CESPRI

Centro di Ricerca sui Processi di Innovazione e Internazionalizzazione
Università Commerciale "Luigi Bocconi"
Via R. Sarfatti, 25 – 20136 Milano
Tel. 02 58363395/7 – fax 02 58363399
http://www.cespri.unibocconi.it

Francesco Lissoni, Fabio Montobbio

Inventorship and Authorship in Patent-Publication Pairs:

an Enquiry into the Economics of Scientific Credit

WP n. 224   **November 2008**

INVENTORSHIP AND AUTHORSHIP IN PATENT-PUBLICATION PAIRS: AN ENQUIRY INTO THE ECONOMICS OF SCIENTIFIC CREDIT[⊗]

Francesco Lissoni^{◇∗}, Fabio Montobbio ^{◻∗}

[◇]Università degli studi di Brescia (Italy)
[◻]Università dell'Insubria, Varese (Italy)
[∗]KITES – Università "L. Bocconi", Milan (Italy)

francesco.lissoni@unibocconi.it, *fabio.montobbio@unibocconi.it*

Abstract

Authorship and inventorship are attribution rights that contribute to the reputation of individual scientists, but have to be distributed across several individuals, due to the importance of teamwork in both science and technology. For academic teams that both publish and patent their research results, we can compare the social and legal norms that regulate the joint distribution of these two types of attribution rights. We use text-mining techniques to identify 681 "patent-publication pairs" (related sets of patents and publications), for a sample of Italian academic scientists. On average, the number of co-authors is larger than the number of co-inventors, especially in medical-related fields. First and last authors have a lower probability of being excluded from inventorship, as suggested by patent laws. However, the probability of exclusion also declines with seniority, as expected from social norms. Long-lasting doubts on the reliability of authorship as a tool for allocating scientific credit are reinforced, and can be extended to inventorship.

JEL: Codes: O31, O34, L30

"Why does your name even appear on the paper?"

"I am the one who suggested the problem [...] I prepared the grant application to the NIH. [...] Without such support [my student] could do nothing. I'm not just talking about the fellowship. I'm talking about all the instruments in my lab, the chemicals, the glassware. [...] I see [her] almost every day; we discuss the progress of the work; I suggest certain techniques; I call important references to her attention [...] there's both a teacher-apprentice relationship and collegiality."

(Djerassi C., *Cantor's Dilemma*, Penguin Books, 1989; pp.50-51).

"I think there's rarely more than one inventor. I mean, if you wake up and you have an idea, that's the invention. And then there's all this work around it, of course ... [The postdoctoral researchers] contributed to the work, but they didn't do any really innovative work [...] They don't have time to think as much, they have a lot of manual labor to do"

(McSherry C., *Who Owns Academic Work?*, Harvard Univ. Press; 2003; p.84)

1. Introduction

Attribution rights play a key role in the economics of science. Academic scientists build up their reputation through their publications, so that authorship credits play an important role in career advancement, access to funding, or admission to learned societies. At a systemic level, the authorship-based incentive system pushes scientists to diffuse openly the results of their research, and the related data and methodologies (Merton, 1957; Dasgupta and David, 1994; Stephan, 1996; Audretsch et al., 2004).

Increasingly, academic scientists also earn scientific credit through patenting: being listed among the inventors of a well-known patent (inventorship) may bring not only economic returns in the form of licensing fees, but also reputational gains. Such a reputation is particularly valuable outside the academic community, where patents are often seen as a "proof of impact" of research conducted with public money (see, for example, the guidelines of governmental research evaluation exercises, such as RAE, 2008, and CIVR, 2006; see also OECD, 2003) Recent studies show that "academic inventors" are most often very prolific authors, whose publishing activity appears to be positively affected by their engagement in patenting (Azoulay et al., 2007; Breschi et al., 2007).

Both authorship and inventorship are forms of intellectual property (legal or 'moral' rights of attribution), which national legislations discipline according to international conventions.[1] At the same time, and especially within the scientific community, they appear to be largely administered on the basis of social norms (Zuckerman, 1968; Fasse, 1992).

---

[1] The two international treaties that matter most in this respect are the International Covenant on Economic, Social and Cultural Rights (ICESCR) and the Berne Convention. The former is a UN-administered agreement which entered into force in 1976, and it contributes to the International Bill of Rights; its article 11 protects 'the moral and material interests resulting from any scientific, literary or artistic production of which [a person] is the author' (UNESCO, 2001). The latter is an international treaty for the harmonization of national copyright laws, dating back to 1886 and now administered by WIPO, the World Intellectual Property Organization (WIPO, 2008); authors' moral rights are protected by article 6. The interpretation of moral rights, as defined by the Berne Convention, is generally much more extensive in continental Europe (esp. France) than in the UK and the USA (Fernandenz-Molina and Pais, 2001; Fisk, 2006).

Well-established studies of scientific authorship suggest that attribution norms allow for negotiations within the research teams, which may result in mis-attributions and omissions (Biagioli et al., 1999). Over the years, widespread concern has also been expressed about the practices of honorary or guest authorship in medical and physics journals (e.g. Mowatt et al., 2002; Ioannidis, 2008). More recent legal case studies suggest that the same may happen with academic inventorship (McSherry, 2003; Seymore, 2006). Since many academic inventions are both patented and described in one or more scientific publications, contrasts may also arise on the criteria to be applied for distributing authorship and inventorship over the same research product (Ducor, 2000; Murray, 2002; Murray and Stern, 2007; Gans et al., 2008).

In this paper, we assess the relative weight of legal and social norms adopted by academic scientists for the joint attribution of inventorship and authorship over "patented research results". We use an original sample of 681 "patent-publication pairs" (PPPs) produced by 308 Italian academic inventors between 1975 and 2002, in the fields of Chemical Engineering, Electronic Engineering and Telecommunications, Pharmacology, and Biology, along with related bibliometric and biographical information on the selected academic inventors and their co-authors.

We find that inventorship attribution follows stricter criteria than authorship attribution. As a result, several authors of scientific publications may be excluded from the list of inventors of the related patents (especially in Biology and Pharmacology). However, such exclusion cannot be entirely explained by the individual scientists' contribution to the research project, as legal norms on inventorship would suggest. Social norms, and in particular seniority, also matter, so that junior co-authors are more at risk of being excluded from inventorship, other things being equal. At the same time, our results are suggestive of the existence of widespread practices of guest and gift authorship attribution. We conclude that existing practices of authorship and inventorship attribution may need to be critically re-examined.

The paper is structured as follows. In section 2 we introduce the concepts of co-inventorship and co-authorship, and develop the hypotheses to test. We do so by assessing the legal and sociological literature on attribution rights, publication guidelines of various scientific journals, and technology transfer office recommendations to potential academic inventors. In section 3 we describe our methodology for the identification of PPPs, and the resulting data sample. In section 4, we provide descriptive statistics on inventorship and authorship attribution at the PPP level. In section 5 we estimate the probability for the co-author of a publication to be absent from a related patent, as a function of the co-author's contribution to the publication, seniority, and academic prestige. Section 6 concludes.

2. Guest authors and ghost inventors: problems of attribution in large research teams

Multiple-authored publications are a common feature of many scientific and technical fields, and the average number of authors per publication keeps increasing (Weeks et al., 2004). Drenth (1998) estimates that in the biomedical field the mean number of authors per paper has increased steadily from 3.21 in 1975 to 4.46 in 1995;[2] similarly, Levsky et al. (2007) calculate a 23 per cent increase in the number of authors from 1995 to 2005. Historians and sociologists of science have explained this trend with the changing nature of the scientific work, which is increasingly based on specialization, inter-disciplinarity, and the sharing of data and facilities (Katz and Martin, 1997). Some have also suggested that the growing "publish-or-perish" pressure on faculty may induce scientists to trade authorship credits in order to keep up their publication record, thus inflating the number of authors per paper (Levsky et al., 2007).

The number of multi-invented patents has also increased significantly over the last three decades, as revealed by European and US Patent Office data.[3] However, comparisons with publication data reveal that the average number of inventors per patent is well below the average number of authors per publication, even for comparable technological and scientific fields (Meyer and Bhattacharya, 2004). One common explanation for this difference is that patents originate mostly from industrial research, funded by business companies and carried out by their employees. The proprietary nature of such research limits the inventors' freedom to choose their research team partners, contrary to what happens to academic scientists.

However, differences in the number of co-authors and co-inventors can also be found when comparing patent-publication pairs, that is patents and publications which originate from the same research team and programme (Ducor, 2000; Murray, 2002). In this case, the only possible explanation for the difference is that the qualifying criteria for being considered either an author or an inventor differ, or that some differences exist in the established practices of attribution.

2.1 Honours, Gifts, and Guests: the Vexed Issue of Co-authorship

Authorship attribution considerably affects the productivity evaluation of researchers and their institutions, as well as individual careers, funding decisions, and public perceptions about science. Accordingly, concerns with the diffusion of co-authorship are of two related kinds. First, "the rapid increase of multi-authored papers [has introduced] ambiguity about the respective contributions of the

---

[2] In particular, Drenth (1998) observes a rise of authorship rate among professors and chair-persons.
[3] The average number of inventors per patent application at the EPO (European Patent Office) has increased constantly, from 1.95 in 1980 to 2.46 in 1999. When considering only patents in science-based fields such as organic chemistry, the figures are respectively 2.76 and 3.88. (as suggested by our own elaborations on the EP-CESPRI database).

joint authors" (Zuckerman, 1968: 277). Second, fears have been expressed over mis-attribution practices, such as 'guest' (or 'honorary') and 'gift' authorship, which occur when a scientist is listed in the authors' by-line of a paper to which he/she has not contributed (Drenth, 1998; Mowatt et al., 2002).[4] The resulting ambiguity about individual contributions makes publications less useful as a signal of scientific credit and threatens the ethical integrity and credibility of the entire research, from the soundness of the applied methods to the quality of data (Biagioli, 1998).

These problems are particularly severe for biomedical research, because of the great importance of responsibility attribution in that field. As a consequence, since 1985, the *International Committee of Medical Journal Editors* has continued to publish the 'Uniform Requirements for Manuscripts Submitted to Biomedical Journals'. In their latest version, the Requirements recommend the following criteria for authorship:

> "Authorship credit should be based on 1) substantial contributions to conception and design, or acquisition of data, or analysis and interpretation of data; 2) drafting the article or revising it critically for important intellectual content; and 3) final approval of the version to be published [...] Acquisition of funding, collection of data, or general supervision of the research group, alone, do not justify authorship" (ICMJE, 2007).[5]

According to the ICMJE Requirements, a heterogeneous set of authors can be listed together in the same by-line. For example, a scientist who has limited himself to an entrepreneurial role (such as chasing grants, "conceiving and designing" the paper, and revising it "critically") could be listed along with a colleague who has carried out most of the work (such as acquiring, analysing and interpreting the data, drafting the manuscript, and providing the technical expertise).[6] Despite such latitude, the ICMJE Requirements have been largely ignored by the scientific community. For example, Bates et al. (2004) find that 60 per cent of 72 articles surveyed in 2002 in the *Annals of Internal Medicine* and 21 per cent of 107 articles in the *British Medical Journal* have at least one author that does not meet the first ICJME criterion. Similar results have been found by Hwang et al. (2003) for the *Journal of Radiology* (see also references therein on studies in the *Lancet* and the *Dutch Medical Journal*).

Authorship attribution remains a highly subjective decision, which is negotiated within research teams, according to customary rules that differ across disciplines and laboratories, and do not necessarily match journal guidelines. In addition, a lack of respect of authorship criteria can be hardly detected or

---

[4] Even worse is the case of 'ghost' authors, typically junior scientists that contribute significantly to the published research, but who are mentioned only in the acknowledgements section of the paper or not mentioned at all (Rennie and Flanagin, 1994). Quantitative accounts of guest and honorary authorship are also provided by Hoen et al. (1998) and Mowatt et al. (2002), and references therein.

[5] See also the Faculty Policies on Integrity in Science of the Harvard Faculty of Medicine http://www.hms.harvard.edu/integrity/index.html. Similar rules, albeit less detailed, can be found in the authors' guidelines of the International Electrical and Electronic Engineering association (IEEE, 2008; Section 8.2.1.A).

[6] The economic logic behind rewarding through authorship a scientific entrepreneur, such as a laboratory head, is well stated in Carl Djerassi's (1989) fictitious portrait of a senior scientist explaining why she deserves co-authorship along with her junior partner, as quoted at the beginning of the paper. On Djerassi's authority as a source of information on authorship practices, see Garfield (1983).

sanctioned by the journal editors: when guest or gift authors are added to a publication, there is little risk of undermining the scientific validity of the article and the reputation that the "true" authors may derive from it.[7]

Name-order in the authors' by-line is often used to identify the individual contributions. Although general authorship guidelines (those of the ICMEJ among others) do not provide mandatory recommendations, two major traditions exist: alphabetical ordering, which is typical, for example, of social sciences, and contribution-related ordering, which is most common in the hard sciences and is explicitly recommended by some scientific societies.[8]

In their study on medical publications, Mowatt et al. (2002) calculate that 76 per cent of by-lines assign the first position to the person who contributed most significantly to the study, while only 2 per cent list authors alphabetically. Of the remaining 22 per cent, seniority criteria were involved, such as listing the senior author last. Zuckerman's (1968) seminal work on Nobel laureate authorship practices revealed early on that name ordering decisions are most often delegated to senior investigators, who base their judgement both on contribution and seniority.[9]

Summing up, the message conveyed by the first and last positions in a non-alphabetical by-line is relatively unambiguous: the first author is usually the junior scientist who has contributed most to the paper; the last is the senior investigator, who runs the lab, chases the grants, and sets the research strategy. The same cannot be said for the authors in between. These may be either effective contributors to the paper (although less important and/or more senior ones than the first author), but they may also be guest authors of many sorts (such as laboratory technicians rewarded for their dedication, or very senior scientists listed as a sign of deference). In fact, name ordering sustains the notion of authorship only partially, provides a mild correction to the mis-attribution problem, and no correction at all to the problem of responsibility.

