institutions	1.033**	1.024	1.018	1.041***
	(0.01)	(0.03)	(0.02)	(0.01)
Constant	0.091***	0.054***	0.090***	0.104***
	(0.01)	(0.01)	(0.01)	(0.01)
Subdomain dummies (10)	Yes	Yes	Yes	Yes
Geographical dummies (5)	Yes	Yes	Yes	Yes
N	448414	30988	184582	232844
Log Likeli.	-107932.6	-6134.4	-49251.0	-52298.6

Exponentiated coefficients
Sample: Papers published in 2001
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.
Robust standard errors are reported in parentheses.
Dependent variable: dummy indicating if the paper is a big hit (top 10%).
Pairwise keyword novelty is a dummy equal to one if the paper is in the top 10% most novel papers.

Table 136 – Correlations between Pairwise Keyword Novelty and Keyword Novelty

		1	**2**	**3**	**4**
1	Pairwise keyword novelty (Author KWs)	1			
2	Keyword novelty (Author KWs)	-0.23*	1		
3	Pairwise keyword novelty (ISI KWs)	0.22*	0.01*	1	
4	Keyword novelty (ISI KWs)	0.01*	0.41*	-0.04*	1

Correlation Table (Full)

Table 137 – Descriptive statistics and correlations

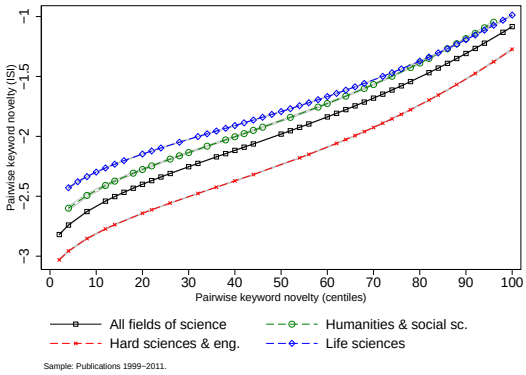
	Mean	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1 Number of citations (3y)	4.49	6.99	1																
2 Big hit papers (top-10%)	0.10	0.30	0.59*	1															
3 Big hit papers (top-5%)	0.05	0.21	0.55*	0.67*	1														
4 Pairwise keyword novelty (AK)	-3.77	2.58	0.16*	0.07*	0.05*	1													
5 Pairwise keyword novelty (KW+)	-2.36	2.25	0.10*	0.03*	0.02*	0.22*	1												
6 Pairwise keyword novelty (top-10%)(AK)	0.17	0.38	0.11*	0.04*	0.03*	0.49*	0.13*	1											
7 Pairwise keyword novelty (top-10%)(KW+)	0.07	0.26	-0.16*	-0.08*	-0.05*	-0.10*	0.35*	-0.04*	1										
8 Keyword novelty (AK)	10.16	1.19	-0.02*	-0.04*	-0.03*	-0.23*	0.01*	-0.17*	0.02*	1									
9 Keyword novelty (KW+)	10.76	1.24	0.11*	-0.02*	-0.02*	0.01*	-0.04*	0.02*	-0.07*	0.41*	1								
10 Commonness (AK)	0.46	0.50	0.13*	0.06*	0.04*	0.49*	0.13*	0.38*	-0.05*	-0.10*	0.05*	1							
11 Commonness (KW+)	0.32	0.47	0.21*	0.08*	0.06*	0.43*	0.45*	0.34*	-0.21*	-0.04*	0.03*	0.68*	1						
12 Journal references novelty	-1.19	2.29	0.18*	0.03*	0.02*	0.12*	0.16*	0.09*	-0.20*	0.12*	0.30*	0.07*	0.17*	1					
13 Journal references novelty (top-10%)	0.11	0.31	-0.12*	-0.07*	-0.05*	-0.10*	-0.08*	-0.06*	0.26*	-0.06*	-0.05*	-0.08*	-0.18*	0.44*	1				
14 Journal impact factor (3y)	4.17	3.47	0.53*	0.24*	0.18*	0.23*	0.18*	0.17*	-0.25*	-0.01*	0.24*	0.16*	0.29*	0.31*	-0.17*	1			
15 Number of keywords	4.64	1.61	0.10*	0.06*	0.05*	0.37*	0.03*	0.09*	-0.11*	0.11*	0.08*	0.10*	0.10*	0.10*	-0.06*	0.12*	1		
16 Number of authors	2.32	1.80	0.20*	0.09*	0.07*	0.11*	0.07*	0.08*	-0.13*	-0.02*	0.08*	0.07*	0.13*	0.16*	-0.09*	0.25*	0.05*	1	
17 Number of institutions	1.94	1.52	0.17*	0.09*	0.07*	0.08*	0.05*	0.06*	-0.10*	-0.03*	0.05*	0.06*	0.10*	0.09*	-0.08*	0.19*	0.04*	0.90*	1

