Novelty and academic impact

N. Carayol[1] A. Lahatte[2] O. Llopis[3]

[1]GREThA (UMR CNRS 5113)
University of Bordeaux

[2]Observatoire des Sciences et Techniques
HCERES, Paris

[3]Rennes School of Business

STI conference, Paris 2017

Motivation of the paper

- The old times of science: the "renaissance man" (Jones, 2009), the "gentlemanly specialist" (Rudwick, 1985), the English "amateur scientist" (Shapin, 2008) or the French "savant".

Motivation of the paper

- The old times of science: the "renaissance man" (Jones, 2009), the "gentlemanly specialist" (Rudwick, 1985), the English "amateur scientist" (Shapin, 2008) or the French "savant".

- Modern times: communities of professional scientists strongly subsidized by the states and organized through formal peer reviewed vetting procedures for recruitment, funding and publishing. A time of "big science" (Price, 1963) which outcome doubles every ten-to-twenty years (Price, 1961; Olesen Larsen and von Ins, 2010), with increasing team size (Jones, Wuchty and Uzzi, 2008) and raising specialization and knowledge complexity (Jones, 2009).

Motivation of the paper

- The old times of science: the "renaissance man" (Jones, 2009), the "gentlemanly specialist" (Rudwick, 1985), the English "amateur scientist" (Shapin, 2008) or the French "savant".

- Modern times: communities of professional scientists strongly subsidized by the states and organized through formal peer reviewed vetting procedures for recruitment, funding and publishing. A time of "big science" (Price, 1963) which outcome doubles every ten-to-twenty years (Price, 1961; Olesen Larsen and von Ins, 2010), with increasing team size (Jones, Wuchty and Uzzi, 2008) and raising specialization and knowledge complexity (Jones, 2009).

- In this context, **does the science system maintain its standards of creativity and innovation?**

Motivation of the paper

- The old times of science: the "renaissance man" (Jones, 2009), the "gentlemanly specialist" (Rudwick, 1985), the English "amateur scientist" (Shapin, 2008) or the French "savant".

- Modern times: communities of professional scientists strongly subsidized by the states and organized through formal peer reviewed vetting procedures for recruitment, funding and publishing. A time of "big science" (Price, 1963) which outcome doubles every ten-to-twenty years (Price, 1961; Olesen Larsen and von Ins, 2010), with increasing team size (Jones, Wuchty and Uzzi, 2008) and raising specialization and knowledge complexity (Jones, 2009).

- In this context, **does the science system maintain its standards of creativity and innovation?**

- Most empirical evidence of a negative bias against groundbreaking and innovative research comes from peer review (Braben, 2004; Chubin and Hackett, 1990; Wesseley, 1998; Heinze et al., 2009); Alberts,

But, what is "novelty"?

- H Poincare first introduced the idea that invention in mathematics proceeds from recombinations of distinct pre-existing ideas ("mathematical entities") in one's mind ((Poincaré, 1910)).

But, what is "novelty"?

- H Poincare first introduced the idea that invention in mathematics proceeds from recombinations of distinct pre-existing ideas ("mathematical entities") in one's mind ((Poincaré, 1910)).

- Economists of innovation: Innovation = original recombination of existing elements (Nelson and Winter 1982, Schumpeter 1942). Example: Edison invention of the "electric candle" as the Cross-pollination of two distant ideas: "candle" + "electricity", and the test of more than 6.000 different materials to find the filament for the bulb.

But, what is "novelty"?

- H Poincare first introduced the idea that invention in mathematics proceeds from recombinations of distinct pre-existing ideas ("mathematical entities") in one's mind ((Poincaré, 1910)).
- Economists of innovation: Innovation = original recombination of existing elements (Nelson and Winter 1982, Schumpeter 1942). Example: Edison invention of the "electric candle" as the Cross-pollination of two distant ideas: "candle" + "electricity", and the test of more than 6.000 different materials to find the filament for the bulb.
- Weitzman (1998) proposes a mechanism for the growth of ideas in the economy that result from binary random combinations of existing ideas (pairwise cross-pollination).

Measuring novelty in science: pairwise journal citation frequencies

- Uzzi et al. (2013) build on this idea to study novelty in science. They employ pairwise journal co-citations in articles' reference lists (Small, 1973) to identify recombinations of previous knowledge. The "atypicality" or the conventionality of those re-combinations are computed through their frequencies of occurrence over the whole period.

Measuring novelty in science: pairwise journal citation frequencies

- Uzzi et al. (2013) build on this idea to study novelty in science. They employ pairwise journal co-citations in articles' reference lists (Small, 1973) to identify recombinations of previous knowledge. The "atypicality" or the conventionality of those re-combinations are computed through their frequencies of occurrence over the whole period.