---

[7] In some cases the lack of application of the ICMJE requirements is due both to ignorance on the scientists' part (Bhopal et al., 1997) and to lack of enforcement by the journal editors. In other cases there are intentional strategies: for example, a few pharmaceutical companies have even managed to publish papers produced by internal ghost-writers, but "authored" and submitted by complacent guest authors from the academic ranks (Ross et al., 2008, and references therein).

[8] Some professional societies, the ICMJE among them, explicitly recommend to list authors according to their contribution. For a review, see Rennie and Flanagin (1994) and Drenth (1998).

[9] In particular, the Nobel laureates interviewed by Zuckerman (who come from all disciplines) point out that a precise measurement of relative contributions is impossible, so that a pure contribution-based ordering effort would create conflicts within the research teams: ambiguity is necessary to temper tensions within the team. A major obstacle that stands in the way of contribution measurement is individuals' tendency to overestimate their own contribution within a team (Hoen et al., 1998). Such a tendency is discussed in economic terms by Van den Steen (2004). Applied sociological analysis of teams provides more general evidence of individuals' overestimation of their own importance (centrality) in teams (Johnson and Orback, 2002). Notice also that Zuckerman (1968) points out that Nobel laureates justify placing their own names last in the by-line with a *noblesse oblige* argument (the stated wish to help younger colleagues' contributions to stand out) but also admit to derogating this when the paper has the potential to attract a good deal of attention from the scientific community.

2.2 The "Muddy Metaphysics" of Co-Inventorship

Unlike scientific authorship, inventorship is a legal concept which bears direct economic consequences. In the USA, a patent may be declared invalid if the designated inventors' contribution does not match the legally defined one.[10] In addition, the inventors' names can be changed after a patent is granted only if the error is made without deceptive intent. This norm also applies to foreign patents, when extended to the USA.

According to section 35 of the US constitution (as amended in 1984), two individuals can be designated as inventors on the same patent only if they have worked "jointly" and provided some kind of "inventive" contribution (Fasse, 1992, pp. 172-173). In particular, *each* person named on a patent must have contributed to the *conception* step in the invention (as defined by the claims). Conception is "the formation, in the mind of the inventor of a definite and permanent idea of the complete and operative invention, as it is to be applied in practice" (Hybritech Inc. v. Monoclonal Antibodies, Inc.).[11]

In Europe, even with patents issued by the European Patent Office (EPO), inventorship is ultimately defined by the various national legislations. For example, in the United Kingdom the inventor is defined as the "actual deviser of the invention...", who in turn is the person(s) who contribute(s) to the novelty (inventive step) of the claims listed in the patent application (s7-3 Patents Act, 1977). In Italy, as in many other countries, no specific definition of inventor is provided by legal texts. As a matter of fact, the legal doctrine on the identification of authors and inventors coincides, with the latter being simply defined as the "author of an invention". Mis-attribution of inventorship does not appear to threaten the validity of the patent, but it may cause re-allocation of the property rights.

Criteria for defining inventorship are more restrictive than those defining authorship.[12] Being involved in the conception of the invention is a requirement that some authors of scientific publications may fail. For example, current interpretations of the US law suggest that "merely suggesting a desired result" or "having entrepreneurial involvement" do not qualify as inventorship.[13] As a result, a scientist who raises funds, conceives the initial experiment, and revises the draft paper would qualify as the author of a project-related paper (according to the ICMJE guidelines we described above), but not as the inventor of any project-related patent. At the opposite end, "following the complete instructions" of a colleague or superior does not qualify anybody as an inventor; and even joining a research team too late, after its members have conceived the key characteristics of the desired invention, may be a cause for exclusion from inventorship. The latter cases bring to mind situations in which a junior scientist or a

---

[10] See for example *Yeda Res. & Dev. v. ImClone Systems Inc.* in 2006
[11] 802 F.2d 1367, 1376 (Fed. Cir. 1986)
[12] The legal opinions provided by university TTOs and IPR consultants are particularly illustrative in this respect. For a representative subset of such opinions see: Bennett and Biswas (1997), Hutchins (2003) and Vinarov (2003).
[13] Fasse, 1992; pp. 192ff.

graduate student may be rewarded with authorship for her brilliant assistantship, but not with inventorship.[14] However, both the concept of inventorship and its application are controversial. In particular, the US literature suggests that no clear criteria exist to decide which scientists do not qualify as inventors of a research-related patent, at least on the sheer basis of their contribution to authorship (Fasse, 1992).

It is also likely that decisions on inventorship attribution, very much like those on authorship, depend upon the discretionary judgement of the most senior scientists in the team, who often manage the economic details of the research and exercise authority (case-study evidence on this point is provided by Colyvas, 2007). These scientists may be more generous towards authorship attribution, which entails only a reputational reward, than towards inventorship, which could lead to tangible economic benefits and legal hiccups.[15]

The practicalities of inventorship attribution also leave room for mistakes and abuses. Very much like journal editors, patent office examiners leave the identification of inventors entirely to the applicants. At most, signed declarations are required. If not challenged in court, these initial attributions remain unscrutinized because patent offices pay attention only to the technical contents of the patents they are called to judge.[16]

2.3 Proposed analysis

Our discussion of the literature suggests that when a team of scientists achieves a research result which has the potential to be both patented and published, the number of authors of the publication(s) will be higher than the number of inventors listed on the related patent(s). In addition, we suggest the that the exclusion of some authors from inventorship will be based upon both legal and social norms.[17]

In particular, we point out three different types of authorship that may not turn into inventorship:

---

[14] For a case of a student's exclusion from a patent, see Fasse (1992; p. 282). More cases of disputes within academic teams are mentioned by Seymore (2006).

[15] In doing so, senior authors may also be affected by a tendency to overvalue their own contribution to patents, a tendency which a questionnaire survey by Jaffe et al. (2000) has shown to be quite common. This tendency applies also to authorship, as already discussed in footnote 9. It is also well illustrated by interviews to several academic inventors, such as the one conducted by MacSherry (2003) and quoted at the beginning of the paper.

[16] Editors will check, through the referees, the relevance, originality and rigour of the paper. Examiners will check for novelty, non obviousness, and industrial applicability. It is also doubtful that a junior scientist, excluded from a patent (but possibly rewarded with authorship), will find it convenient to sue a senior colleague, upon whom her career prospects may depend heavily. Similarly, we can hardly expect the same junior scientist, who has signed both a paper and a patent, to initiate a legal action against the inclusion of an illegitimate inventor in the patent.

[17] Although it may also be possible that some patent inventors are excluded from authorship in the related publications, we do not deal with this case, for two reasons. First, as discussed above, this case is not as frequent as the opposite one. Second, the available data do not allow us to enquire in this direction. The main constraint is the absence of information on co-inventors' relative contribution to the patent, due to the lack of informative value of name ordering in the inventors' list. As discussed above, while many papers list authors according to their contribution and/or seniority, the same does not apply to patents, which usually list inventors in alphabetical order.

I. *Authorship without sufficient inventive contribution.* This category includes all scientists and laboratory technicians whose contribution has been creative enough to qualify them for authorship (according to journal guidelines), but not for inventorship (according to the rule of law). It also encompasses senior scientists whose contribution to the research project has been largely entrepreneurial.

II. *"Guest" and "gift" authorship.* This category includes laboratory technicians and other assistant figures (including graduate students and junior scientists in charge of minor tasks), who have been rewarded for their dedication or resourcefulness (gift authors) or senior scientists honoured for their seniority or prestige (guests). More generally, we include in this category all scientists who contribute marginally to a publication and should not be included according to the standard guidelines. According to legal norms on patents, none of these authors qualify as inventors.

III. *Authorship subject to arbitrary exclusion.* This category includes junior scientists who qualify both for authorship and inventorship, but who are arbitrarily excluded from the latter as a consequence of a team's decision. Although quantitative evidence on this occurrence is not available, a number of borderline cases are mentioned by McSherry (2003) and Seymore (2006).

In what follows, we try to assess the relative importance of these types of exclusion, by exploiting the information provided by the by-line of scientific publications (when the order of authors is non-alphabetical). In particular, we assume that first authors in the by-line contribute creatively and conceptually to the research effort, whereas last authors provide an "entrepreneurial contribution", possibly (but not necessarily) coupled with a creative or conceptual one. We also assume that the contribution of "middle" authors is less significant than that of other authors. Last and middle authors may include cases of "guest" and "gift" authorship.

The key social norm we consider in our analysis is the scientists' consideration for seniority. We assume that seniority affects negatively the probability of exclusion from a patent (controlling for contribution), because senior scientists' hierarchical position allows them to exercise some control over all attribution matters (authorship and inventorship).

On the basis of these assumptions, we can put forward two statements that lend themselves to empirical analysis.

1. The exclusion of first authors cannot be grounded on legal rules, and may be the result of the application of social norms; that is, it produces exclusions of type III. This intuition is strengthened if the probability of first authors to be excluded decreases with their seniority: younger scholars are more at risk of arbitrary exclusion.

2. The exclusion of last and middle authors is explained by lack of creative contribution, that is, on legal grounds. This is compatible with both type I and type II exclusion. Even in this case,

however, finding that senior scientists are less at risk of exclusion would confirm the importance of social norms.

Seniority is correlated to professional experience, as measured by an author's publication stock. We expect scientists with more experience to give a more substantial (creative, conceptual) contribution to the research project. In our analysis, therefore, we control for each author's cumulative number of publications at the time of the patent.

3. Data and Methodology.

Our database results from the integration of two different sources: the KEINS database on Italian academic inventors, and the *ISI Science Citation Index*, from which we collected the publication records of both the academic inventors and their co-authors. In section 3.1 we describe the data and in section 3.2 we explain the text-mining methods used to select the patent-publication pairs in four different disciplinary fields.

3.1 Description of the Sample

The KEINS database contains information on all academic scientists designated as inventors on EPO patent applications filed either by universities, public research organizations or business companies, for a number of European countries (Lissoni et al., 2006, 2008). It also contains information on individual characteristics of the scientists (such as age, affiliation, academic rank, discipline) and the information that can be found on the front page of their patents (including title and abstract). In this paper, we focus on Italian academic inventors who were already active in 2000, in the four disciplinary fields with the highest share of academic inventors over the total number of professors in the field, namely: Chemical Engineering (which includes technology of materials, such as macromolecular compounds), Biology, Pharmacology, and Electronic Engineering & Telecommunications, for a total of 308 academic inventors and 552 patents.[18]

For these academic inventors we collected publication data from the 1975-2003 on-line version of the ISI *Science Citation Index*.[19] Further data from the same source were collected for all their co-authors, in order to establish the latter's first year of activity (first year in which a publication in their names appears in the *Science Citation Index*) and number of publications year by year (see Appendix 1).

---

[18] Italian scientists listed in the KEINS database include professors from all ranks (assistant, associate and full). The database does not include PhD students or post-docs and other non-tenured positions

[19] A more detailed description of this initial sample can be found in Breschi et al. (2007, 2008).

3.2 Patent-publication pairs: methodology

Our methodology is based on the identification of *patent-publication pairs* (PPPs). A patent and a paper form a pair when the same idea is described to some extent in both documents, and at least one author and one inventor are the same person. This occurs when a new scientific idea coincides with a solution to a technical problem and has some degree of industrial applicability.[20] Scientific papers and patents differ widely in content, since scientific publications describe a set of theories and/or experimental results. Publications are aimed at emphasizing the originality and neatness of the results, whereas patents describe the features of a new product or process, of which they emphasize the novelty and utility, by laying out a list of claims. However, in so-called "science-based" technologies ('Pasteur's Quadrant', described by Stokes, 1997) and in engineering, it is often the case that a patentable advancement is also worth publishing in refereed journals. In this case, we may expect highly specific words to be present both in the patent and in the publication; inventors and their lawyers may also cut-and-paste a few sentences from one document into the other.

To our knowledge, only Ducor (2000), Murray (2002) and Murray and Stern (2007) have tried to build PPP databases, although with different methodologies. Ducor (2000) performed a manual search of various databases for proteins with specific genetic or aminoacid sequences, finding 40 pairs. Murray's (2002) study concerned a single patent-paper pair on tissue engineering in cartilage. Murray and Stern (2007) compared 340 articles published in *Nature Biotechnology* between 1997 and 1999 with their authors' patents at the USPTO, ending up with169 PPPs, all of them selected through careful reading of both types of documents.

The number of patents and publications needed for our analysis is so large that we could not rely on manual search and reading. So we decided to apply established quantitative methods of data mining and information retrieval. We first matched our inventors' patents to all their scientific articles published two years before or two years after the priority date of the patents. The resulting matches could be considered a set of *potential* PPPs for which we then selected the *actual* PPPs by comparing the titles and abstracts of patents and publications through co-word analysis (Bassecoulard and Zitt, 2004; Leopold et al., 2004).

