DOUBLE DISCLOSURES AND THE NEGOTIATION OF SCIENTIFIC CREDIT IN RESEARCH TEAMS

Francesco Lissoni[1, 4,*], Fabio Montobbio[2,4], Lorenzo Zirulia[3,4]

[1] GREThA UMR CNRS 5113 - Université de Bordeaux (France)
[2] Dipartimento di Economia e Statistics "Cognetti de Martiis" – Università di Torino (Italy)
[3] Dipartimento di Scienze Economiche – Università di Bologna (Italy)
[4] CRIOS – Università "L. Bocconi", Milano (Italy)

* Corresponding author: francesco.lissoni@u-bordeaux.fr

Incomplete draft, not to quote nor circulate (this version: 15/05/2016)

Abstract: We jointly examine the issues of research team formation and of the allocation of scientific credit to individual team members in a dynamic setting, with reference to "double disclosure" instances (the same research result is both published and patented). Senior and junior scientists decide whether to collaborate over an extended time horizon and bargain over the allocation of attribution rights (authorship and inventorship). Seniors make take-it-or-leave-offers, which juniors can either accept or sanction by exiting the team. Sustainable equilibria are found in which juniors trade inventorship for authorship, and opt to stay in the team. We test our theoretical predictions against an original dataset of "patent-publication pairs" produced by academics in seven European countries from 1997 to 2007. Younger and female authors are found to be more likely than older and male ones not to appear on patents, irrespective of the country and the technological field. First authors are more likely than middle authors to appear on patents, but when excluded they are less likely to quit the team, which we interpret as a sign of compliance with a successful negotiation outcome over attribution rights.

Keywords: economics of science; intellectual property; patent-publication pairs; scientific credit; authorship

JEL: Codes: O31, O34, L30

Acknowledgements: We gratefully acknowledge financial support from the Sloan Foundation's Research Program on the Economics of Knowledge Contribution and Distribution and the "Chaire d'Accueil" programme of the Regional Council of Aquitaine. Hospitality and seminars at UC Davis (through the "BIS-Californie" programme of the University of Bordeaux) and the Swinburne University of Technology (through the EU Centre on Shared Complex Challenges) allowed Francesco Lissoni to progress with the paper at critical times. Data for countries other than Italy and France were made available to us by Christian Stummer, Malwina Mejer, Catalina Martinez, Evangelos Bourelos, and Cornelia Lawson, under the terms of the APE-INV Research Network Programme, funded by the European Science Foundation. Gianluca Tarasconi provided outstanding technical assistance for data collection and data linkage. Maurizio Polenghi contributed to data checking and cleaning. All errors are ours.

1. Introduction

Once a classic topic in the sociology of science (Merton, 1968, 1988), scientific credit attribution has very recently come to the forefront of economic analysis (Bikard et al., 2015; Gans and Murray, 2014; Gans and Murray, 2013; Häussler and Sauermann, 2014, 2015; Lissoni et al., 2013). This resurgence of interest follows the recognition that scientific research, as measured by scientific publications, is increasingly conducted by teams, rather than individuals, and that such teams have been incessantly increasing in size as well as in geographical and organizational scope (Agrawal et al., 2013; Jones, 2009; Jones et al., 2008; Wuchty et al., 2007). At the same time, individual scientists still feed their careers with personal credit, including that for their role in research-related task such as grant-chasing and the commercialization of inventive results.

We contribute to this literature by focussing on « double disclosure » instances, which occur when scientific research results originate both a publication and a patent output, thus giving birth to a patent-publication pair (Murray and Stern, 2007), or a set of interrelated pairs. Solid evidence exists, which shows that authors of publications in the pairs generally outnumber inventors of related patents, so that some research team members are attributed authorship, but not inventorship. Lissoni et al. (2013) suggest that exclusion from inventorship may result from a bargaining process within the team, in which senior and male scientists (ideally, the chiefs of labs) make take-it-or-leave-it offers to junior and female ones. Evidence from a sample of Italian patent-publication pairs was shown to be compatible with the theoretical setting.

In this paper we extend Lissoni et al. (2013) in both a theoretical and an empirical direction. We develop a barganing model to a dynamic setting, in which junior inventors, if unhappy with the decisions, do not just have a voice option (litigating the senior's decisions), but an exit one (they can leave the team). In this way, we can examine how the distribution of attribution rights may affect if not team formation, at least team stability.

On the empirical side, we both explore the implications of our theory for the distribution of inventorship rights among team members and team stability, with data from Austria, France, Germany, Spain, Sweden, and the UK.

The results we obtain confirm, for all countries considered, the role of seniority and gender in the distribution of attribution rights, and suggests that exclusion from inventorship does not undermine team stability to the extent it is traded for first authorship.

In the reminder of the paper we first review the most recent literature on scientific credit distribution in teams and team formation (section 2). We then present our theoretical model (section 3) and our data (section 4). Section 5 contains our econometric analysis, section 6 concludes.