- Lee et al, 2015 proposed a measure of novelty of journal references which is time-varying and more simple.

Measuring novelty in science: pairwise journal citation frequencies

- Uzzi et al. (2013) build on this idea to study novelty in science. They employ pairwise journal co-citations in articles' reference lists (Small, 1973) to identify recombinations of previous knowledge. The "atypicality" or the conventionality of those re-combinations are computed through their frequencies of occurrence over the whole period.

- Lee et al, 2015 proposed a measure of novelty of journal references which is time-varying and more simple.

- Wang, Veugelers and Stephan (2017) use the sum of (completely) new pairwise reference combinations weighted negatively by the cosine similarity of the two journals.

The idea: Novelty based on the originality of the keyword combinations

- We propose to capture the "different angle" of a research article by looking at the use frequencies of the keyword combinations in scientific articles.

The idea: Novelty based on the originality of the keyword combinations

- We propose to capture the "different angle" of a research article by looking at the use frequencies of the keyword combinations in scientific articles.
- The originality of keyword combinations are intended to capture the "thematic novelty" of the scientific papers (their propensity to raise questions that are new to their corresponding scientific field).

The idea: Novelty based on the originality of the keyword combinations

- We propose to capture the "different angle" of a research article by looking at the use frequencies of the keyword combinations in scientific articles.
- The originality of keyword combinations are intended to capture the "thematic novelty" of the scientific papers (their propensity to raise questions that are new to their corresponding scientific field).
- Novelty /disruption results more from intentional specific investigations rather than from random combinations of already existing pieces of knowledge.

Main questions

- How could (scientific) novelty be observed/defined?

Main questions

- How could (scientific) novelty be observed/defined?
- Is novelty a good leverage for excellence? Is it risky?

Main questions

- How could (scientific) novelty be observed/defined?
- Is novelty a good leverage for excellence? Is it risky?
- Does it pay to be novel in science? Provided you get results, and get published, do you have higher impact?

The data

- Our dataset includes all research articles published from 1999 to 2013 and indexed in Thomson Reuters Web of Science (WOS): 10 million articles (7.8 million before 2011) having more than two keywords.

The data

- Our dataset includes all research articles published from 1999 to 2013 and indexed in Thomson Reuters Web of Science (WOS): 10 million articles (7.8 million before 2011) having more than two keywords.
- We collected all citations to these papers: 26.1 million citations (3y) & 43.3 million citations (5y).

The data

- Our dataset includes all research articles published from 1999 to 2013 and indexed in Thomson Reuters Web of Science (WOS): 10 million articles (7.8 million before 2011) having more than two keywords.
- We collected all citations to these papers: 26.1 million citations (3y) & 43.3 million citations (5y).
- These papers are classified in three major research areas: humanities and social sciences (7.18%), life sciences (46.68%) and hard sciences and engineering (46.14%).

The data

- Our dataset includes all research articles published from 1999 to 2013 and indexed in Thomson Reuters Web of Science (WOS): 10 million articles (7.8 million before 2011) having more than two keywords.
- We collected all citations to these papers: 26.1 million citations (3y) & 43.3 million citations (5y).
- These papers are classified in three major research areas: humanities and social sciences (7.18%), life sciences (46.68%) and hard sciences and engineering (46.14%).
- A number of associated data for each paper which will be used to build our list of independent and control variables.

The indicator

Novelty: atypicality of a papers' keywords combinations in year t and scientific domain c.

- Step 1: retrieve all papers in WOS (1999-2013)

The indicator

Novelty: atypicality of a papers' keywords combinations in year t and scientific domain c.

- Step 1: retrieve all papers in WOS (1999-2013)
- Step 2: extract and clean all keywords from all papers, deleting irrelevant keywords

The indicator

Novelty: atypicality of a papers' keywords combinations in year t and scientific domain c.

- Step 1: retrieve all papers in WOS (1999-2013)
- Step 2: extract and clean all keywords from all papers, deleting irrelevant keywords
- Step 3: attribute an indicator of commonness for each pair of keywords, year and subject category.

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

with N_{ct} the number of (non-distinct) keyword combinations in papers published in c and year t. The terms N_{ict}, N_{jct} and N_{ijct} give the number of such (non-distinct) keyword combinations in which respectively keyword i, keyword j, and both keywords i and j appear.

The indicator (II)

- Step 2: attribute a unique *novelty of keyword combinations* indicator for each paper in our sample (most articles have more than 2 keywords).

$$com_c = 10thPercentile\left(Com_{ijct} \,|\forall ij \in K\right) \tag{2}$$

The indicator (II)

- Step 2: attribute a unique *novelty of keyword combinations* indicator for each paper in our sample (most articles have more than 2 keywords).

$$com_c = 10thPercentile\left(Com_{ijct} \,|\forall ij \in K\right) \tag{2}$$

- Step 3: From commonness to novelty: inverse logarithmic transformation of commonness to have the novelty of a given paper in a given subject category c.