Given t the priority year of a patent and i the individual listed among its designated inventors, a *potential* patent-publication pair is defined as the association between the patent and a publication that has individual i among the authors and has been published in the period [*t-2, t+2*]. After excluding all duplications (which may occur when two or more patents or two or more publications have the same co-

---

[20] The first formal definition of a patent-publication pair can be found in Murray (2002): "…in periods when scientific and technical constructs become intertwined (when scientific ideas represent not only new insights but also new technical solutions), the same idea is often inscribed in both a patent and a paper (publication), thus forming a patent–paper pair. These two "documents" form a natural experiment because they transcribe the same idea and yet the texts are distinct—a paper describes experimental results, while a patent defines utility and makes claims on inventiveness" (p. 1392).

inventors or co-authors and title), all publications with no abstracts, and all patents which their inventors declared to be unrelated to any publication of theirs, the final sample of *potential* patent-publication pairs is composed of 6810 pairs, 218 individuals, 389 patents and 2838 publications.[21]

A description of this sample is provided in Tables 1, 2 and 3 and in Figure 1. Table 1 shows the number of academic inventors by field and year of birth. Table 2 shows the number of patents by priority date and type of applicant (note that 63 per cent of patents are owned by business companies). Table 3 shows the number of publications by year of publication and author's discipline (note that 60 per cent of the publications in the sample are from scientists in the fields of Biology and Pharmacology).

[TABLE 1 here]

[TABLE 2 here]

[TABLE 3 here]

[FIGURE 1 here]

Figure 1 shows the observed frequencies for the number of authors and inventors in each of the 2838 publications and 389 patents constituting the *potential* PPPs. The distribution of the number of authors has a fatter tail to the right. The median number of authors per publication is between 4 and 5. The median number of inventors per patent is between 2 and 3. Moreover, the maximum number of inventors per patent is 21 whereas there are 23 publications with a number of authors greater than 21 (the top two publications have 337 and 517 authors respectively).

For all documents in this *potential* PPP set we examine the title and abstract, and transform them into comparable information sets. The first step of the transformation consists of removing the uninformative words (so called *stop words*) from both titles and abstracts: these are very frequent terms that are not relevant for the identification of PPPs, such as pronouns, conjunctions, and the most frequent nouns and verbs.

In the second step, we apply a traditional data-mining method, the so-called *bag of words* method, which the literature considers both simple and effective (Salton and McGill, 1983; p.13); Leopold et al., 2004). For each disciplinary field we build a complete set of words from all titles and abstracts of the patents and publications (with the exclusions of the stop words), so that each document j (patent or publication) can be represented by a vector. Each cell (i,j) in the vector has a value equal to 1 if the word i appears in document j, and 0 otherwise (Bassecoulard and Zitt, 2004). This vector

---

[21]Academic inventors' declarations on the existence (or non-existence) of publications related to their patents were collected by means of structured phone interviews. Among other things, interviewees were asked, with reference to each of their patents, whether or not they had published the results of the research leading to the patent itself. Responses were obtained from 154 out of 308 inventors, for a total of 372 patents out of 552. Overall, interviewees confirmed the existence of a patent-related publication for 86 per cent of the patents.

representation may be used to produce a large number of "similarity measures" between patents and publications. The most common one, which we use here, is the *cosine similarity measure (S)*.

If x_{ij} is the value of the binary variable for document j and word i, S measures the similarity between a document k and s as follows:

$$S(k,s) = \sum_i x_{ki} x_{si} \left/ \sqrt{\sum_i x^2_{ki}} \sqrt{\sum_i x^2_{si}} \right.$$

Theoretical values of S are in the continuous [0,1] range. In our application, S takes values comprised between 0 and 0.75. For our analysis, we select those PPPs whose S value falls in the top 10 per cent of the distribution, which is comprised between 0.145 and the maximum, for a total of 681 cases.[22]

Different to methodologies employed by other authors, our bibliometric approach does not ensure a one-to-one match between patents and publications (one patent corresponding to just one publication, and *vice versa*). This occurrence is limited to 44 cases out of 681 PPPs. Many more cases are the result of "one-to-many" matches, that is matches between one patent and several publications (more precisely, we have 76 patents matched to 271 publications, which originate as many PPPs); or of "many-to-many" matches, in which *n>1* patents jointly match *m>1* publications, for a total of 346 cases. The "many(patents)-to-one(publication)" case is much rarer, with 6 publications associated with 20 patents, for a total of 20 PPPs.

The large number of many-to-many PPPs is not unexpected: a good research project will certainly produce more than one result worthy of publication, and possibly more than one patent. As for the prevalence of one-to-many over many-to-one cases, one can reasonably presume that scientists facing patentable research results will tend to publish them separately (in order to keep the length of articles under control, or simply to follow a "salami slicing" strategy), but to patent them jointly. In fact, the patent system provides many incentives to pool several claims in a single application: for example, the application and renewal fees are set per patent, not per claim; and broader patents are more effective in protecting the economic exploitation of the invention.

---

[22] In order to check the robustness of the matching method we also used three other different selection methods to find the actual patent-publication pairs. We proceeded as follows: (1) For each potential patent-publication pair, we compared the patent and publication abstracts, and calculated the number of words that are the same in the two documents. Then we calculated the share of words that are the same in the total number of words in the patent abstract. (2) This method selected the patent-publication pairs simply on the basis of the answers to a questionnaire, filed for a subset of 154 academic inventors. In particular, we retained as actual patent-publication pairs only those by academic inventors who confirmed that the scientific research leading to the patented invention also produced technical or scientific publications. The total number of patent-publication pairs in this case is 3380. (3) It is equivalent to the bag of words, with the only difference being that cells in the vectors do not contain dummies but frequencies, that is, the number of occurrences for each word in the documents. After calculating S with this type of vectors, we selected once again the patent-publication pairs in the top 10 percentile (with S ranging from 0.206 to 0.81). The descriptive results found were similar in all respects and are available from the authors on request.

The large number of one-to-many and many-to-many PPPs also has implications for our analysis. In particular, it suggests that the appropriate unit of analysis may be the overall team of authors (inventors) listed in the set of related publications (patents). This is because, within a research team, the distribution of authorship and inventorship over different publications and patents may reflect a specific division of labour and/or some team agreements over IPRs. Accordingly, focusing on individual patents or publications within a given one-to-many or many-to-many PPP may be misleading, since an author who has been excluded from one patent can be included in a related one. Therefore we investigated the determinants of authors' exclusion both from individual patents in a PPP, and from each set of related patents.

4. Descriptive analysis

We present two types of results. In section 4.1, we consider all the selected PPPs and document the extent of the author-inventor bias. In section 4.2, we analyse author exclusion; in order to do so, we take into account only the publications whose authors are not listed in alphabetical order, and look at the position in the by-line of the excluded authors.

4.1 Number of authors and number of inventors

Table 4 shows that the average number of inventors per PPPs is 3.6, while the average number of authors is significantly higher (equal to 5). Figure 2 reports the frequency distribution of the difference between the number of authors and the number of inventors per PPP, both for the initial set of potential PPPs and for the PPPs selected through the *bag of words* method: it shows that the average and median differences between authors and inventors are reduced when moving from the first one to the latter, in particular, the distribution is more concentred around the mean, which in turn is closer to one.

[TABLE 4 here]

[FIGURE 2 here]

These results are indicative of the existence of an exclusion process. We also counted, for all many-to-many PPPs, the total number of authors and inventors to see whether an exclusion pattern at the group level could be detected, and obtained results which are very close to those of Figure 2 and Table 4: this means that even when the same publication is related to more than one patent, it may happen that one of the co-authors is excluded from *all* patents.

However, significant differences exist across disciplines. We have 178 patent-publication pairs in Biology, 201 in Pharmacology, 62 in Chemicals and Material Technology and, finally, 240 in Electronics

and Telecommunications. Figure 3 and Table 4 show that the average difference between the number of authors and the number of inventors is significantly higher than zero only in *Biology* and *Pharmacology*. In Chemical Engineering and Material Technology and Electronics & Telecommunications we find that the average number of authors and inventors are roughly the same, and the median value of the author-inventor difference across PPPs is equal to 0.

4.2 Patterns of Exclusion

In order to investigate whether a specific pattern of exclusion emerges, we select those PPPs in which the authors of the publications are *not* listed in alphabetical order. We rank the publications by number of authors and then single out the position in the by-line of the excluded authors. Results are displayed in Table 5. The table shows the number of exclusions by position of the excluded author in the by-line and number of authors per publication. For the sake of simplicity, but without loss of relevant information, we include only the publications with up to 14 authors.

Table 5 indicates that the last author listed in the by-line of the paper title, who we may presume to be the scientist heading the research team, appears to have the lowest probability of being excluded, followed by the first author. Authors in between the first and the last position are excluded relatively more often. When considering the four disciplinary fields separately, we do not detect any significant difference across fields.[23] This stylized evidence shows that the position of an author in the by-line is informative on his/her probability of exclusion from the patent, as explained in section 2.

[TABLE 5 here]

5. Estimation of the probability of exclusion

In this section we estimate the probability of an author's exclusion from inventorship in a patent-publication pair, as a function of both the author's contribution to the research effort and personal (biographical, professional) characteristics. The sample we use for the estimation is built up as follows. We start from the 681 PPPs selected using the *bag of words* method. From these we exclude: (1) all publications with only one author; (2) all the publications whose author by-line is in alphabetical order and with a number of authors greater or equal to the number of inventors; (3) all the academic inventors from the KEINS database, because for these scientists the probability of being excluded is zero

---

[23] Results are not displayed but are available from the authors.

by construction;[24] (4) two publications whose number of authors made the data collection effort daunting (36 and 42 authors, respectively). This leaves us with 467 patent-publication pairs, 184 patents, 320 publications and 899 authors. The resulting sample contains 1842 observations, each of them representing a scientist.[25] The observations are the individual authors for each PPP and therefore each scientist may enter the sample more than once if he/she has more than one publication, and/or these are related to more than one patent.[26]

Having excluded from the database all the academic inventors whose patents originated our search, the scientists included are only the co-authors. For all of them, we know the number of their yearly publications and the year of their first publication recorded in the ISI *Science Citation Index*, which we assume to mark the start of their academic career.

Our dependent variable y is the exclusion event, which is $y_{ij}=1$ if the author i of the publication is excluded from the inventorship of the related patent j and $y_{ij}=0$ otherwise. The overall percentage of exclusions in our sample is 81.92. $\Pr(y_{ij}=1 \,|\, x)$ is the probability that author i is excluded from the related patent j, conditional on a set of variables x. To estimate the determinants of this probability we take into account several variables, some related to the individuals' characteristics, others to characteristics of the PPPs. Table 6 provides summary statistics. Appendix 3 provides the correlation matrix.

[TABLE 6 here]

For the individuals' characteristics, we consider:

- The contribution to the publication, that is, the individual's position in the author by-line, as measured by two dummy variables: LAST and FIRST (middle authors being the reference case). Following the discussion in section 2 and the evidence provided in section 4.2, we expect both to bear a negative sign.

- Seniority, measured either in absolute terms or relative to the other authors in the publication. In absolute terms, we measure individual i's SENIORITY at the time of the invention, as the difference between the priority year of patent j (t_{patj}) and the year of his/her first publication (t_{fpi}). As for relative seniority, we measure it with a continuous variable, ranging from 0 to 1, defined as:

---

[24] By construction, all the publications in the sample have been authored or co-authored by the academic inventors; therefore, we may never observe their exclusion.

[25] There are 12 observations related to 10 publications with only two authors. We kept these observations in the sample. Their exclusion does not change the econometric results in any respect.

[26] If scientist i is the author of two publications, both related to the same patent B, he/she will enter our database twice; if scientist i is the author of two publications, both of them related to patents A and B, he/she will enter our database 4 times; if scientist i is the author of one publication related to just one patent, he/she will enter our database just once; the latter is the most common case that covers 32.3 per cent of the number of observations.

RELATIVE SENIORITY$_{ij}$=(t$_{fpi}$ -t$_{0j}$)/(t$_{1j}$-t$_{0j}$)

where t$_{fpi}$ is the year of the first publication of individual i, and t$_{0j}$ and t$_{1j}$ are the years of the first publication of, respectively, the most and the least experienced among all the authors of the publication.

Alternatively, we measure relative seniority with two dummy variables, MOST_SENIOR and MOST_JUNIOR, the former being equal to one for RELATIVE SENIORITY=1, the latter for RELATIVE SENIORITY=0. Following the discussion in section 4.2, we expect SENIORITY, RELATIVE_SENIORITY and MOST_SENIOR to bear a negative sign, and MOST_JUNIOR to bear a positive one.

- The professional experience of scientists is also measured in absolute and relative terms. In absolute terms, we use the stock of individual i's publications (STOCK) one year before the patent priority date (t$_{patj}$-1). In relative terms, we build the continuous variable, ranging from one to zero:

RELATIVE EXPERIENCE$_{ijtpat}$=(STOCK$_{tpati}$-STOCK$_{tpat0j}$)/(STOCK$_{tpat1j}$ - STOCK$_{tpat0j}$)

where STOCK$_{tpati}$ is individual i's stock of publications at the priority date of the patent and STOCK$_{tpat1j}$ and STOCK$_{tpat0j}$ are respectively the highest and the lowest stock of publications at t$_{patj}$ among the stocks of all the authors of all the publications related to patent j. Alternatively, we employ two dummies for the scientists with the highest and lowest scientific experience, respectively (TOP_SCHOLAR, BOTTOM_SCHOLAR). Following the discussion in section 4.2, we expect STOCK, RELATIVE_EXPERIENCE and TOP_SCHOLAR to bear a negative sign, and BOTTOM_SCHOLAR to bear a positive one.