2. Background literature

Scientific credit is ultimately assigned to scientists working in team by third parties such as: the authors of follow-up publications, from whom citation rates depend; the funding agencies and their peer reviewers, to whom individuals submit their publication lists when grant-chasing; and perspective employers and consumers of consultancy services, who inspect the scientists' CVs and other references. To the extent that none of these actors can directly observe each team member's contribution, some heuristics or social norms have to exist that translates the collective achievement into a set of individual ones. A classic example in this sense is Merton's Matthew Effect, by which credit in a team is distributed according to the relative status of team members, with the most accomplished and senior scientists getting a disproportionate share of it compared to junior fellows (Azoulay et al., 2013; Jin et al., 2013; Simcoe and Waguespack, 2011). An example of social norm is fractional counting, by which each individual in an N-authored paper gets the same share of credit, where the latter is any inverse function of N (1/N in simple fractional counting). Official instances of use of such norm can be found, for example, in Italian public competitions for professorial jobs of the recent past (Checchi, 1999). Non-alphabetical name-ordering conventions, some of which are explicitly recommended by scientific journals, serve the same purpose, but go in the direction of stressing inequalities in each author's contribution (Engers et al., 1999; Rennie et al., 2000). In the same direction it goes the proposal, by an increasing number of journal editors, to dispose of authorship and replace it with "contributorship", in order to distribute credit according to the specific tasks performed by each team member (Biagioli et al., 1999; Frische, 2012).

The existing research on scientific credit attribution has kept distinct two different aspects of team-based research activity. First it investigates to what extent the prevailing social norms and heuristics shape the scientists' decisions to work in teams[1] (REF) and secondly investigates what consequences follow for the distribution of credit among team members, conditional on the prevailing heuristics and norms, as well as their strategic use by scientists (REF). This second stream of research assumes that some exogenous forces push scientists to work in teams. It is however important to ask how and why the distribution of scientific credit in teams affects the process of team formation.

The two most important drivers of teamwork in science are the increase in fixed costs (equipment, infrastructure) and in the quantity of knowledge scientists need to muster in order to make any progress in their research (de Solla Price, 1963; Galison and Hevly, 1992; Jones, 2009). In this context, strong evidence exists that multi-authored papers rank higher than single-authored ones according to several impact metrics, which makes collaboration attractive. Still, it remains to be seen whether all team members reap the benefits of the collaboration. Setting aside all issues of "guest" or "ghost" authorship (Lissoni and Montobbio, 2015; Mowatt et al., 2002), one wishes to know whether the prevailing heuristics and norms for credit attribution work in the direction of favouring a distribution or a concentration of credit, and of their welfare implications. Based on a large dataset of scientific publications, Jin et al. (2013) show that the Matthew effect not only bestows more credit to more reputed scientists, but also protect them from the

[1] Notice that some ambiguity permeates the entire literature, with reference to the concept of teams. While most authors abide nominally to Stephan's, 2012, authoritative description of teams as hierarchical entities – with chiefs of labs recruiting and disposing of their team members - the individual decisions to join a team or to admit a new member in his own team are often modelled as instances of collaboration among peers. We will come back to it later.

eventual blame for errors and omissions, which falls disproportionately on junior authors. This is coherent with the general finding that the rise of team science has gone hand in hand with a decline of concentration of research output at the departmental level, but an increase at the individual one (Agrawal et al., 2013).

Such asymmetries in credit distribution beg the question of what incentives are left to junior and less reputed scientists to collaborate or join teams. In a series of papers, some with co-authors, Joshua Gans and Fiona Murray assume away all positive returns to teamwork and collaboration, and model the emergence of co-authorship as the result of social norms distributing credit. Key assumptions concern the distribution of skills and roles among perspective team members, as well as the coordination costs and the degree of cumulativeness in the scientific enquiry.

Bikard et al. (2013) adapt a model of collaboration choices by Becker and Murphy (1992) to the case of an individual scientist who decides whether to allocate her time between writing a paper with N co-authors and one single-authored paper. Scientists draw their utilities from reputation, which take the form of citations received by their papers, through several possible function $s(N)$ (share of credit for an N-authored paper). For a given amount of time dedicated by the scientist to the single-authored and the multi-authored paper, the latter produces more total citations, increasing in N, but at the cost of increasing coordination costs. An empirical test follows, based on which the authors conclude that a sharing rule such as $s=1/\sqrt{N}$ is compatible with a rational allocation of time to collaboration. Notice that $1/\sqrt{N} > 1/N$, which implies that a collaboration premium exists. Based on the further assumption that the social value of each publication coincides with the total number of citations received, the authors conclude that collaboration comes at a cost for society, namely the loss of single-authored publications that would have generated the same individual credit, but with more research output. A policy recommendation follows, which is either to limit collaboration or to find way to affect the social norms on credit distribution (such as replacing $1/\sqrt{N}$ with $1/N$).