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

The indicator (II)

- Step 2: attribute a unique *novelty of keyword combinations* indicator for each paper in our sample (most articles have more than 2 keywords).

$$com_c = 10thPercentile\left(Com_{ijct} \,|\forall ij \in K\right) \tag{2}$$

- Step 3: From commonness to novelty: inverse logarithmic transformation of commonness to have the novelty of a given paper in a given subject category c.

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

- Step 4: Take the max novelty over the subject categories c.

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

The indicator (III) - variations

- Consider 2/3 years.

The indicator (III) - variations

- Consider 2/3 years.
- At the paper level: take the max instead of the 10th percentile.

The indicator (III) - variations

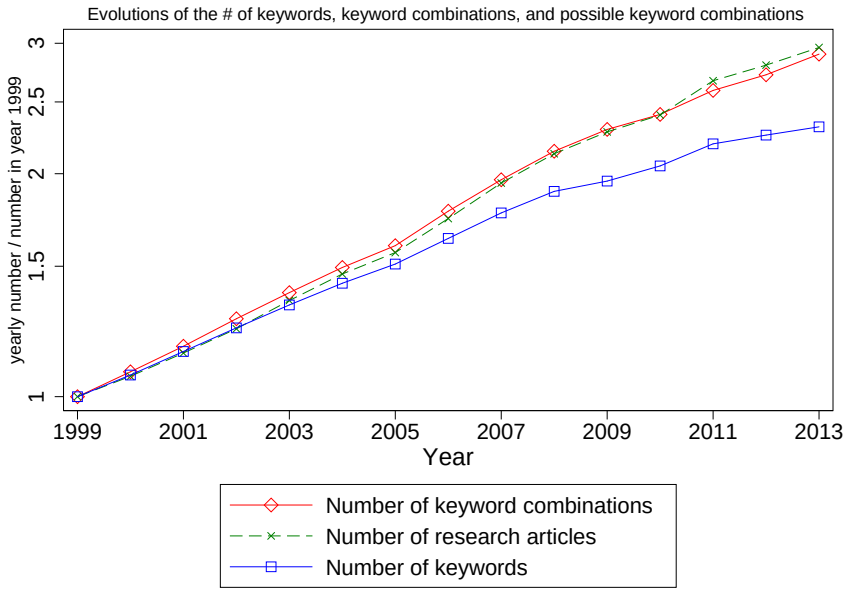
- Consider 2/3 years.
- At the paper level: take the max instead of the 10th percentile.
- Not field (subject category) specific.

The indicator (III) - variations

- Consider 2/3 years.
- At the paper level: take the max instead of the 10th percentile.
- Not field (subject category) specific.
- Using ISI keywords.

The evolution of scientific knowledge

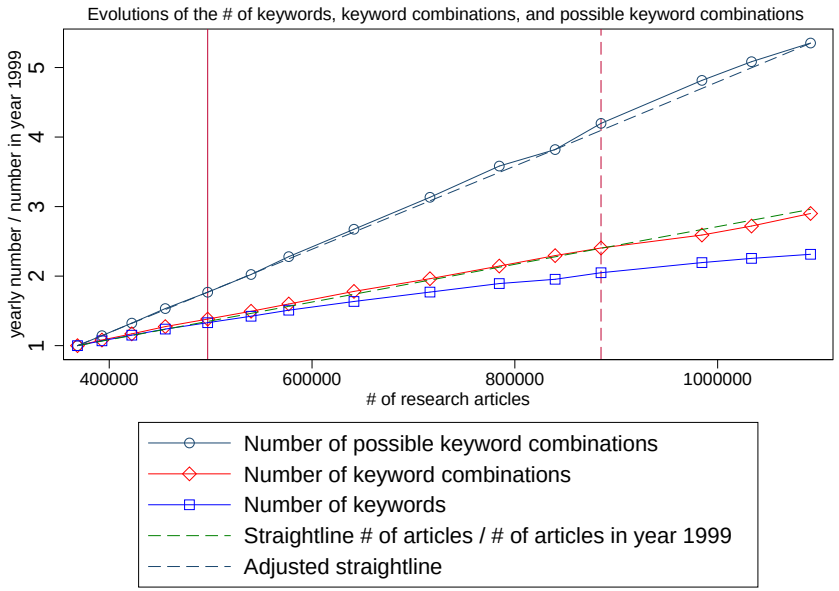
Figure: The evolution of the number of distinct keyword combinations, number of distinct keywords, and number of research articles



- Evolution of the number of distinct keyword combinations follows a very similar growth pattern as the number of research articles (about 290% growth from 1999-2012).