As control variables, we consider several characteristics of each PPP:[27]

- The number of authors of the publications related to patent j (N_AUT$_j$): the larger the team of scientists, the higher the probability that some authors will be excluded, due to dilution of contributions, or more disputes over them.

- The academic inventor's discipline (ELECTRONICS, PHARMACOLOGY, BIOLOGY and CHEMISTRY). We do not know each individual's disciplinary field, but we assume it is the same as that of the academic inventor of whom he/she is co-author.

- Journal dummies: journals may differ in their tolerance of multiple authorship, with some journals adopting stricter authorship criteria than others. In addition, journal dummies may serve as

---

[27] We also control for the ownership of the patent through a series of dummy variables that indicate whether the patent application was filed respectively by one or more of the inventors, a business company, or either a university or a public research organization (including government agencies; see INDIVIDUAL, PRIVATE, and OPEN SCIENCE in Table 6). These variables are never significant in the regressions and, not having strong a priori on the expected sign of the relationship, they have been excluded from the analysis. Results are available from the authors.

controls for the disciplinary field, so we use them as an alternative to controls for the scientist's discipline

- The difference between the publication year and the priority year of the patent (DELTA_YEAR$_j$ = t_{pubj} -t_{patj}), which controls for the accuracy of our matching exercise, and is also sensitive to interpretation related to scientists' patenting strategies (see discussion in section 5.2, below)

- Time dummies for the priority dates of the patent, which may capture changes in the practice of listing inventors in patents or authors in publications over time.

The correlation matrix in Appendix 3 shows how high the correlation is between seniority and experience, and between the absolute and relative measures of the two variables. We also notice that all the listed explanatory variables appear to be correlated, either negatively or positively, with the exclusion event. Finally, we note that seniority is correlated with FIRST and LAST, respectively with a negative and a positive sign: this provides support for our assumption that first authors are likely to be junior team partners, and last authors the most senior ones.

In what follows, we first examine the relationship between the position of the author in the publication by-line and the likelihood of exclusion from the patent. Second, we examine the impact of the absolute and relative seniority and scientific status of the authors on the likelihood of exclusion. Finally, we check the robustness of our results through a number of regressions based upon a more restrictive definition of PPPs.

5.1 Results

Tables 7 and 8 display the results of our regression analysis. In particular, Table 7 reports the results for the Logit regressions and Table 8 for linear probability models. In both exercises we assume that observations are independent across individuals, but not necessarily across publications and patents by the same individual scientists. As control variables, in column (1) of both Tables 7 and 8 we consider the authors' years of academic activity (SENIORITY), scientific experience (STOCK(t-1)), the number of authors in the publication (N_AUT) and the distance in time between the publication and the related patent (DELTA_YEARS).

In columns (2) and (3) we also control for the authors' seniority and the scientific experience *relative* to their co-authors [MOST_JUNIOR, MOST_SENIOR, TOP_SCHOLAR, BOTTOM_SCHOLAR in column (2); RELATIVE_SENIORITY and RELATIVE_EXPERIENCE in column (3)]. In columns (4), (5) and (6) we replicate the same regressions of columns (1) to (3), but with the addition of dummy variables for the scientific journals.[28]

---

[28] To avoid the problem of perfect collinearity, we included dummy variables only for those journals with at least 15 observations in our sample.

[TABLE 7 and 7b here]

[TABLE 8 here]

Our results show that both first and last authors have a significantly lower probability of being excluded from inventorship than middle authors. This result holds across all specifications in Tables 7 and 8. First authors are less likely to be excluded than last ones. In Table 7b we calculate the changes in the predicted probability of exclusion for a discrete change in FIRST and LAST (with all other variables held at their mean value), based upon Logit regression (1) in Table 7: we obtain values equal to -0.17 and -0.13, respectively, which are not far from the marginal effects derived from the linear probability model.

Assuming that first authors have contributed the most, and most creatively, to the publication, these results suggest that their lower probability of exclusion reflects the rule of law. The same explanation is consistent with the assumption that last authors contribute to the research effort more and more creatively than middle authors, but less than the first ones, so that their probability of exclusion is lower than the former and higher than the latter.

Table 7 also shows that the probability of exclusion decreases significantly with the scientist's years of activity. In Logit regressions (1) and (3) of Table 7, the estimated coefficient of SENIORITY is negative and significantly different from zero; the same applies to RELATIVE_SENIORITY, in columns (3)and (6); in all these cases, controls for experience are not significant. However, when we measure seniority and experience in relative terms, but with dummy variables, it is experience that turns out to be significant, in particular BOTTOM_SCHOLAR. This suggests that, other things being equal, the scientist with the lowest publication stock in the team (in many cases, no publications at all) has the highest probability of exclusion from inventorship: this may be either a very junior scientist, or a laboratory technician. Results from linear probability regressions in Table 8 once again are in line with those of Logit regressions, and lend themselves to be interpreted as marginal values.

Table 9, Figure 4 and Figure 5 show the predicted probabilities of exclusion based upon Logit regression (1) in Table 7, for different levels of SENIORITY. The analysis of the marginal effect of SENIORITY *for individuals who are first in the by-line* shows that the first ten years of activity help to decrease the probability of exclusion by approximately 0.15. The same analysis *for individuals who are last in the by-line* suggests that the same increase in seniority decreases the probability of exclusion by approximately 0.14. The following ten years of activity (that is, from the 10th to the 20th) reduce the probability of exclusion of first and last authors respectively by 0.20 and 0.18.

[TABLE 9 here]

[FIGURES 4 and 5 here]

These results indicate that, given the same contribution to the publication (position in the by-line), a junior scientist is significantly more at risk of being excluded from the patent than a senior one. Figures 4 and 5 and the calculated marginal effects illustrate this point. Among authors who are first in the by-line, a 10-year increase in seniority gives a substantial premium in terms of reduced probability of exclusion.

A first implication of our results is that seniority matters so much compared to contribution that it is difficult to rule out the possibility of arbitrary exclusion of younger scholars from inventorship (type III exclusion).

The second implication comes from the evidence that last authors also benefit greatly from seniority: a 10-year increase in years of activity provides them with a substantial premium in terms of reduced probability of exclusion. This may be either because senior scientists' contribution is extremely valuable (so that we presume they also contribute to the invention) or because of the social norms attached to seniority. However, we control for the scientists' stock of publications, which captures experience. So we can argue that in our regressions seniority signals leadership, which places the last author in a position to decide over his/her own inclusion in the inventor list.

In section 3, we remarked that one-to-many and many-to-many patent-publication associations may originate from broad research projects where the distribution of authorship and inventorship credits is negotiated at the team level, having in mind the overall set of project deliverables. In the regression sample we have 72 publications of this type. Therefore, we have taken care of building a subsidiary database, in which we consider all the patents linked to one publication. In this case the exclusion event concerns the whole set of inventors related to all the patents matched with one single publication.

Table 10 reports the results of a set of Logit regressions identical to those of Table 7, but in which $\Pr(y_{ij}=1 \mid x)$ is the conditional probability that author i is excluded, not just from one patent, but from all the patents related to his/her publication (that is, j does not represent one of the patents related to i's publication, but the entire set of patents related to it; the set of explanatory variables x does not change).

[TABLE 10 here]

We note immediately that the sign and significance of the estimated parameters for FIRST does not change, although their magnitude decreases. Also, the estimated parameter for LAST maintains its

sign, and remains significant, albeit only at 95 per cent or 90 per cent confidence levels. All proxies for seniority and experience, both absolute and relative, maintain their sign and significance, and increase slightly in absolute terms. Only the control variable DELTA_YEARS becomes not significant. We conclude that the core of our results remains unchanged when we alter our definition of "exclusion from inventorship".

Additional controls include category variables for the most important journals, the priority year of the patent, and finally, dummy variables indicating whether the ownership of a patent is assigned to open science institutions, companies or individuals (as described in footnote 27). For most of these controls, coefficients are not statistically different from zero and no regular patterns emerge.

5.2 Robustness Check

Robustness of our results depends on the quality of our PPP dataset. A possible cause of concern is the potential mix of both false positives (patents and publications we have paired under the same PPP, while in reality they are unrelated) and false negatives (patents and publications we failed to match, but which are indeed closely related). In particular, to the extent that senior authors and laboratory heads sign more papers than junior scientists, we are concerned with the possibility that false positives could bias some of our results, particularly the estimated coefficients of last author, seniority and scientific experience.

In order to control for these potential problems, we build a more restrictive PPP sample, by selecting only the patent-publication pairs with an S similarity score in the top 5 per cent of the distribution. This leaves us with only 341 PPPs, with a minimal value of S equal to 0.174. We then run a set of Logit regressions identical to those of Table 7, the results of which are reported in Table 11. Our results confirm the negative effect of SENIORITY (or, alternatively, of RELATIVE_SENIORITY) on the probability of being excluded from inventorship, with a slight increase in the estimated coefficient. However, they do not confirm the role of scientific experience (see the lack of significance for BOTTOM_SCHOLAR). In addition, the estimated coefficients FIRST and LAST maintain the sign and significance of the original exercise, although with a somewhat larger order of magnitude

[TABLE 11 here]

If we raise the bar further, and select only the PPPs whose similarity scores fall within the top 1 per cent of the distribution (the minimal level of S reaches 0.25), we again obtain similar results. In this case we are left with 68 PPPs and 156 observations in the regression sample. In particular, estimated

parameters from the Logit regression for FIRST, LAST and SENIORITY become, respectively, -2.25***, -2.35*** and -0.19***.[29]

In order to perform a further robustness check, we may consider a different way to restrict our sample of PPPs. In particular, we may consider only the publications appearing after the priority date of the related patents, namely those for which the variable DELTA_YEARS takes a null or positive value. The rationale behind this restriction is that research teams, especially if well advised by their TTO, are more likely to publish their papers after filing the patent, in order to avoid endangering its novelty. So, we could suspect that PPPs where DELTA_YEARS<0 are more likely to be false positive than those for which DELTA_YEARS≥0.

Table 12 replicates the Logit regression of column (1) of Table 7, for two different PPP samples, one for observations with DELTA_YEARS≥0, the other one for the complementary set of observations (DELTA_YEARS<0).

[TABLES 12 and 12b here]

The results for the former sample are similar, in terms of sign and significance of the estimated parameters, to those of Table 7; the main difference being only the magnitude of FIRST and LAST parameters, which is respectively lower and higher than expected (the SENIORITY parameters also appear smaller). By contrast, the regression for DELTA_YEARS<0 returns a very high coefficient for FIRST and a non-significant one for LAST, which is consistent with the possibility that part of our results in Table 7 were affected by a bias due to the methodology followed for the creation of our PPP sample. Note, however, that the observations with DELTA_YEARS<0 account for just one third of the sample, so that we can still declare ourselves confident of the overall robustness of the results.

One reason not to exclude altogether from our sample the publications with DELTA_YEARS<0 is that there may be an alternative explanation for the results of Table 12, one that is substantive rather than purely statistical. Such an explanation is based on the possibility that, within a team of scientists, the decision to file a patent may come along two different routes, which affect differently the distribution of inventorship credits. In particular, we suspect that publications with DELTA_YEARS≥0 may be the result of a deliberate strategy, imposed by the leading authors (FIRST and LAST), which aims at avoiding the risk of invalidating the patent by publishing the invention-related research results

---

[29] The complete results are not displayed but are available from the authors.

too early.[30] In this case we can expect the probability of exclusion to be altogether low, and not much different for first and last authors, as indeed suggested by Table 12.

Conversely, when we observe a scientific article published before the filing of the related patent, we may suspect that scientists did not set out looking for a patent, but changed their minds after finding promising scientific results. In this case, the patent could be generated by the specific additional activity of the author who has contributed most to the research activity, i.e. the first author. This interpretation is consistent with the very high absolute value of the negative coefficient for FIRST (as opposed to the lack of significance for LAST) for PPPs with DELTA_YEARS<0.

The "substantive interpretation" of results in Table 12 is in line with some recent results from studies on academic inventors' publishing activity. Breschi et al. (2008) show that those academic inventors who have signed just one patent have a peak in their publications one year before the patent, which suggests that their patents are a one-off by-product of a successful, non-patent-targeted research project. On the contrary, serial academic inventors, who make joint plans for publishing and patenting, reach their publication peak at a later date, from one to three years after the first patent. Consistently, Azoulay et al. (2007) show that most patenting events are preceded by a flurry of publications.

6. Conclusions

In this paper, we have investigated the relative weight of social norms and legal rules on the determinants of authorship and inventorship attribution in patent-publication pairs. First, our analysis confirms that in patent-publication pairs the number of authors is usually larger than the number of inventors, and that this is due to the exclusion of several authors of a publication, or set of related publications, from the related patent or set of patents. We find also that the difference between the number of authors and inventors is significantly positive only in Biology and Pharmacology.

Second, we contribute to the emerging technical literature on PPPs by proving the usefulness of text-mining techniques for matching patents and publications. Our application of such techniques suggests a more complex picture of the patent-publication relationship than envisaged so far. We find that complex combinations of patents and publications are likely: not only one-to-one correspondences between individual patents and publications, but also (and mainly) several publications connected to a single patent or several publications associated with several patents.

---

[30] This is particularly true in Italy (to which our data refer) and more generally in Europe, where the "grace period" rule does not apply. On the contrary, in countries where the rule applies, such as the USA, patent offices allow inventors to submit an invention-related article to a scientific journal before filing the patent, as long as they take care to file the patent within six months after the publication of their article

Third, we provide evidence on three possible causes of exclusion from inventorship. Authors of a publication may be excluded either because their authorship contribution was substantial, but insufficient to qualify as inventorship, in accordance with legal criteria; or because their authorship contribution was by itself negligible, as in the case of "gift" and "guest" authorship; or, finally, because of mere seniority considerations, which put junior scientists more at risk of being denied inventorship, compared to senior colleagues, other things being equal.