Gans and Murray (2013) consider a focal scientist facing a cumulative research project, to be undertaken in two steps, either alone or in collaboration, and with the possibility to module the publication output. The scientist can choose between three research strategies: (i) *integration* (he/she undertakes both steps alone and publishes one single-authored paper), (ii) *collaboration* (he/she collaborates with other scientists on both steps, and publishes one multi-authored paper), and (iii) *publication* (he/she completes the first step alone, publishes a single-authored paper with limited results of that step only, and get cited by other scientists' follow-up papers on the second step).

The publication strategy is the least expensive, as it does not entail any coordination costs nor any cost to acquire the necessary competencies to perform the second step. Which of the three strategies turns out to be privately optimal depends once again on the attribution norms, that is the functional form of $s(N)$. And once again it is found that collaboration premia induce over-collaboration.

Finally, Gans and Murray (2014) move further towards the description of teams not as mere instances of collaboration among peers, but as hierarchical organizations. Two scientists are considered, alternatively portrayed as a pioneer and a follower, or a senior and a junior (the latter working in the lab managed by the

former). The overall quality of the two-step research output depends on quality of research at each step, which is a probabilistic function of effort. Some specialization exists, to the extent that the pioneer is the only scientist who can undertake the first step, and only the follower can achieve top quality in the second step. Several collaboration schemes are considered, depending on whether the pioneer must commit to co-authorship from the start, or can postpone the final decision at the end of either the first step (after observing the outcome of his own effort) or the second one (after observing also the outcome of the follower's effort). Society can correctly assign an economic value to whatever research output the two scientists produce, but not to each scientist's contribution (asymmetric information). Thus, it applies some probability-based heuristics to estimate the relative contributions of the two scientists, which end up affecting the latter's efforts. The Pareto optimal equilibrium occurs under the regime in which the pioneer (senior) can dispose of co-authorship until the second step, which (quite intuitively) results in a maximum effort deployed the follower (junior) and in the largest expected value of research.

The literature on team formation and credit attribution rightly attracts our attention to how attribution norms may shape team formation, but it is quite limited in its description of the team's structure and activity. Even in Gans and Murray (2014), the senior and the junior scientist's look much less a chief-of-lab and his/her postdoc than two scientists with different experience (and possibly from different labs or universities) who have the possibility, and no obligation, to join efforts. The relationship exists only insofar the collaboration occurs, and the collaboration extends only to the joint paper.

As Stephan (2012) reminds us, the reality of research teams, to be intended as labs, is different. Contractual arrangements exist, which imply some sunk costs for the creation of team, and bind the team members for some time and to a multiplicity of deliverables. These may include a variety of outputs from research contracts (reports, proofs of concepts, prototypes…) as well as inventions stemming from fundamental research (Häussler and Sauermann, 2014). This broaden up the possibility to negotiate the distribution of scientific credit over and beyond the instance of a single paper. It also brings us back to Merton's original idea of authorship as symbolic intellectual property, which members of a team can dispose of according to economic considerations. (Merton stressed the cases of a senior scientist giving it up when wishing to increase the visibility of a junior colleague, or reclaiming it all for himself when a scientific prize was at stake).

Lissoni et al. (2013) consider the case of teams involved in the joint production of research results that are both amenable to be patented and published, as it is often the case in fields such as biotechnology or materials science (Fehder et al., 2014; Gans et al., 2013; Haeussler and Sauermann, 2013). Senior and junior scientists are bound by a long-term relationship and negotiate over the attribution of authorship (in particular, first authorship) and inventorship. The relative value of the two attribution rights differ across scientists. Authorship is relatively more valuable for juniors, as it is the currency they need to build their academic career. Seniors, whose career are established, attach instead greater relative value to inventorship and, possibly, getting exclusive control of the patent. Gender may combine with seniority, to the extent that female scientists may attach less value to inventorship, due to the well-documented obstacles they face when engaging in commercialization efforts (Ding et al., 2013; Thursby and Thursby, 2005; Whittington

and Smith-Doerr, 2005). Authorship and inventorship are distributed by senior scientists: they make take-it-or-leave-it offers to juniors, whose only alternative to acceptance is (costly) litigation.

It is then showed and tested that junior and female scientists may find it optimal to trade inventorship for authorship, irrespective of their effective contribution. This sends a false signal to the scientists' perspective customers, employers, and research partners, which entails a social cost. Still, the model falls short of investigating how the junior scientist may anticipate the senior's offer and decide in advance whether to join the team or not; or at least to exercise an exit option, alternative to litigation, which would allow us to observe (at least in principle) some instability in teams' composition as a function of internal negotiations of attribution rights. It is the second route that we take in this paper.