The evolution of scientific knowledge

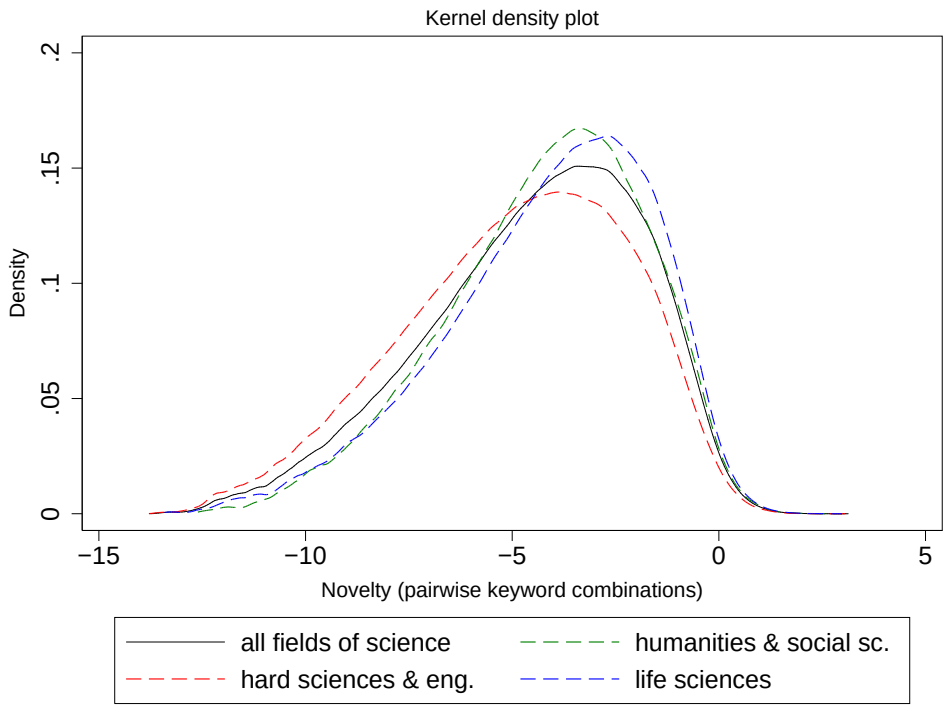
Figure: The evolution of the number of possible keyword combinations, the "explored" keyword combinations and keywords with respect to the number of research articles.



- Evolution of the number of distinct keyword combinations follows a very similar growth pattern as the number of research articles (about 290% growth from 1999-2012).

The distribution

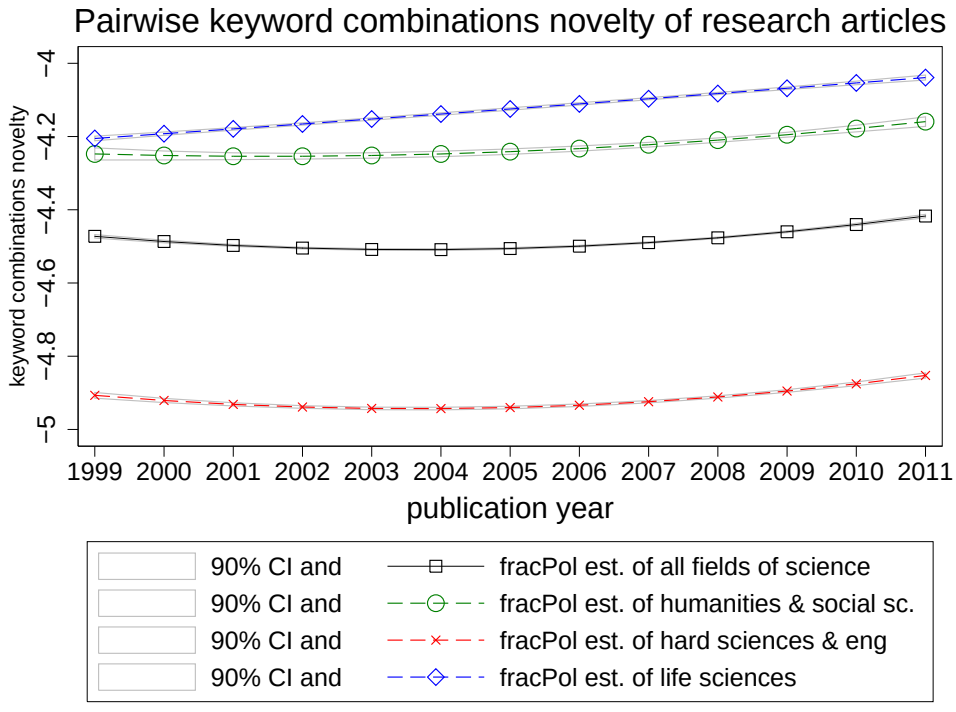
Figure: The distribution of keyword combinations novelty (3-year). For all articles and for the three domains of science



- Most articles have intermediate levels of novelty.
- Similar across all three fields of science.