Our results are of both practical and theoretical interest. On the practical side, we observe that inventorship attribution cannot be entirely explained by a scientist's contribution to a research project, as it should be according to the rule of law. Social norms are important, so that seniority matters considerably. To some extent, this may be simply the result of the difficulties any team of scientists meets when it comes to quantifying individual contributions to a project: borderline cases may be solved with a favourable decision for senior scientists, and a less favourable one for junior associates. However, on the basis of our results we cannot exclude that senior scientists with insufficient contribution are granted inventorship, and junior scientists with sufficient contribution are not. If this is the case, the application of social norms leaves room for litigation and generates serious concern for the university and its technology transfer offices, and all the policy makers who in recent times have greatly encouraged academic patenting. This finding reinforces mounting doubts on the efficiency of reputational systems in science, especially in medical-related fields. It sends a word of caution to all policy makers who are currently pushing to link the distribution of research funds to automated or quasi-automated bibliometric assessments of scientific productivity: such exercises rely too heavily on the questionable concept of authorship to be accepted as sensible solutions to the complexity of the economics of science.

On the theoretical side, our results contribute to the criticism of the economic and social value of the concept of scientific authorship, and extend such criticism to that of "inventorship". With the transformation of science into a collective enterprise, no immediate correspondence exists anymore between papers and authors, so that the concept of authorship becomes increasingly problematic. Our results are in line with the idea that authorship is the result of a complex web of social and legal conventions. While academic science still cherishes the idea that the scientific discovery is the result of an individual's spark of genius, other fields of human creativity have abandoned it (Fisk, 2006). In movie-making, for example, it is taken for granted that a division of labour exists between the various professional figures, so that specialized credits are awarded to each of them (directors, screenwriters, choreographers, sound-makers….); this does not prevent the existence of some prestige ranking (think of directors and stars, as opposed to more technical figures), but it allows due credit to be distributed to all participants in the creative act. Some steps in the same direction have been undertaken by several scientific journals, especially in medical sciences (*JAMA*, *Lancet*, *British Medical Journal*, *Radiology* and the *Journal of Public Health* among others), which now require scientists not merely to identify

themselves as "authors", but also to specify the exact contents of their contribution according to pre-determined categories, as recommended by the 'Authorship Task Force' set up by the Council of Science Editors in 1998 (Rennie, 1998; Biagioli et al., 1999; Hwang, 2003). A key argument put forward by the task force was precisely that in modern science the concept of 'authorship' is irreparably obsolete, and that 'contributorship' should replace it.

Patent laws may incur the same type of problems, and consider the same solutions, when it comes to dealing with inventorship. The legal figure of the inventor is also an obsolete one that dates back to a time – the XIX century – when the existence of patents had been put into question, and was defended by the creation of a public image of inventors as "heroes of the industrial revolution", that is individuals whose rights ought to be defended (Machlup and Penrose, 1950; Long, 1991; Bracha, 2005; MacLeod, 2008). To the extent that the division and specialization of the "inventive labour" is increasing (very much like the division and specialization of the "scientific labour" has been increasing over half a century), a greater articulation of the different tasks in the inventive activity is needed to avoid room for abuses and litigation. In that respect, our work can be considered not only as an exploration in the field of attribution rights in academic science and technology, but also as a first step in the direction of investigating the overall adequacy of present norms of inventorship attribution.

We think that further research is needed in order to make use of biographical and career data, and not just bibliometric information, to confirm the validity of our inferences on the role of social norms in defining authorship and inventorship, and to investigate the adequacy of those norms for organizational changes in research activities. Moreover, the analysis of inventorship attribution could be extended to corporate patents, and to the impact that mis-attribution may have on the labour market for industrial researchers. It would also be of great interest to explore differences in scientific authorship attribution across academic institutions and countries, as a result of differences in the institutional profile of academic systems, funding, and labour markets.

References

Audretsch D.B., Bozeman B., Combs K.L., Feldman M., Link A.N., Siegel D.S., Stephan P., Tassey G., Wessner C. (2004), "The Economics of Science and Technology", *Journal of Technology Transfer* 27/2, pp.155-203

Azoulay P., Ding W., Stuart T. (2007), "The determinants of faculty patenting behavior: Demographics or opportunities?", *Journal of Economic Behavior & Organization* 63/4, pp.573-576

Bassecoulard E., Zitt, M. (2004), "Patents and Publications. The Lexical Connection", in: Moed H.F., Glänzel W., Schmoch U., *Handbook of Quantitative Science and Technology Research*, Kluwer. Dordrecht, Ch. 30.

Bates T., Anić A., Marušić, M. Marušić A. (2004), "Authorship Criteria and Disclosure of Contributions" *Journal of American Medical Association* 292(1), pp.86-88

Bennett V.C., Biswas S.J. (1997), "Protecting the patentability of your collaborative research", *Nature Biotechnology* 15, pp. 472-473

Bhopal R., Rankin J., McColl E., Thomas L., Kaner E., Stacy R., Pearson P. (1997), "The vexed question of authorship: views of researchers in a British medical faculty", *British Medical Journal* 314, p. 1009.

Biagioli M. (1998), "The Instability of Authorship: Credit and Responsibility in Contemporary Biomedicine", *FASEB Journal* 12, pp.3-16

Biagioli M., Crane J., Derish P., Gruber M., Rennie D., Horton R. (1999), *Authorship Task Force White Paper*, Council of Science Editors (http://www.councilscienceeditors.org/services/atf_whitepaper.cfm, last accessed: May 2008)

Bracha O. (2005), *Owning Ideas: A History of Anglo-American Intellectual Property (Ch.4: United States Patents)*, S.J.D. Dissertation, Harvard Law School

Breschi S., Lissoni F., Malerba F. (2003) "Knowledge Relatedness in Firm Technological Diversification", *Research Policy* 32/1, pp.69-87

Breschi S., Lissoni F., Montobbio F. (2008). University patenting and scientific productivity. A quantitative study of Italian academic inventors. European Management Review. *The Journal of the European Academy of Management* 5(2): 91-109

Breschi S., Lissoni F., Montobbio F. (2007), The scientific productivity of academic inventors: new evidence from Italian data, *Economics of Innovation and New Technology* 16/ 2, pp.101-118

CIVR (2006), *Valutazione triennale della ricerca 2001-2003. Relazione Finale*, Comitato d'Indirizzo per la Valutazione della Ricerca, Roma (http://vtr2006.cineca.it/index_EN.html; last access: May 2008)

Colyvas J.A. (2007), "From divergent meanings to common practices: The early institutionalization of technology transfer in the life sciences at Stanford University", *Research Policy* 36, pp. 456-76

Dasgupta P., David P.A. (1994), "Toward a new economics of science", *Research Policy* 23, pp.487-521

Djerassi C. (1989), *Cantor's Dilemma*, Penguin Books, London

Drenth JP. (1998), "Multiple authorship: the contribution of senior authors", *Journal of the American Medical Association* 280, pp. 219-21

Ducor P. (2000), "Coauthorship and Coinventorship", *Science* 289, pp.873-875

Engelsman E. C., van Raan A.F.J. (1992), "A patent-based cartography of technology", *Research Policy* 23, pp. 1-26.

Fasse W.F. (1992), "The Muddy Metaphysics of Joint Inventorship: Cleaning Up after the 1984 Amendments to 35 U.S.C. § 116", *Harvard Journal of Law and Technology* 5, pp.73-74

Fernandenz-Molina J.C., Pais E. (2001), "The Moral Rights of Authors in the Age of Digital Information", *Journal of the American Society for Information Science and Technology* 52/2, pp. 109-117

Fisk C.L. (2006), "Credit Where It's Due: The Law and Norms of Attribution", *Georgetown Law Journal* 95/1, pp.49-118

Gans JS, Murray, F., Stern S. (2008); Patents, Papers & Privacy: The Disclosure Of Scientific And Commercial Knowledge. Paper presented at the DRUID Conference 2008, Copenhagen

Garfield E. (1983), "Carl Djerassi: Chemist and Entrepreneur", *Chemtech* 13, pp. 534-538

Hoen W.P., Henk C.W., Overbeke A.J.P.M. (1998), "What Are the Factors Determining Authorship and the Order of the Authors' Names?", *Journal of American Medical Association* 280, pp. 217-218

Hutchins M. (2003), "Common mistakes that undermine patent protection and how to avoid them", *International Journal of Medical Marketing* 3, pp. 204-211

Hwang S.S. et al. (2003), "Researcher Contributions and Fulfillment of ICMJE Authorship Criteria: Analysis of Author Contribution Lists in Research Articles with Multiple Authors", *Radiology* 22, pp.16-23

IEEE (2008), *IEEE Publication Services and Products Board Operations Manual (revised version)*, (http://www.ieee.org/portal/cms_docs_iportals/iportals/publications/PSPB/opsmanual.pdf)

ICMJE (2007), *Uniform Requirements for Manuscripts Submitted to Biomedical Journals: Writing and Editing for Biomedical Publication*, International Committee of Medical Journal Editors (http://www.icmje.org)

Ioannidis J.P.A. (2008) Measuring Co-Authorship and Networking-Adjusted Scientific Impact. *PLoS ONE 3*(7): e2778. doi:10.1371/journal.pone.0002778

Jaffe A.B., Trajtenberg M., Fogarty M.S. (2000) "Knowledge Spillovers and Patent Citations: Evidence from a Survey of Inventors", *American Economic Review* 90/ 2, pp. 215-218

Johnson J.C., Orback M.K. (2002), "Perceiving the political landscape: ego biases in cognitive political networks", *Social Networks* 24, pp.291-310

Katz J.S., Martin B.R. (1997), "What is research collaboration?", *Research Policy* 26, pp. 1-18

Klavans R., Boyack K.W. (2006), "Identifying a Better Measure of Relatedness for Mapping Science", *Journal of the American Society for Information Science and Technology* 57/2, pp.251-263.

Leopold E., May M., Paaß (2004), "Data Mining and Text Mining for Science & Technology Research", in: Moed H.F., Glänzel W., Schmoch U. (eds.), *Handbook of Quantitative Science and Technology Research*. Kluwer, Dordrecht

Levsky M.E., Rosin A., Coon T.P., Enslow W.L., Miller M.A. (2007), "A Descriptive Analysis of Authorship within Medical Journals, 1995-2005", *Southern Medical Journal* 100/4, pp. 371–375.

Lissoni F., Llerena P., McKelvey M., Sanditov B. (2008), "Academic Patenting in Europe: New Evidence from the KEINS Database", *Research Evaluation* (forthcoming)

Lissoni F., Sanditov B., Tarasconi G. (2006), "The Keins Database on Academic Inventors: Methodology and Contents", *CESPRI Working Paper* 181, Bocconi University

Long P.O. (1991), "Invention, Authorship, 'Intellectual Property', and the Origin of Patents: Notes toward a Conceptual History", *Technology and Culture* 32/4, pp.846-884

MacLeod C. (2008), *Heroes of Invention: Technology, Liberalism and British Identity, 1750-1914*, Cambridge University Press

Machlup F., Penrose E. (1950), "The Patent Controversy in the Nineteenth Century", *Journal of Economic History* 10/1, pp. 1-29

McSherry C. (2003), *Who Owns Academic Work*, Harvard University Press, Cambridge MA

Merton R.K. (1957), *Social Theory and Social Structure*, Free Press, Glencoe, Ill.

Meyer M., Bhattacharya S. (2004), "Commonalities and differences between scholarly and technical collaboration. An exploration of co-invention and co-authorship analyses", *Scientometrics* 61, pp. 443-456

Mowatt, G., Shirran, L., Grimshaw J.M., Rennie D., Flanagin A., Yank V., MacLennan G., Gotzsche P.C., Bero L.A. (2002), "Prevalence of Honorary and Ghost Authorship in Cochrane Reviews", *Journal of the American Medical Association* 287, pp.2769-2771

Murray F. (2002) "Innovation as co-evolution of scientific and technological networks: exploring tissue engineering", *Research Policy* 31, pp. 1389–1403

Murray F., Stern S. (2007) "Do formal intellectual property rights hinder the free flow of scientific knowledge? An empirical test of the anti-commons hypothesis", *Journal of Economic Behavior & Organization* 63/4, pp.648-687

OECD (2003), *Turning Science into Business: Patenting and Licensing at Public Research Organisations*, Organization for Economic Co-operation and Development, Paris

RAE (2008), *Research Assessment Exercise*, Higher Education Funding Council for England (http://www.rae.ac.uk/; last accessed; May 2008)

Rennie D. (1998), "Freedom and Responsibility in Medical Publication: Setting the Balance Right", *Journal of American Medical Association* 280, pp.300-302

Rennie D., Flanagin A. (1994), "Authorship! Authorship! Guests, ghosts, grafters, and the two-sided coin", *Journal of American Medical Association* 271, pp. 469-471.

Ross J.S., Hill K.P., Egilman D.S., Krumholz H.M. (2008), "Guest Authorship and Ghostwriting in Publications Related to Rofecoxib. A Case Study of Industry Documents from Rofecoxib Litigation", *Journal of American Medical Association* 299, pp. 1800-1812

Salton G., McGill M.J. (1983), Introduction to Modern Information Retrieval, McGrawHill, New York

Seymore S.B. (2006), "My Patent, Your Patent, or Our Patent? Inventorship Disputes within Academic Research Groups", *Albany Law Journal of Science and Technology* 16, pp.125-167

Stephan P. (1996), "The Economics of Science," *Journal of Economic Literature*, Vol XXXIV, pp. 1199-1235.