3. Negotiation over Authorship and Inventorship and Team Dynamics: A 2-Scientist Model

3.1 The model

We extend the model by Lissoni et al. (2013), to a dynamic setting. Time is discrete, with periods $t = 0, 1, \dots, \infty$.. We consider a team composed by a senior scientist (S) and a junior one (J), who engage in repeated collaborations, best described as a sequence of projects, one at each time t (in first instance, we also assume both scientists to be male). Each project, which requires the participation of S and J as necessary inputs, originates with certainty one paper and one patent. Besides, each project consists of two stages, namely:

1. the *team formation stage*, in which S and J decide to form the team, based on the expected allocation of attribution rights and the consequent payoffs
2. the *attribution rights allocation stage*, in which the team delivers its output and the attribution rights are definitely assigned.

S has full control over the allocation of attribution rights, that is he can decide who will be listed as author of the paper, and in which position, and who will be included as inventor in the patent. His decision cannot be litigated (for example, because this would prohibitively costly), so that the allocation of attribution rights that legal norms would prescribe according to the contribution of each scientist is in fact irrelevant.

In the first stage of project *S* proposes to J an allocation of attribution rights *A*, and *J* decides whether joining the team or not. We assume, however, that S cannot commit to the allocation A he proposes (that is, he can renege on it at the time of filing the patent application or submitting the paper for publication).

A is fully described by i) the identity of scientist who should be listed as first author; ii) the identity of the scientists who should be listed in the patent. For example *A*(*S*,*SJ*) stands for an allocation in which *S* is the first author, and both *S* and *J* are included in the patent (the order of the inventors on patents is irrelevant, as it conveys no information on individuals' contribution). We denote with R_1^S the value of publication for *S* if he is listed as first author, and R_{N1}^S if he is not. The corresponding values for *J* are R_1^J and R_{N1}^J. The common value from a patent is v/n, where $n = 1,2$ is the number of inventors listed in the patent.

In case J, based on *A*, does not join the team, both J and S get a reservation utility equal to 0. If the team is formed, then in the second stage the paper and the patent are produced, S take a final decision on the allocation of attribution rights, and the corresponding payoffs are assigned.

For *J*, the relationship with S implies a sunk cost I ($R_1^J > I$.) at the time when the team is formed, while no investment in the relationship is required for *S* [2]. *S* and *J* are characterized by an intertemporal utility function with discount factor δ^S and δ^J. δ^J assumes a value $\delta > 0$ with probability 1, while δ^S may assume two values, $\delta > 0$ with probability μ and 0 with probability $1 - \mu$. δ and μ are common knowledge. In other words, there are two types of senior scientists: the forward looking type (if $\delta^S = \delta$) , to whom we will refer as the $\delta - type$ and the myopic type or $0 - type$. In the first stage at t=0 S's type is not observed by *J*, which only knows the type distribution. Alternatively, δ may be interpreted as the objective probability that the relationship between *S* and *J* "survives" each period t, instead of being terminated for exogenous reasons (such as a shortage of funds). In this interpretation, *S* estimates correctly the value of δ with probability μ, and sets it mistakenly at 0 with probability $1 - \mu$,.

We assume that in a non-repeated interaction, the unique (subgame perfect) equilibrium is such that the team is not formed. This boils down to assume $R_{N1}^J < I$. In fact, once the team is formed (and the patent and paper are produced) *S* maximizes its payoff by assigning himself the first position in the paper and excluding *J* from the patent. In this case, *J* gets R_{N1}^J. By proceeding backwards, in the team formation stage *J* correctly predicts that its payoff if he joins the team will be $R_{N1}^J - I$, which is lower than the outside option. Without loss of generality, we shall assume from now on that $R_{N1}^J = 0$.

S maximizes its current payoff also in a repeated relationship when is a $0 - type$. Therefore, *J* would prefer not to form a team with *S* if μ is too small. However, when *S* is a $\delta - type$, different outcomes may be sustained as equilibria in the supergame, involving the formation of the team and different attribution rights.

First of all, we observe that since no investment is required by *S*, the latter always prefer to team up with *J* (independently from *S's* type). So, in each period *t*, the only relevant action for *S* is choosing the allocation of attribution rights. We denote each action with a pair of element: the first is the scientist who is first author, the second is the set of scientists included in the patent. For *J*, instead, the relevant action choice is to join the team or not. In the repeated game, *S*'s and *J*'s actions at time *t* are a function of game history.

As usual in repeated games, we will proceed by considering under which conditions on the parameters (expressed in terms of threshold for δ) a specific allocation of attribution rights is sustainable, in that a $\delta - type$ for S has no profitable deviation in the supergame. In a repeated interaction, *J* can threaten to leave the team if "exploited" (denied the attribution rights he was promised), which is bad for the $\delta - type$ in the long run. Such equilibrium may be interpreted both as an implicit contract between *S* and *J*, or a

[2] We justify this assumption as follows: J both scientists have learning costs when it comes to start a new project, by J's are much larger than S's, which we set at zero.

social norm prevailing in the scientific community. We focus on three possible allocations of attribution rights, to which we refer as *norms*, which S may be willing to sustain as part of the equilibrium in the supergame:[3]

- *J* is first author, but he is excluded from the patent (norm 1)
- *J* is first author, and he is included in the patent (norm 2)
- *S* is first author, and J is included in the patent (norm 3)

We shall assume that both S and *J* play trigger strategies, i.e. they switch forever to the unique equilibrium of the non-repeated interaction if a deviation is observed. The following Proposition summarizes the main result (the proof is the Appendix C).