Novelty in science: Time evolution

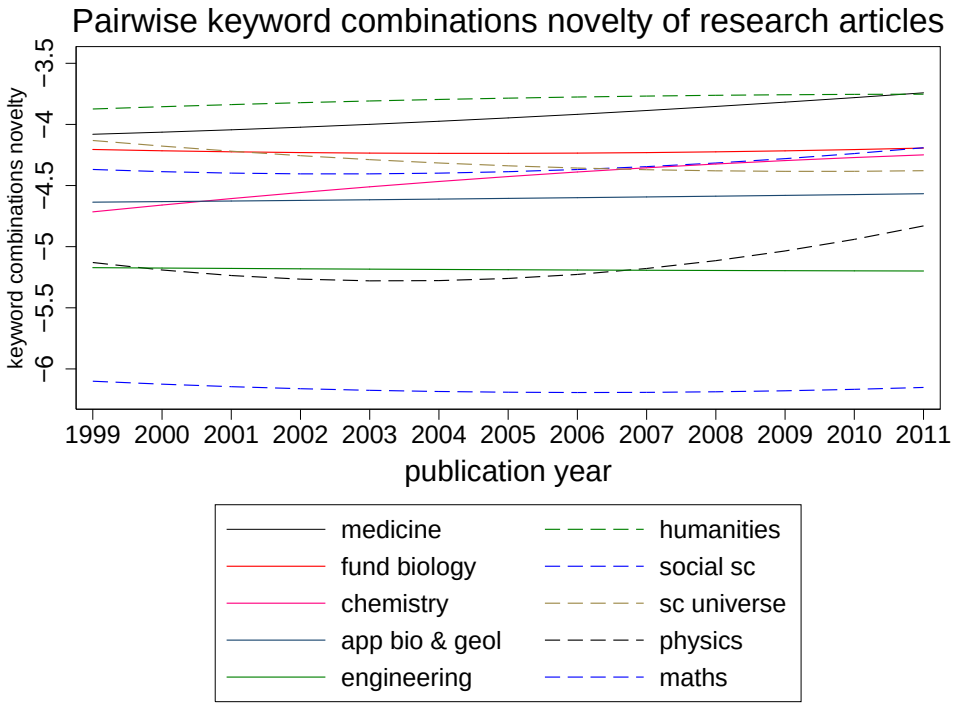
Figure: Evolution of keyword combinations novelty for the three large fields of science



- Quite stable over the period. Highest for life sciences.

Novelty in science: Time evolution

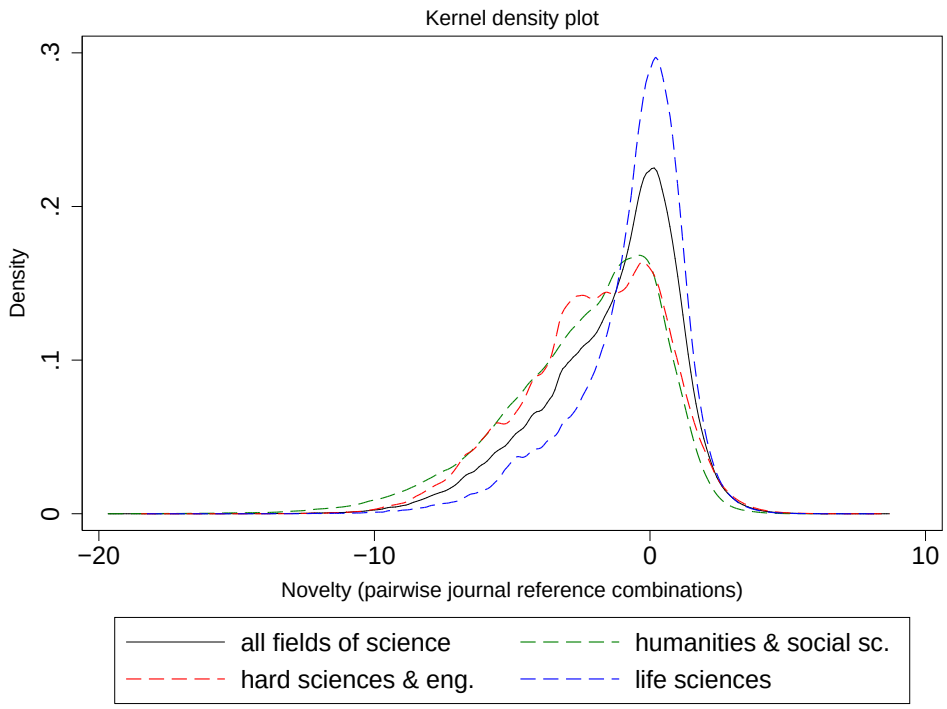
Figure: Evolution of keyword combinations novelty for the three large fields of science



- Quite stable over the period. Highest for life sciences.