Stokes, D. (1997), *Pasteur's Quadrant: basic science and technological innovation*. Washington D.C.: The Brookings Institution.

UNESCO (2001), *A Guide to Human Rights. Institutions, Standards, Procedures*, United Nations Educational, Scientific and Cultural Organization, Paris

Van den Steen E. (2004), "Rational Overoptimism (and Other Biases)", *American Economic Review* 94/4, pp. 1141-1151

Vinarov S.D. (2003), "Patent protection for structural genomics-related inventions", *Journal of Structural and Functional Genomics* 4, pp. 191-209

Weeks W.B., Wallace A.E., Kimberly B.C.S. (2004), "Changes in authorship patterns in prestigious US medical journals", *Social Science & Medicine* 59, pp.1949-1954

WIPO (2008), *Berne Convention for the Protection of Literary and Artistic Works*, World Intellectual Property Organization, Geneva (http://www.wipo.int/treaties/en/ip/berne/; last accessed: May 2008)

Zuckerman H.A. (1968), "Patterns of Name Ordering among Authors of Scientific Papers: A Study of Social Symbolism and Its Ambiguity", *American Journal of Sociology* 74/3, pp.276-291

TABLES

Table 1 Number of academic inventors by field and years of birth

Fields	*Years* 1925-1939	1940-1954	1955-1969	Total
Pharmacology	15	24	21	60
Biology	12	30	17	59
Chemical Engineering. & Materials Technology	10	21	14	45
Electronics and Telecom Engineering	4	24	26	54
Total	41	99	78	218

Table 2. Number of patents by priority date and type of assignee

Years	*TYPE* INDIVIDUAL	OPEN	PRIVATE	n.a.	Total
1988-1991	37	29	115	22	203
1992-1994	3	33	108	25	169
1995-1997	12	41	144	22	219
1998-2000	5	10	62	9	86
Total	57	113	429	78	677

The total no. of patents considered is 389. Double counting occurs for multi-applicant patents

Table 3. Number of publications by year of publication and fields

Years	*Fields* Pharmacology	Biology	Chemical Eng. & Materials	Electronics & Telecom Eng.	Total
1990-1993	195	263	121	144	723
1994-1996	282	407	169	278	1136
1997-1999	229	340	87	301	957
2000-2002	43	37	16	92	188
Total	749	1047	393	815	3004

The total number of publications considered is 2838. Double counting occurs for publications co-authored by two or more scientists from different disciplines.

Table 4: Summary statistics on potential and selected patent-publication pairs, total samples and by scientists' field.

	Average no. of authors	MIN no. of authors	MAX no. of authors	Average no. of inventors	MIN no. of inventors	MAX no. of inventors	Avg difference between authors and inventors[a]	MIN diff. between authors and inventors[a]	MAX diff. between authors and inventors[a]	Median diff. between authors and inventors[a]
Potential PPPs	*8.51*	*1*	*517*	*3.62*	*1*	*21*	*4.89*	*-18*	*515*	*1*
Selected PPPs (Bag of words) ↳of which:	5.00	1	42	3.36	1	21	1.64	-18	37	1
Pharmacology	6.47	2	14	3.75	1	10	2.71	-5	12	2
Biology	6.32	2	42	3.60	1	21	2.72	-18	37	3
Chemical Eng. & Materials Tech.	4.54	1	8	4.67	2	11	-0.13	-4	5	0
Electronics and Telecom	3.63	1	19	2.99	1	6	0.63	-3	16	0

(a) This column refers to average min, max and median values of the difference between the number of authors and inventors across each patent-publication pair.

Table 5 Count of exclusions from inventorship, by position of the author in the by-line and no. of authors per publication (publications with up to 14 co-authors)

Number of authors	Number of PPPs	Position of the excluded author in the by-line														None*
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	
2	19	9	6													3
3	97	47	63	35												6
4	98	52	57	68	39											6
5	69	40	49	50	44	29										
6	58	29	37	40	44	37	26									3
7	51	25	37	27	38	36	36	26								1
8	19	13	12	13	15	14	12	9	9							
9	26	12	21	19	18	20	20	20	17	8						
10	8	4	4	6	6	6	5	6	5	3	4					
11	13	10	9	12	11	12	12	12	13	11	11	5				
12	14	7	9	6	10	11	13	13	12	11	10	11	11			
13	2	0	0	1	0	1	1	0	1	1	1	0	0	0		
14	1	0	0	0	1	1	0	1	1	1	1	1	1	1	1	

* No co-author excluded

Table 6 Summary statistics for the regression sample.

Variable	Obs	Mean	Std. Dev.	Min	Max
Y	1842	0.82	0.385	0	1
First	1842	0.17	0.372	0	1
Last	1842	0.13	0.340	0	1
Seniority	1842	7.59	7.67	-2	26
Stockt_1	1842	18.40	35.08	0	299
N_aut	1842	7.25	3.33	2	19
Delta_years	1842	0.48	1.29	-2	2
Most_junior	1842	0.36	0.478	0	1
Most_senior	1842	0.14	0.357	0	1
Top_scholar	1842	0.09	0.289	0	1
Bottom_scholar	1842	0.33	0.469	0	1
Relative seniority	1842	0.38	0.384	0	1
Relative experience	1842	0.23	0.318	0	1
Chemistry	1842	0.04	0.189	0	1
Electronics	1842	0.234	0.423	0	1
Pharma	1842	0.18	0.386	0	1
Biology	1842	0.55	0.498	0	1
Open	1842	0.22	0.413	0	1
Private	1842	0.68	0.467	0	1
Individual	1842	0.12	0.331	0	1

Table 7. Probability of exclusion from inventorship: Logit regressions.

Dep. Var. $=y_{ij}$

	(1)	(2)	(3)	(4)	(5)	(6)
FIRST	-1.11***	-1.02***	-1.10***	-1.07***	-1.01***	-1.08***
	(0.29)	(0.26)	(0.28)	(0.28)	(0.26)	(0.28)
LAST	-0.81***	-0.83***	-0.80***	-0.82***	-0.83***	-0.80***
	(0.24)	(0.26)	(0.24)	(0.23)	(0.24)	(0.23)
N_AUT	0.055*	0.071**	0.062**	0.06	0.072*	0.065*
	(0.032)	(0.031)	(0.032)	(0.04)	(0.037)	(0.039)
DELTA_YEARS	-0.21***	-0.21***	-0.18***	-0.26***	-0.27***	-0.24***
	(0.056)	(0.059)	(0.058)	(0.06)	(0.062)	(0.060)
SENIORITY	-0.08***			-0.08***		
	(0.018)			(0.02)		
STOCK(t_1)	0.0041			0.004		
	(0.0032)			(0.003)		
MOST_JUNIOR		0.59*			0.45	
		(0.31)			(0.29)	
MOST_SENIOR		-0.20			-0.27	
		(0.36)			(0.31)	
TOP_SCHOLAR		0.29			0.27	
		(0.34)			(0.33)	
BOTTOM_SCHOLAR		0.97***			0.90***	
		(0.33)			(0.31)	
RELATIVE SENIORITY			-1.53***			-1.39***
			(0.39)			(0.38)
RELATIVE EXPERIENCE			0.40			0.29
			(0.40)			(0.42)
Constant	-0.38	-0.94	0.48	1.54	0.63	1.68***
	(1.74)	(2.02)	(1.89)	(0.59)	(0.60)	(0.61)
Observations	1842	1842	1842	1842	1842	1842
Pseudo R-squared	0.14	0.16	0.14	0.15	0.16	0.15
Time dummies	Y	Y	Y	Y	Y	Y
Fields dummies	Y	Y	Y	N	N	N
Journal dummies	N	N	N	Y	Y	Y

Note: Robust standard errors in parentheses (clustered for individuals)

*** $p<0.01$, ** $p<0.05$, * $p<0.1$

Table 7b. Change in predicted probabilities of exclusion from inventorship, for changes in the author's position in the by-line, as from regression (1)

	Excluded	Non Excluded
First	0.71	0.29
Not first	0.88	0.12
Difference	-0.17	
Last	0.75	0.25
Not Last	0.87	0.13
Difference	-0.13	

Table 8. Probability of exclusion from inventorship: Linear Probability regressions

Dep. Var. $= y_{ij}$

	(1)	(2)	(3)	(4)	(5)	(6)
FIRST	-0.15***	-0.15***	-0.15***	-0.15***	-0.15***	-0.16***
	(0.05)	(0.051)	(0.055)	(0.05)	(0.051)	(0.053)
LAST	-0.12***	-0.12***	-0.11***	-0.11***	-0.12***	-0.12***
	(0.04)	(0.044)	(0.042)	(0.04)	(0.042)	(0.041)
N_AUT	0.0065*	-0.025***	-0.020***	0.0084*	0.0073*	0.0075*
	(0.004)	(0.0070)	(0.0067)	(0.004)	(0.0041)	(0.0042)
DELTA_YEARS	-0.025***	0.0078**	0.0076**	-0.030***	-0.032***	-0.028***
	(0.007)	(0.0037)	(0.0038)	(0.007)	(0.0080)	(0.0076)
SENIORITY	-0.011***			-0.011***		
	(0.003)			(0.003)		
STOCK(t_1)	0.00064			0.00083*		
	(0.0004)			(0.0004)		
MOST_JUNIOR		0.060*			0.043	
		(0.035)			(0.035)	
MOST_SENIOR		-0.034			-0.041	
		(0.055)			(0.050)	
TOP_SCHOLAR		0.039			0.036	
		(0.053)			(0.054)	
BOTTOM_SCHOLAR		0.11***			0.10***	
		(0.037)			(0.035)	
RELATIVE SENIORITY			-0.20***			-0.18***
			(0.061)			(0.061)
RELATIVE EXPERIENCE			0.049			0.038
			(0.067)			(0.070)
CONSTANT	0.55*	0.40	0.56*	0.67***	0.63***	0.75***
	(0.3)	(0.34)	(0.32)	(0.1)	(0.11)	(0.12)
Observations	1842	1842	1842	1842	1842	1842
Pseudo R-squared	0.14	0.15	0.13	0.13	0.14	0.13
Time dummies	Y	Y	Y	Y	Y	Y
Fields dummies	Y	Y	Y	N	N	N
Journal dummies	N	N	N	Y	Y	Y

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 9. Change in predicted probability of exclusion from inventorship, for changes in author's position in the by-line and different levels of SENIORITY, as from regression (1) in Table 7

	Last author		First author	
Seniority	Non Excluded	Excluded	Non Excluded	Excluded
0	0.15	0.85	0.18	0.82
5	0.21	0.79	0.25	0.75
10	0.29	0.71	0.33	0.67
15	0.37	0.62	0.43	0.57
20	0.47	0.53	0.53	0.47

Table 10. Probability of exclusion from inventorship (entire set of publication-related patents): Logit regressions.

	(1)	(2)	(3)	(4)	(5)	(6)
FIRST	-0.77**	-0.69**	-0.76**	-0.74**	-0.68**	-0.74**
	(0.3)	(0.30)	(0.30)	(0.3)	(0.31)	(0.32)
LAST	-0.51*	-0.56*	-0.50*	-0.53*	-0.58**	-0.52*
	(0.3)	(0.29)	(0.26)	(0.3)	(0.29)	(0.27)
N_AUT	0.081**	0.10**	0.087**	0.12**	0.13***	0.12***
	(0.04)	(0.042)	(0.040)	(0.05)	(0.048)	(0.047)
DELTA_YEARS	-0.062	-0.074	-0.020	-0.094	-0.11	-0.052
	(0.06)	(0.064)	(0.063)	(0.08)	(0.081)	(0.079)
SENIORITY	-0.096***			-0.096***		
	(0.02)			(0.02)		
STOCK(t_1)	0.0050			0.0044		
	(0.004)			(0.004)		
MOST_JUNIOR		0.48			0.44	
		(0.35)			(0.34)	
MOST_SENIOR		-0.31			-0.32	
		(0.52)			(0.46)	
TOP_SCHOLAR		0.46			0.33	
		(0.47)			(0.44)	
BOTTOM_SCHOLAR		1.21***			1.12***	
		(0.28)			(0.26)	
RELATIVE SENIORITY			-1.73***			-1.69***
			(0.56)			(0.53)
RELATIVE EXPERIENCE			0.47			0.35
			(0.57)			(0.55)
Constant	0.71	-0.78	0.78	1.21*	0.13	1.33*
	(1.6)	(1.90)	(1.72)	(0.7)	(0.68)	(0.73)
Observations	1842	1842	1842	1842	1842	1842
Pseudo R-squared	0.13	0.14	0.12	0.15	0.15	0.14
Time dummies	Y	Y	Y	Y	Y	Y
Fields dummies	Y	Y	Y	N	N	N
Journal dummies	N	N	N	Y	Y	Y

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 11. Probability of exclusion from inventorship: Logit regressions. (restricted PPP sample: top 5% of S score)