Proposition 1 *(Norm 1) Suppose that the players' strategies are such that i) the team is formed at t=0; ii) if S is* $\delta - type$*, he goes for A(J,S) in each period in which the team operates, if in all previous periods the team operated with allocation A(J, S), and A(S,S) otherwise. iii) J joins the team a t=0; iv) for each t>0, J stays in the team if S plays A(J,S) and quits forever otherwise. This is an equilibrium in the supergame if:*

$$\delta \geq max\left\{\frac{I-\mu R_1^J}{I-\mu I}, \frac{R_1^S-R_{N1}^S}{R_1^S+v}\right\}.$$

(Norm 2) Suppose that the players' strategies are such that i) the team is formed at t=0; ii) if S is $\delta - type$*, he plays A(J, SJ) in each period in which the team operates, if in all previous periods the team operated under A(J, SJ), and A(S,S) otherwise. iii) J joins the team a t=0; iv) for each t>0, J stays in the team if S plays A(J,SJ) and quits forever otherwise. This is an equilibrium in the supergame if:*

$$\delta \geq max\left\{\frac{I-\mu(R_1^J+\frac{v}{2})}{I-\mu I}, \frac{R_1^S-R_{N1}^S+\frac{v}{2}}{R_1^S+v}\right\}$$

(Norm 3) Suppose that the players' strategies are such that i) the team is formed at t=0; ii) if S is $\delta - type$*, he plays A(S,SJ) in each period in which the team operates if in all previous periods the team operated under A(S, SJ), and A(S,S) otherwise. iii) J joins the team a t=0; iv) for each t>0, J stays in the team if S plays A(S,SJ) and quits forever otherwise. This is an equilibrium in the supergame if:*

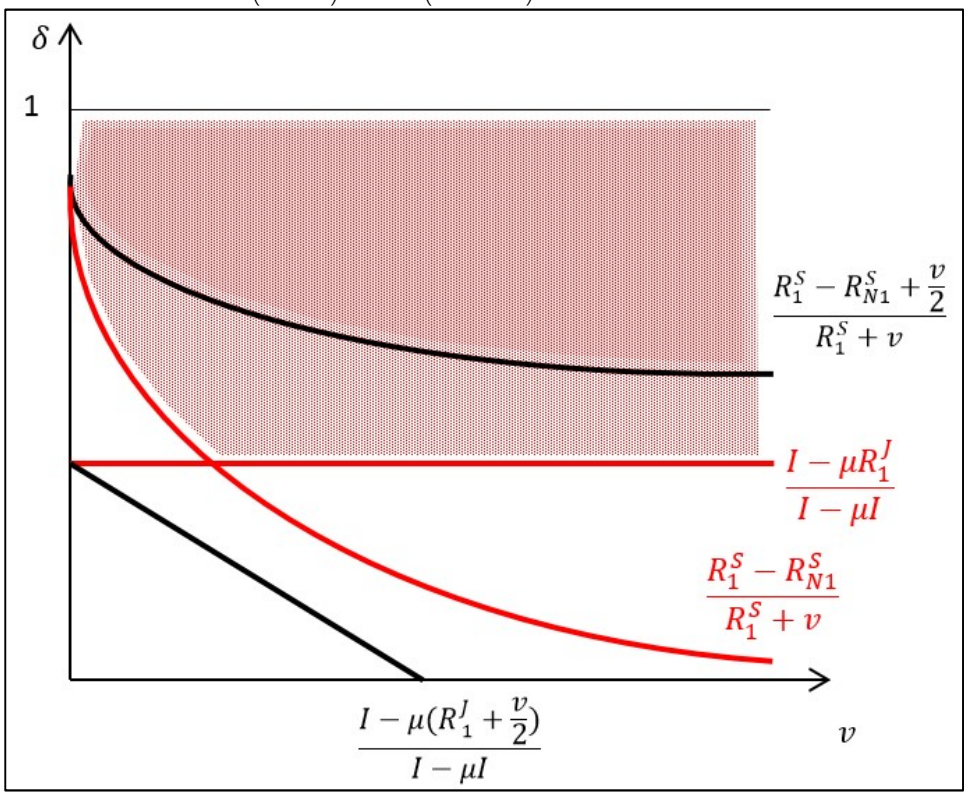
$$\delta \geq max\left\{\frac{I-\mu\frac{v}{2}}{I-\mu I}, \frac{\frac{v}{2}}{R_1^S+v}\right\}$$