A pairwise journal reference benchmark: the distribution

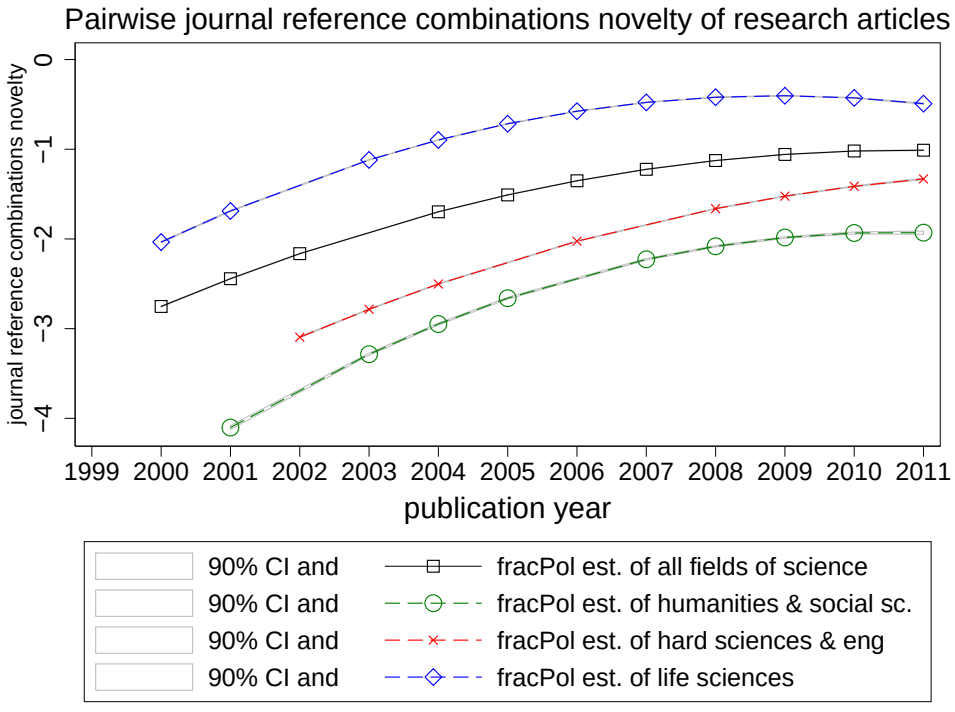
Figure: The distribution of journal reference combinations novelty (3-year). For all articles and for the three domains of science



- Most articles have intermediate levels of novelty.
- Similar across all three fields of science.

A pairwise journal reference benchmark: Time evolution

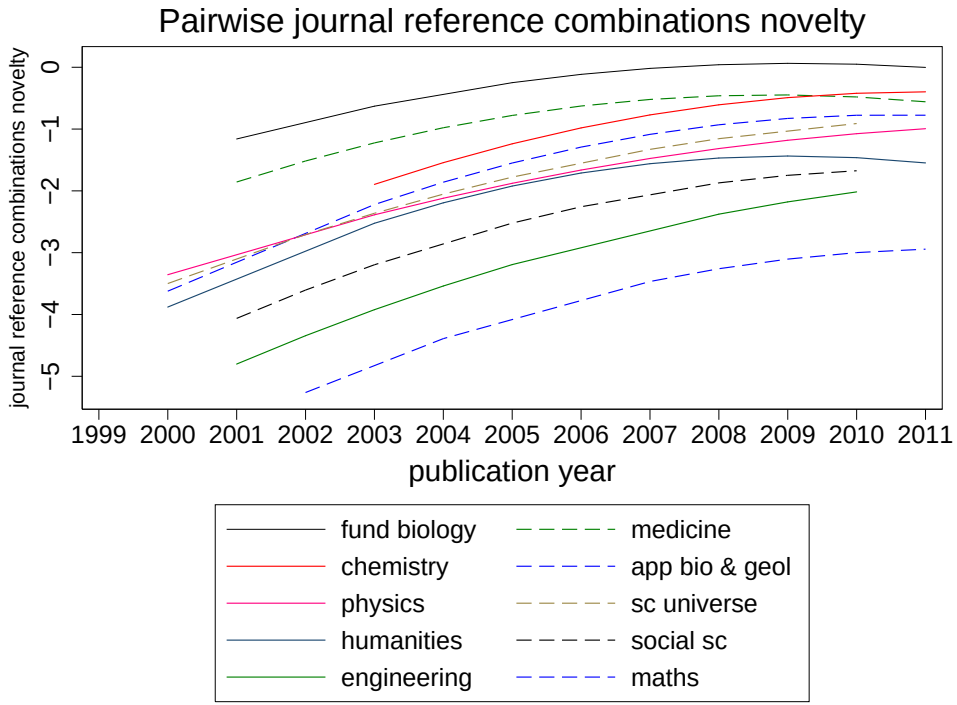
Figure: Evolution of keyword combinations novelty for the three large fields of science



- Quite stable over the period. Highest for life sciences.