Dep. Var. $=y_{ij}$

	(1)	(2)	(3)	(4)	(5)	(6)
FIRST	-0.96***	-0.87***	-0.97***	-0.93***	-0.86***	-0.96***
	(0.4)	(0.34)	(0.36)	(0.4)	(0.33)	(0.35)
LAST	-0.95***	-0.98***	-0.98***	-0.96***	-1.02***	-0.98***
	(0.3)	(0.31)	(0.29)	(0.3)	(0.29)	(0.28)
N_AUT	0.015	0.045	0.032	0.083*	0.11**	0.098**
	(0.05)	(0.050)	(0.053)	(0.05)	(0.042)	(0.045)
DELTA_YEARS	-0.25***	-0.23***	-0.22***	-0.29***	-0.26***	-0.24***
	(0.08)	(0.078)	(0.077)	(0.08)	(0.083)	(0.080)
SENIORITY	-0.094***			-0.092***		
	(0.02)			(0.02)		
STOCK(t_1)	0.0086**			0.010**		
	(0.004)			(0.004)		
MOST_JUNIOR		0.99*			0.86*	
		(0.52)			(0.48)	
MOST_SENIOR		-0.20			-0.20	
		(0.38)			(0.36)	
TOP_SCHOLAR		0.25			0.44	
		(0.41)			(0.40)	
BOTTOM_SCHOLAR		0.35			0.34	
		(0.49)			(0.46)	
RELATIVE SENIORITY			-2.05***			-1.97***
			(0.41)			(0.45)
RELATIVE EXPERIENCE			1.18**			1.16**
			(0.50)			(0.54)
CONSTANT	0.80	1.70*	0.67	1.20	0.19	1.25
	(1.1)	(0.99)	(1.08)	(0.8)	(0.80)	(0.86)
Observations	878	878	878	878	878	878
Pseudo R-squared	0.15	0.15	0.15	0.14	0.13	0.14
Time dummies	Y	Y	Y	Y	Y	Y
Fields dummies	Y	Y	Y	N	N	N
Journal dummies	N	N	N	Y	Y	Y

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 12.   Probability of exclusion from inventorship: Logit regressions, by values of DELTA_YEARS

Dep. Var. $=y_{ij}$

	DELTA_YEARS>0	DELTA_YEARS<0:
FIRST	-0.80***	-2.51***
	(0.3)	(0.5)
LAST	-0.96***	-0.22
	(0.2)	(0.6)
SENIORITY	-0.069***	-0.16***
	(0.02)	(0.04)
STOCK(t_1)	0.0050	0.00038
	(0.004)	(0.005)
N_AUT	0.063**	0.082
	(0.03)	(0.1)
DELTA_YEARS	0.10	-0.41
	(0.09)	(0.3)
Constant	-0.70	4.10
	(1.6)	(2.8)
Observations	1358	484
Pseudo R-squared	0.30	0.16
Time dummies	Y	Y
Fields dummies	Y	Y
Journal dummies	N	N

Note: Robust standard errors in parentheses (clustered for individuals)
*** p<0.01, ** p<0.05, * p<0.1

Table 12b. Change in predicted probabilities of exclusion from inventorship, for changes in the author's position in the by-line, by values of DELTA_YEARS

	Excluded	Non Excluded
With DELTA_YEARS≥0		
First	0.72	0.28
Not first	0.84	0.15
Difference	-0.12	
With DELTA_YEARS<0		
First	0.73	0.27
Not first	0.97	0.03
Difference	-0.24	

FIGURES
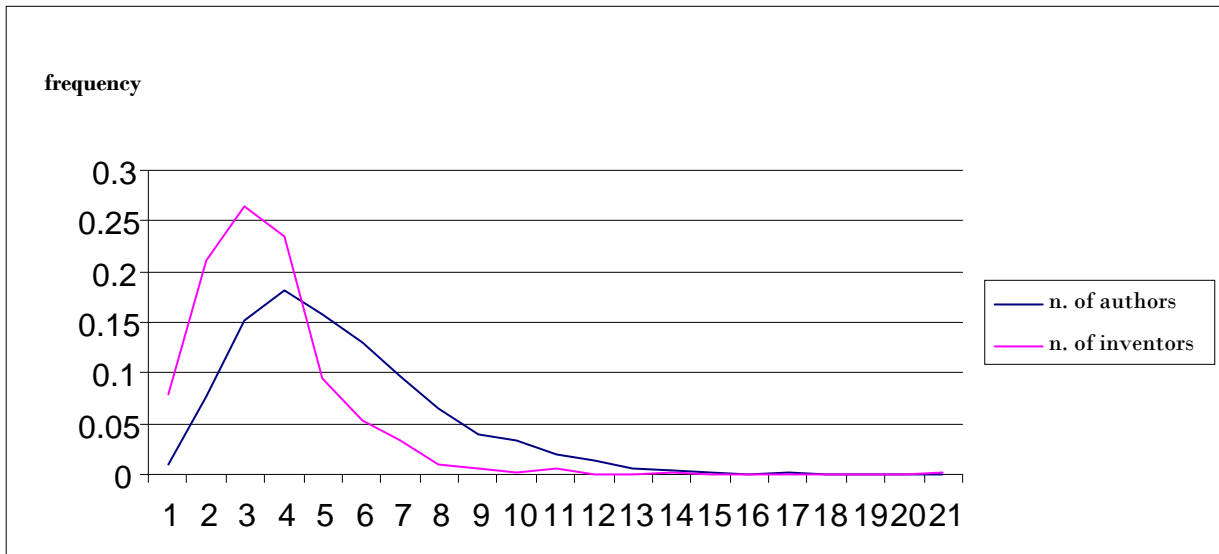
Note: This figure refers to the total sample of 2838 publications and 389 patents. The maximum number of co-inventors is 21. There are 23 publications with a number of authors greater than 21 that are not included in the figure.
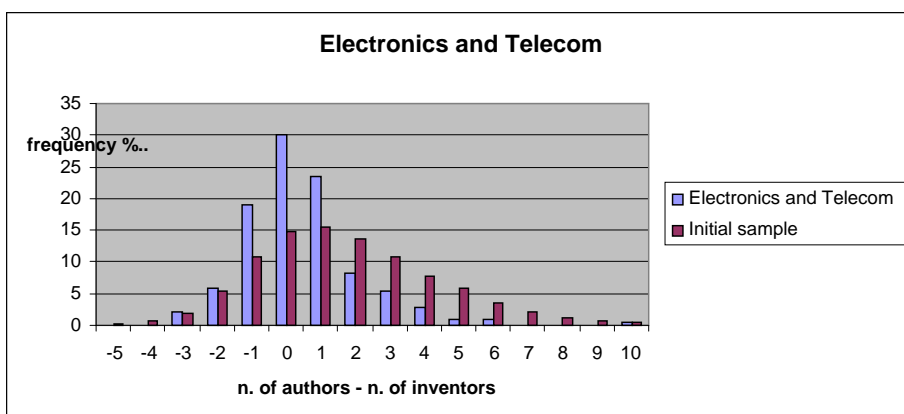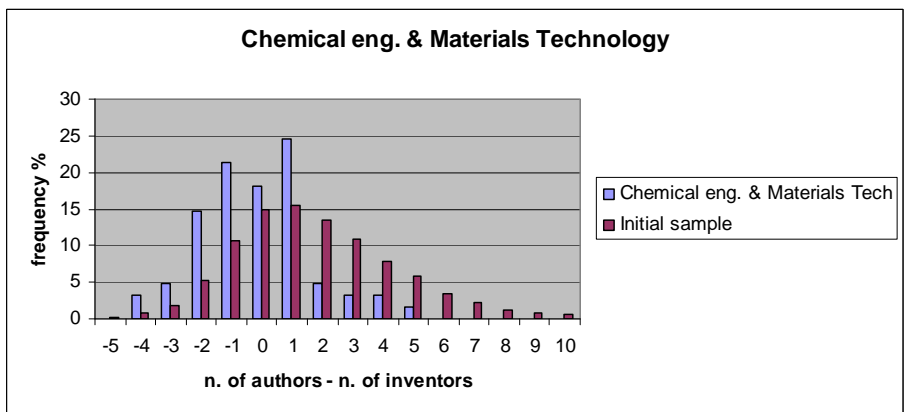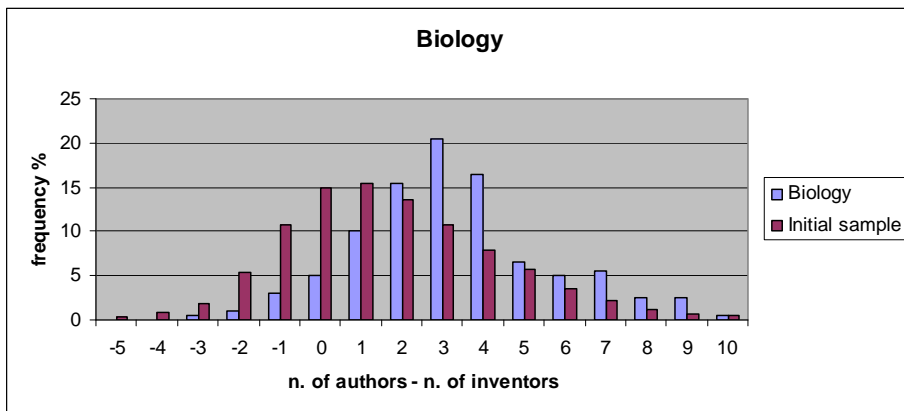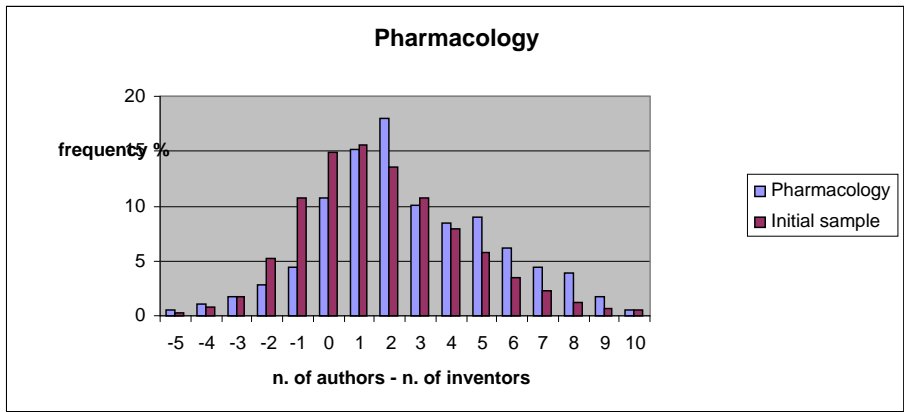
Figure 1: Distribution of number of authors per publication and number of inventors per patent.
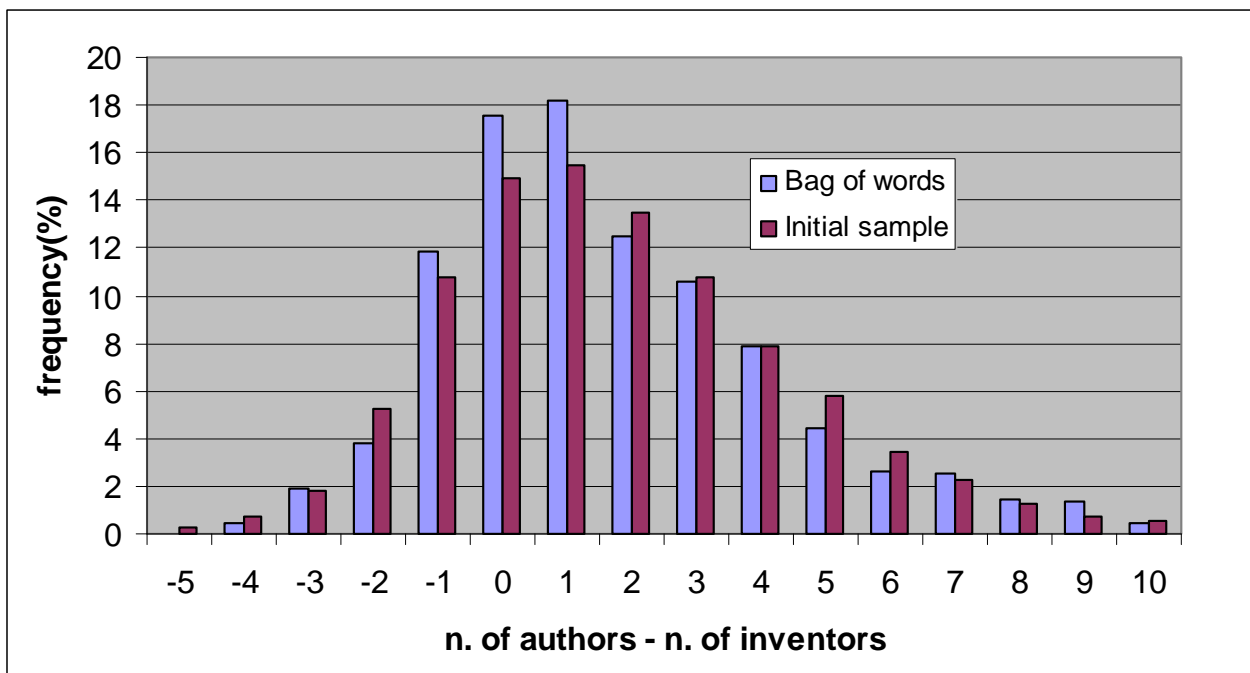
Note: Initial sample refers to the complete set of 6810 potential patent-publication pairs

Figure 2. Observed frequency of the gap between the number of authors and the number of inventors

Note: Initial sample refers to the complete set of 6810 potential patent-publication pairs