As it is usual the case for repeated games, the model allows for multiple equilibria, i.e. there are parameter regions in which more than one norm is sustainable as equilibrium. However, a graphical comparison of norm 1 and norm 2 (see respectively the red and black lines in Figure 1) show that when norm 2 is an equilibrium, norm 1 is an equilibrium, too. Since norm 1 yields a higher payoff to *S* than norm 2, we expect

[3] We do not consider here norms in which S is excluded from the patent. That is, he is always listed as an inventor

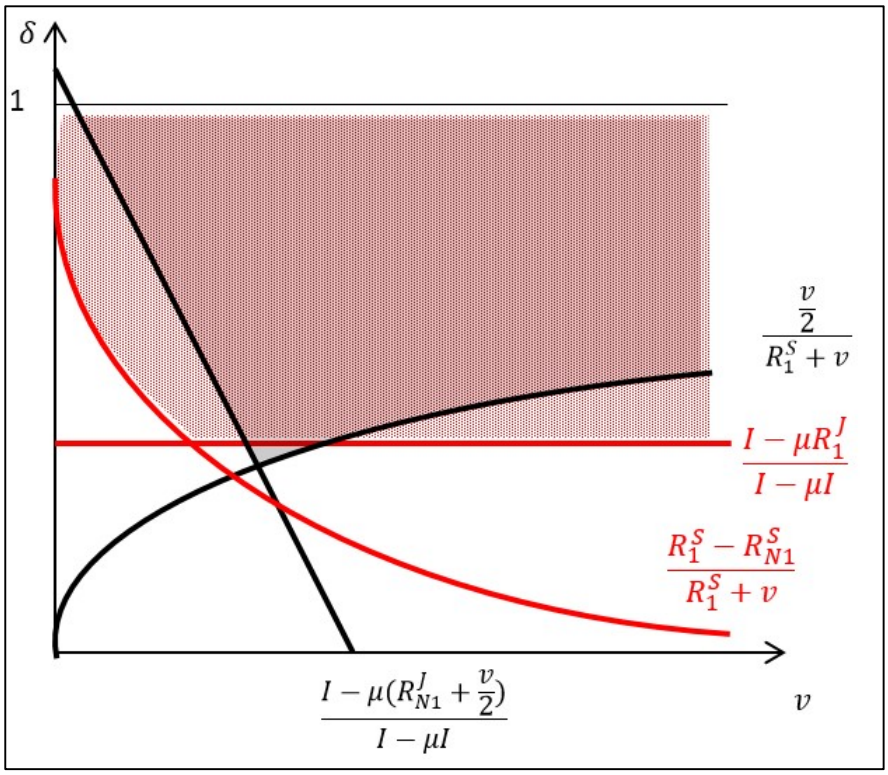
norm 1 to prevail over norm 2, in which case the exclusion of *J* from the patent does not lead him to leave the team as long as he is the first author in the paper.

Figure 1. Allocation of attribution rights: equilibrium conditions for Norms 1 (in red) **and 2** (in black)



Still, there can be (small) parameter regions in which norm 3 is the only possible equilibrium (grey triangle in Figure 2,). In this case, *J* is included in the patent, but loses the first position in the paper in favour of *S*.

Figure 2. Allocation of attribution rights: equilibrium Norms 1 (in red) **and 3** (in black)



For any given value of δ and $\frac{I-\mu R_1^J}{I-\mu I}$, the equilibrium conditions for Norm 1 are more likely to hold the larger the value v of the patent and the smaller S's marginal value of first authorship ($R_1^S - R_{N1}^S$), so that S is keen on trading the latter for exclusive inventorship. This may occur for very valuable patents and/or when S is a very senior and highly reputed scientist, whose marginal value of first authorship is pretty small. For v large and/or ($R_1^S - R_{N1}^S$) small, the equilibrium is more likely to hold the lower $\frac{I-\mu R_1^J}{I-\mu I}$, that is the larger R_1^J (J is very junior and attaches great value to first authorship due to career concerns), and the lower I (J has not invested much in the partnership, so he may content himself with just first authorship).

3.2 Intuitive extensions and testable propositions

Our 2-scientist model does not lend itself to produce immediately testable propositions. The main reason for that is that most scientific teams in our fields of interest include more than 2 scientists, each of which contributes differently, both quantitatively and qualitatively, to the research effort.

In these cases, several scientists may find themselves in the J's position, namely as juniors who negotiate attribution rights with S (or a set of seniors, jointly running the lab). At the same time, the literature suggests that a division of labour may exist between team members, so that not all members contribute equally to the research effort and may reasonably claim first authorship. At an extreme, the contribution can be so weak that legal norms clearly prevent inventorship to be granted (see the discussion in Lissoni et al., 2013). Under these circumstances, several J scientists may be either ready to trade inventorship with

just middle (neither first nor last) authorship, or simply to reclaim "guest" or "gift" authorship from S, on the basis of circumstances external to the specific research project from which the publication draws.