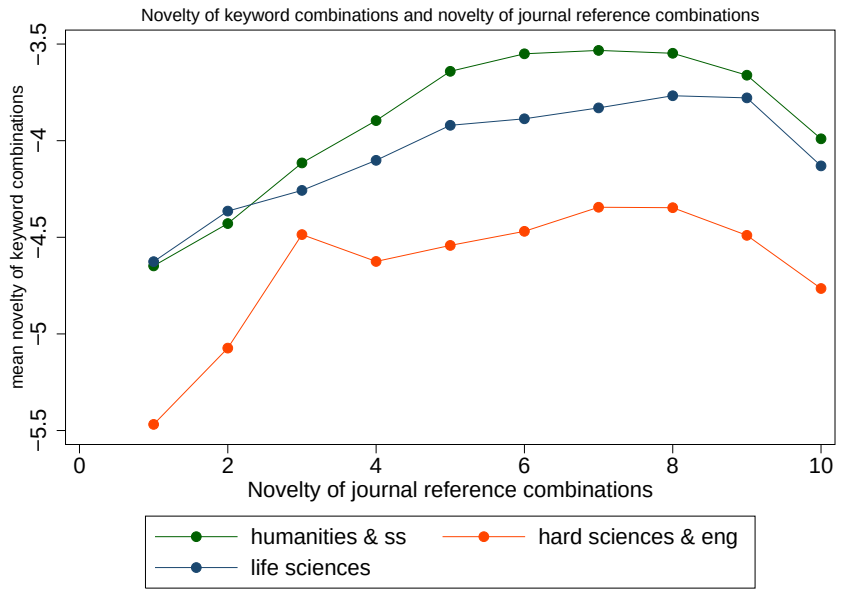
A pairwise journal reference benchmark: Time evolution

Figure: Evolution of journal reference combinations novelty for the three large fields of science



- Quite stable over the period. Highest for life sciences.

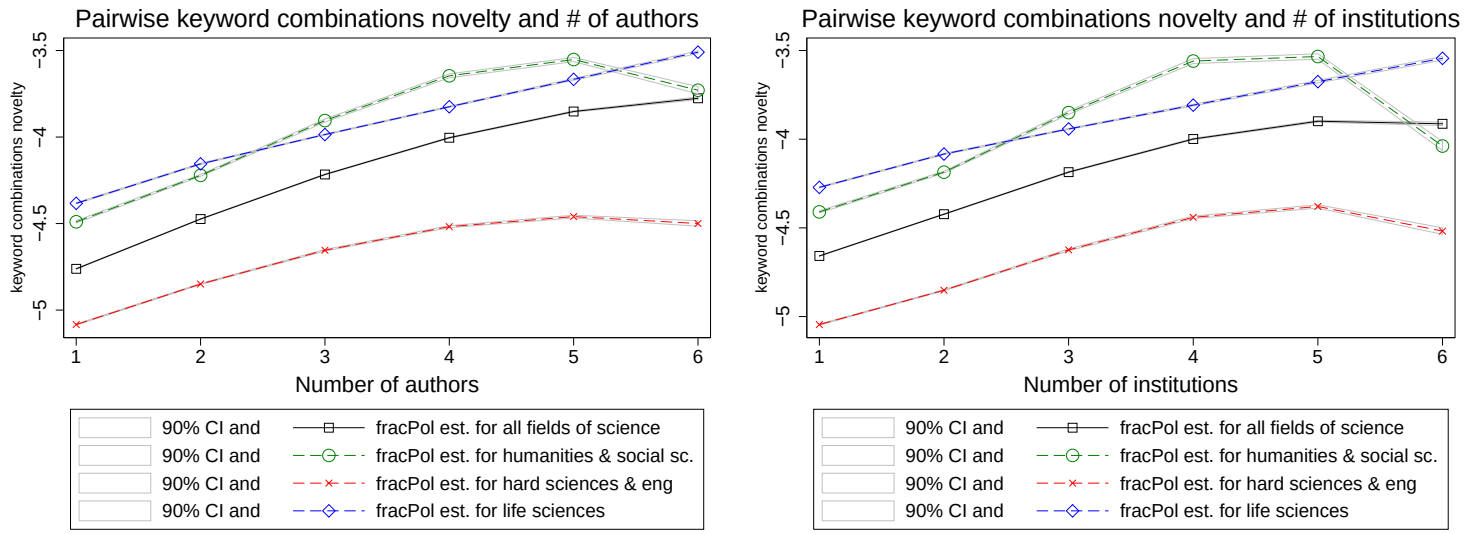
Correlation between keyword combinations novelty and pairwise journal reference combination novelty



- Quite stable over the period. Highest for life sciences.