* denote significance at the 5% level.

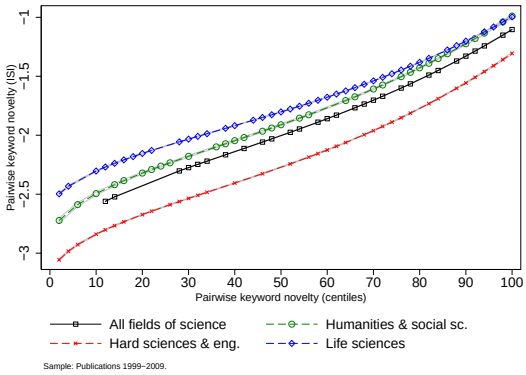
Correlation between Pairwise Author Keywords Novelty and Pairwise KeyWords Plus Novelty

Figure 20 – Pairwise Keyword Novelty (Author KW) and Pairwise Keyword Novelty (KW+)

(a) Sample: 1999-2011



(b) Sample: 1999-2009



Cahiers du GREThA
Working papers of GREThA

GREThA UMR CNRS 5113

Université de Bordeaux

Avenue Léon Duguit
33608 PESSAC - FRANCE
Tel : +33 (0)5.56.84.25.75
Fax : +33 (0)5.56.84.86.47

http://gretha.u-bordeaux.fr/

Cahiers du GREThA (derniers numéros – last issues)

2018-20: BRESCHI Stefano, LISSONI Francesco, MIGUELEZ Ernest: Return migrants' self-selection: Evidence for Indian inventors

2018-21: GRAVEL Nicolas, MAGDALOU Brice, MOYES Patrick: Inequality Measurement with an Ordinal and Continuous Variable

2018-22: FUENTES ESPINOZA Alejandro, HUBERT Anne, RAINEAU Yann, FRANC Céline, GIRAUD-HERAUD Eric: Variétés résistantes et acceptabilité par le marché : une évaluation par l'économie expérimentale

2018-23 : BONIN Hubert : La Société générale en 1890-1914 : d'une forte croissance à la crise de son modèle économique ?

2018-24 : BOKINO Régis, GANO Moustapha : Degré d'indépendance et responsabilisation au sein du comité de politique monétaire de la BCEAO

2018-25 : KLEBANER Samuel : Production normative et dynamique institutionnelle : comment le programme de recherche de la Théorie de la Régulation peut se nourrir des concepts de «l'école d'Histoire du Droit de Francfort»

2019-01 : BALLET Jérôme : Les jugements évaluatifs entre positif et normatif : Pour une économie axiologique

2019-02 : BONIN Hubert : L'année 1819, symbole du réveil économique de la France après la défaite de l'Empire

2019-03 : USECHE Diego, MIGUELEZ Ernest, LISSONI : Highly skilled and well connected: Migrant inventors in Cross-Border M&As

2019-04: BROUILLAT Eric, SAINT JEAN Maïder: Dura lex sed lex: Why implementation gaps in environmental policy matter?

La coordination scientifique des Cahiers du GREThA est assurée par Valerio STERZI.
La mise en page et la diffusion sont assurées par Julie VISSAGUET