In this case, the equilibrium solution for our model suggests that any form of authorship will lead to the scientist's decision to stay in the team. At the same time, though, if the team host several team members whose research effort is such that they all rightfully compete for first authorship (that is, in J's position), those who do not get it may decide to quit if not compensated for the loss with inventorship.

Therefore middle-authorship can both predict a higher probability of getting inventorship (as compensation for the loss of first authorship) or a lower one (the contribution to the research effort being at most sufficient for getting authorship, but not inventorship). Early evidence by Lissoni et al. (2013) suggest that the second case is the relevant one.

Also from Lissoni et al (2013) we derive the suggestion that our model can be either extended to or reinforced in the case of J being a female scientists. As suggested by the literature, female scientists face both more difficulties in their academic careers (which makes first authorship particularly valuable) and in commercializing their research results (which lowers the subjective value of the patent). They can therefore be readier than men to trade inventorship for first authorship, and stay in the team if the bargain goes through.

Based on our 2-agent model and this discussion, we can put forward the following testable propositions:

Proposition 1. Junior (younger, lower-ranked) and female authors of a patent-related publication have higher probability not to appear on the patent than more senior and male ones, controlling for their position in the author by-line.

Proposition 2. First authors' probability to stay in the team (to publish again with the senior co-authors) is unaffected by their inclusion/exclusion in the patent's inventor list

Proposition 3. Middle authors who have been excluded from the patent have a lower probability to stay in team, conditional on their contribution to the research effort being the same as the first author's

We run two sets of regressions. In the first set we examine the probability of the author of a scientific publication to be excluded from the related patent(s), in order to test Proposition 1. We will refer to it as the "exclusion" regression set.

In the second set of regression we examine the probability of the author of a patent-related scientific publication to publish again with one or more members of the same patent-related research team (namely, the co-authors of the specific publication or of other publications connected to the same patent). We test propositions 2 and 3. We will refer to this as the "renewed co-authorship" regression set.

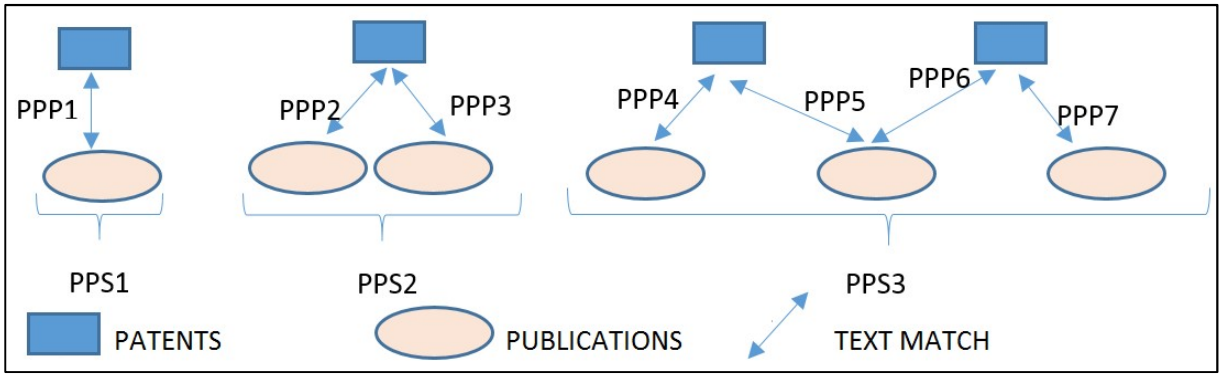
4. Data

4.1 Methodology and contents

Our dataset results from a fairly complex data-mining methodology, which we illustrate at length in Appendix 1. Here we provide a summary of the key issues and descriptive statistics.

The basic units of analysis in our regressions are authors of papers included in patent-publication "sets" (PPSs). Each set groups from one to several related patent-publications "pairs" (PPPs). PPPs are the empirical equivalent of "double disclosures" in our theoretical model. They consist of 1-to-1 patent-publication matches, whose lists of inventors and authors share at least one surname-and-initial and, according to text analysis of titles and abstracts, are likely to deal with the same research result. PPSs that include more than one PPP can be of a 1-to-N type (one patent and several publications) as well as of N-to-1 or N-to-N type, as illustrated in Figure 3

Figure 3. Patent-publication pairs (PPPs) and patent-publication sets (PPSs): definition



Our main source for patent data is the APE-INV programme (http://www.academicpatenting.eu), whose database contains information on academic patents filed at the European Patent Office (EPO), from several countries and priority dates comprised between 1997 and 2007. By academic patent we mean a patent application (whether granted or not) concerning an invention by one or more university-affiliated scientists, regardless of whether the applicant is a university, a business company, or an individual. The APE-INV data for each country were collected with a similar methodology, based on matching names of inventors and scientists, but on the basis of rather heterogeneous data sources on scientists (Lissoni, 2013). As a result, our sample cannot be considered representative at the cross-country level, due especially to under-representation of Belgium, France, and the UK, and over-representation of Italy and Spain.