Figure 3. Observed frequency of author-inventor number gap, by scientific field.
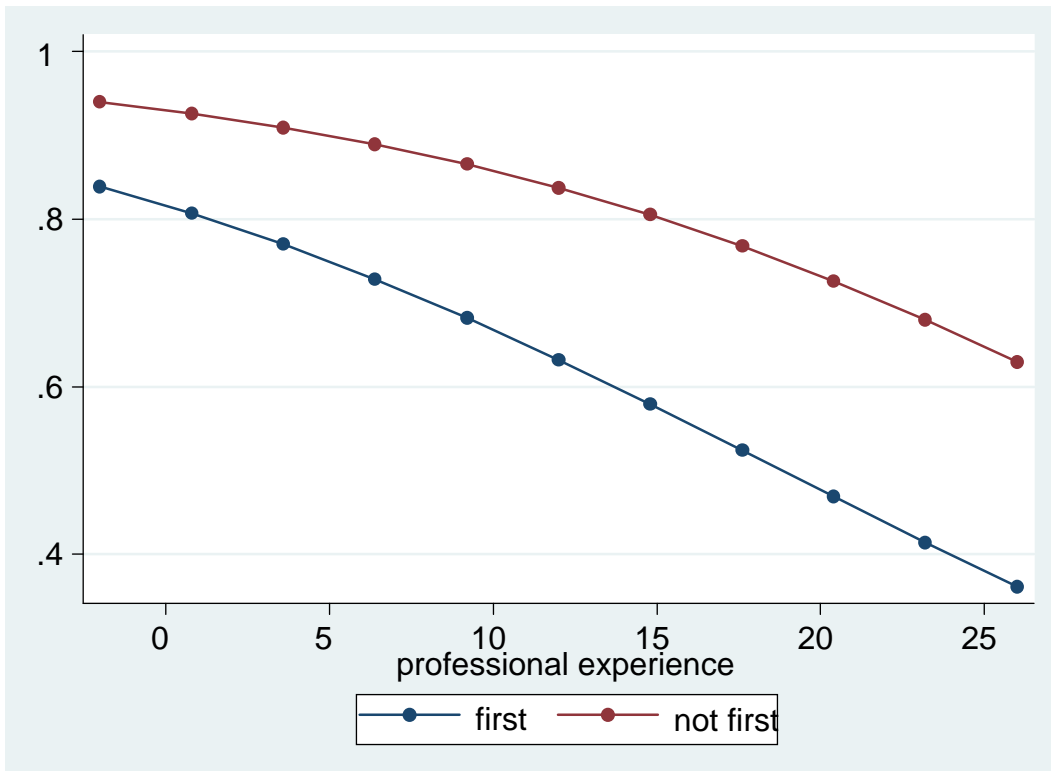
Figure 4. Predicted probabilities from the Logit model (1) by seniority and position in the by-line
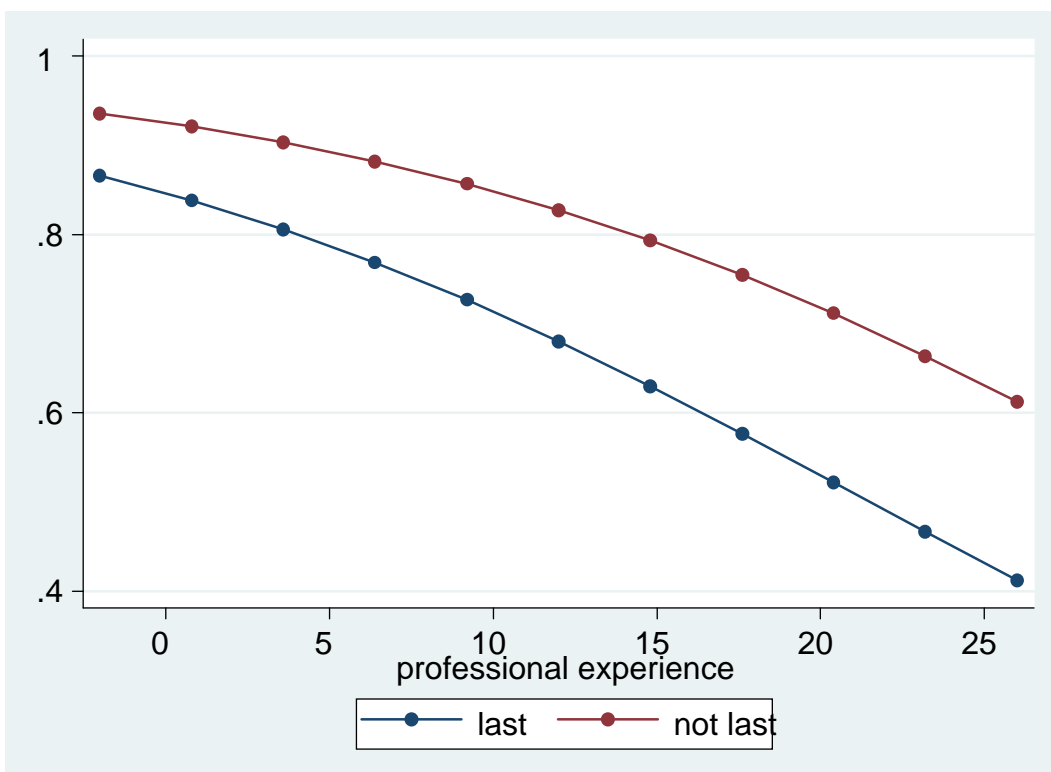

Figure 5. Predicted probabilities from the Logit model (1) by seniority and position in the by-line

Appendix 1 – Methodology for the collection of publication data

The collection of publication data followed two steps, both of which applied to the ISI Web of Science database (ISI-WoS).

In the first step, data were collected on the publications of the selected academic inventors, with the ultimate purpose of identifying the patent-publication pairs (PPPs) necessary to our analysis. In this step, the collection effort was limited to the journal articles published in the period from two years before to two years after the filing date of the relevant patents, authored or co-authored by scientists whose names and surnames coincided with those of the academic inventors in which we were interested. No restrictions were applied in terms of the journal's field, but journals whose title was clearly linked to disciplines different from that of the academic inventor were manually deleted.

In the second step, data were collected for the publications of scientists listed as co-authors of the academic inventors' publications included in selected PPPs (academic inventors were not included in the regressions, their probability of exclusion being zero, by definition). Due to the high number of scientists, serious problems of homonymy arose which we attempted to reduce by selecting, for each co-author, only the publications which fell in *similar fields* to those of the related academic inventors. [Note that we could not select only the co-authors' publication from the same fields as the academic inventors, given the possibility that the former and the latter are from different subdisciplines, with a different set of reference journals]

In order to do so, we applied a procedure originally proposed by Engelsman and Van Raan (1992) (see also Breschi et al., 2003). In particular, we applied such a procedure to the ISI-WoS classification of the journals wherein the academic inventors' articles were published, for a total of 99 subfields. [31]

Let I be the set of academic inventors. Let $F_{ik} = 1$ if professor i signed at least one article in a journal whose subfield is k ($k = 1, ...99$), and $F_{ik} = 0$ otherwise. The number of professors who published at least one article in subfield k can be defined as $N_k = \sum_i F_{ik}$. We can thus calculate the number of professors who have articles in both scientific subfields w and z as:

$C_{wz} = \sum_i F_{iw} F_{iz}$.

Counting joint occurrences of all possible pairs of subfields produces a square (99 x 99) symmetrical matrix of co-occurrence, whose generic cell contains C_{wz} as defined above. In turn, this matrix provides the input for the production of a symmetrical "similarity matrix" of the same size, whose generic cell contains S_{wz}, a similarity score between subfields w and z, based upon the cosine index measure we

---

[31] The ISI-WoS fields are: Biology, Biomedical Research, Chemistry, Clinical Medicine, Earth & Space, Engineering & Technology, Health Sciences, Mathematics, Physics, Professional Fields, Psychology and Social Sciences. Every field is then disaggregated into subfields, for a total of 99.

already employed to measure the similarity between patents and publications in a given PPP (section 3.2). In this case:

$$S_{wz} = \sum_{k=1}^{99} x_{kw} x_{kz} \left/ \sqrt{\sum_k x^2_{kw}} \sqrt{\sum_k x^2_{kz}} \right.$$

Swz ranges between 0 and 1 and is not sensitive to the absolute size of the scientific subfields. It is equal to one when the two subfields *w* and *z* identically co-occur with all other scientific subfields, while it is equal to zero for pairs of subfields that do not coincide.

We then used the similarity matrix as an input to a multidimensional scaling algorithm, which arranges the various fields on a 2-D plane by reducing the dimensionality of the data (Klavans and Boyack, 2006). The mapping of the subfields over the first two dimensions suggested the existence of four meta-fields: Biomedicine, Pharmaceuticals, Materials Chemistry and Engineering.

Once we reduced the 99 subfields to these 4 meta-fields, we reclassified the ISI-WoS journals accordingly. Finally, we compared the fields of the "reference" articles (the articles of the academic inventors) with those of the article signed by the co-authors and retained only those articles for which the fields corresponded.

Appendix 2 – Example of a patent-publication pair

PATENT EP1012301	*PUBLICATION* ISI:000074208600018
Title Total synthesis and functional overexpression of a $i(Candida rugosa) lip1 gene coding for a major industrial lipase 5 inventors	**Title** Design, total synthesis, and functional overexpression of the Candida rugosa lip1 gene coding for a major industrial lipase 5 authors
Abstract The dimorphic yeast Candida rugosa has an unusual codon usage which hampers the functional expression of genes derived from this yeast in a conventional heterologous host. Lipases produced by this yeast are extensively used in industrial bioconversions, but commercial lipase samples contain several different isoforms encoded by the lip gene family. In a first laborious attempt the lip1 gene, encoding the major isoform of the C. rugosa lipases (crls), was systematically modified by site-directed mutagenesis to gain functional expression in S. cerevisiae. As an alternative approach, the gene (1688 bp) was completely synthesised with an optimised nucleotide sequence in terms of heterologous expression in yeast and simplified genetic manipulation. The synthetic gene was functionally overexpressed in Pichia pastoris. The recombinant crl was produced at high level and purity, accounting for 90-95 per cent of the secreted proteins. The physical-chemical and catalytic properties of the recombinant lipase were compared with those of a commercial, non-recombinant C. rugosa lipase preparation.	**Abstract** The dimorphic yeast Candida rugosa has an unusual codon usage that hampers the functional expression of genes derived from this yeast in a conventional heterologous host. Commercial samples of C. rugosa lipase (crl) are widely used in industry, but contain several different isoforms encoded by the lip gene family, among which the isoform encoded by the gene lip1 is the most prominent. In a first laborious attempt, the lip1 gene was systematically modified by site-directed mutagenesis to gain functional expression in Saccharomyces cerevisiae. As an alternative approach, the gene (1647 bp) was completely synthesized with an optimized nucleotide sequence in terms of heterologous expression in yeast and simplified genetic manipulation. The synthetic gene was functionally expressed using both S. cerevisiae and Pichia pastoris as hosts, and the effect of heterologous leader sequences on expression and secretion was investigated. In particular, using P. pastoris cells, the synthetic gene was functionally overexpressed, allowing recombinant lip1 of high purity to be produced for the first time at a level of 150 U/mL culture medium. The physicochemical and catalytic properties of the recombinant lipase were compared with those of a commercial, non-recombinant C. rugosa lipase preparation containing lipase isoforms.

Appendix 3 – Correlation matrix for variables in regression exercise (std. errors in brackets)

	EXCLUSION	N_AUT	DELTA_YEARS	FIRST	LAST	SENIORITY	STOCK(t_1)	MOST_JUNIOR	MOST_SENIOR	TOP_SCHOLAR	BOTTOM_SCHOLAR
EXCLUSION	1										
N_AUT	0.1036*	1									
	(0.0000)										
DELTA_YEARS	-0.0844*	0.1463*	1								
	(0.0003)	(0.0000)									
FIRST	-0.1323*	-0.2113*	0.0009	1							
	(0.0000)	(0.0000)	(0.9697)								
LAST	-0.1183*	-0.1541*	-0.0310	-0.1749*	1						
	(0.0000)	(0.0000)	(0.1840)	(0.0000)							
SENIORITY	-0.1674*	-0.0316	-0.0969*	-0.0632*	0.1867*	1					
	(0.0000)	(0.1750)	(0.0000)	(0.0066)	(0.0000)						
STOCK(t_1)	-0.0613*	-0.0341	-0.0332	-0.0612*	0.2099*	0.6064*	1				
	(0.0085)	(0.1434)	(0.1544)	(0.0086)	(0.0000)	(0.0000)					
MOST_JUNIOR	0.1752*	-0.0617*	-0.0307	0.0200	-0.0958*	-0.6926*	-0.3655*	1			
	(0.0000)	(0.0081)	(0.1875)	(0.3912)	(0.0000)	(0.0000)	(0.0000)				
MOST_SENIOR	-0.0877*	-0.0820*	0.0092	-0.0909*	0.1874*	0.5698*	0.3218*	-0.2970*	1		
	(0.0002)	(0.0004)	(0.6917)	(0.0001)	(0.0000)	(0.0000)	(0.0000)	(0.0000)			
TOP_SCHOLAR	-0.0463*	-0.0840*	0.0131	-0.0675*	0.1855*	0.4031*	0.6326*	-0.2306*	0.4463*	1	
	(0.0468)	(0.0003)	(0.5733)	(0.0038)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)		
BOTTOM_SCHOLAR	0.1783*	-0.0074	0.0872*	0.0062	-0.0925*	-0.6372*	-0.3446*	0.8041*	-0.2603*	-0.2151*	1
	(0.0000)	(0.7520)	(0.0002)	(0.7910)	(0.0001)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	

Obs. 1842
* p<0.01