Which teams produce more novel papers?

Figure: Novelty of keyword combinations and number of authors by fields (3-year window).



- Novelty increases with the number of authors across all fields of science.
- Novelty is higher with US authors, lower for Asian authors.

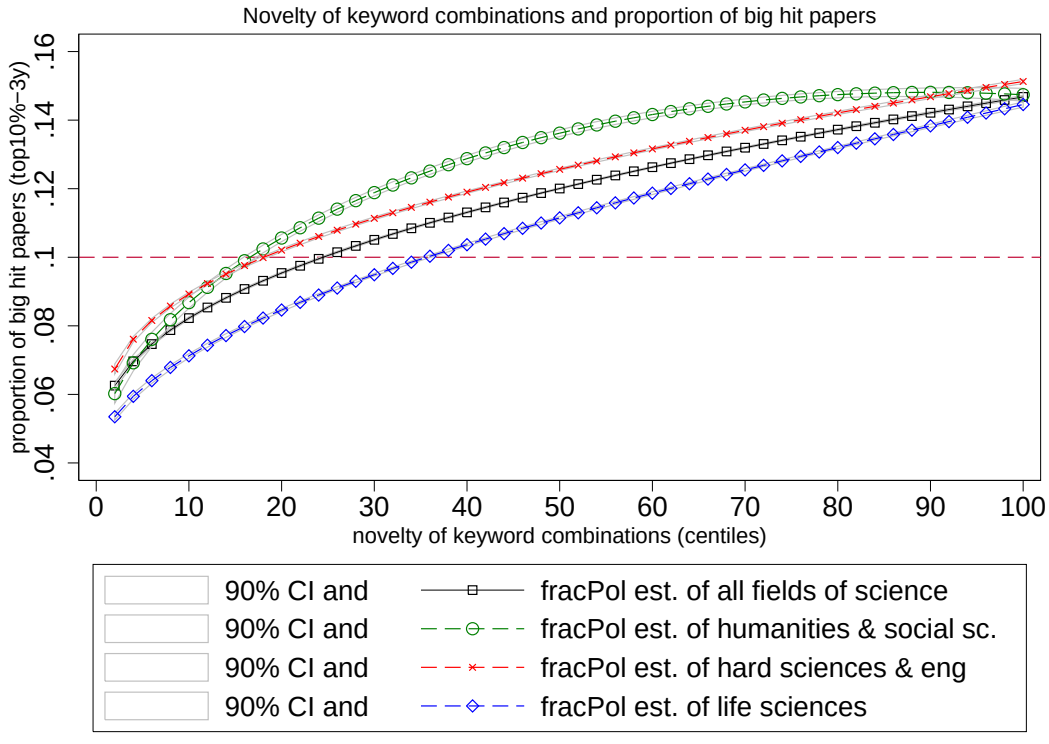
Is novelty "popular" in science?

Do highly novel papers attract outstanding attention from peers?

- Dependent variable: *"big hit paper"*
- Dummy taking the value 1 if the paper belongs to the top-10% most cited in its scientific field and publication year.
- Robustness checks with top5% and top1% most cited.
- Logistic regressions

Empirical evidence (I)

Figure: Novelty and big hits



- Papers ranked according to their novelty score.
- The average paper in top centiles of novelty has 2 to 3 times more chances to be in the top-10% most cited in its field.

Empirical evidence (II)

Table: Predicting citations and big hit probabilities (Pairwise keyword novelty)

	big hit (10%)		big hit (5%)		neg.bin (coeff.)		neg.bin (disp.)	
	3y	5y	3y	5y	3y	5y	3y	5y
Full sample	42%	45%	41%	44%	38%	37%	-15%	-4%
Human and social sciences	25%	28%	25%	29%	30%	32%	-21%	-10%
Hard science and engin	45%	48%	42%	46%	44%	39%	-15%	-4%
Life sciences	45%	48%	46%	48%	33%	33%	-15%	-4%

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

take away

- Keyword combination frequency is a way to measure article novelty
- Novelty is performed in larger and boundary spanning teams
- Novel papers attract significantly more citations.

THANK YOU!!!

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

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.

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.

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.

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.

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

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

Price, Derek John de Solla. 1963. *Little science. Big Science.* New York: Columbia University Press.

Rudwick, M. J. S. 1985. *The Great Devonian Controversy*. Chicago: University of Chicago Press.

Shapin, S. 2008. *The scientific life: A moral history of a late modern vocation*. Chicago: University of Chicago Press.

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

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.