As for publications, we searched the Web of Science (WoS©, by Thomson Reuters) for authors with the same name and initials of the academic inventors, within a time range of two years before and two years after the patent's priority year. We then allocated all the resulting patent-publication matches to seven technological fields. For each field, we conducted a separate text analysis of all matches, based on a field-specific vocabulary. This produced a "similarity index" equal to the cosine distance between the vectors of

words contained in the titles and abstracts of the matched patents and publications, as in Lissoni et al. (2013)[4].

, We had access only to the publications dated from 1999 onward. This returned over 4.6 million matches with non-zero values of the similarity index, based on ~140000 publications and ~9000 patents. Out of these, we first retained for further analysis only the matches with similarity index greater than 0.2, which is less than half the observations in the first percentile of the frequency distribution of the index values. In order to minimize the probability to incur in false positives (see discussion in Appendix 1) we further restricted our analysis to the top 10% of the first percentile, for a total of 952 PPSs (3p90 sample). One further restriction (to the top 5% of the first percentile) reduces the PPSs to 561 (2p95 sample), which we use to conduct robustness tests. Around half the PPSs, whether in the 2P95 or 3P90 sample, is of the 1-to-1 type; the rest is predominantly of the 1-to-N type, and hardly include more than 5 publications in the most restrictive sample (2p95) and 10 publications in the second most restrictive one (3P90), and almost never more than 2 patents (see table 1). These statistics reassure us about the selectivity of our algorithm when it matches publications to patents, and vice versa.

Table 1. PPS frequency distribution, by nr of patents and publications in the PPS

	By nr of patents		**By nr of publications**	
	2p95	*3p90*	*pps_2p95*	*pps_3p90*
1	483	813	363	524
2	62	109	96	187
3	13	23	44	96
4		2	24	48
5	1	2	10	30
6	1		6	19
7			4	16
8			3	8
9		1		2
10			3	1
11-20	1	2	5	16
21-50			3	3
51-100				2
Tot patents/publications	668	1154	1142	2299

3p90: PPSs based on PPPs with similarity index values in the top 10% of the first percentile (952 obs)
2p95: PPSs based on PPPs with similarity index values in the top 5% of the first percentile (561 obs)

Another useful information on the quality of our data can be obtained by calculating the difference between the number of inventors and the number of authors in each PPSs (an information we use in our regression exercises). In the case of 1-to-1 PPSs the difference is simply the difference between the number of inventors on the only patent and the number of authors on the only publication in the pair. In the case of 1-to-N, N-to-1, and N-to-N PPSs the difference is that between the total number of distinct inventors of the

[4] When a patent was classified under more than one technological fields, we computed more than one similarity index for each one of its publication matches, and retained the maximum value.

patents and the total number of distinct authors of the publications in the same PPS. Negative values indicate that we have more authors than inventors. As expected, they prevail (see table 2). Very high absolute values, which refer to 1-to-N or N-to-N cases in which publications list many authors each, are rare: around 2% with ≥50, less than 20% with ≥10.

Table 2. - % distribution of PPSs, by value of the difference between number of inventors and authors

Nr inventors – nr authors	PPS: 2p95	PPS: 3p90
-700,-100	0.7	0.7
-99,-50	1.8	1.3
-49,-10	17.1	20.7
-9,-5	20.7	21.1
-4	6.2	6.4
-3	7.1	8.0
-2	11.4	8.9
-1	9.3	9.3
0	11.6	9.9
+1	4.6	5.0
+2	4.3	3.5
+3	1.8	1.6
+4	1.2	1.3
+5,+9	1.8	1.8
+10,+49	0.4	0.5

When examining the technological field of the patents, negative values clearly prevail in three out of four science-based fields: "Instruments", "Chemicals & Materials", and "Pharma & Biotech" (with the latter exhibiting the largest differences, as also found by Fehder et al., 2014). This is line with our expectation to find more instances of "double disclosures" in science-based technologies, where the inventive activity of academic scientists may be a straightforward consequence of their research activity and our methodology is less likely to produce false positives. While in the other fields, many academic inventions stem out of targeted applied research or consultancy, or even extra-academic activities, which are less likely to be captured by publications.

Finally, in order to measure the authors' seniority, we went back to the WoS and collected all the publications by authors with the same surname and initials of the individuals included in the 3p90 PPS sample. We limited our search to journals relevant for the technological fields of the patents in the PPS and to authors with not exceedingly common surnames (such as Smith, Muller, or Park). We then calculated the year of the authors' first publication and their stock of publication in each following year.

Information on gender could not be retrieved by any automatic mean. Except for the authors with at least one patent (whose documentation reports full names), we had access only to name initials, as derived from WoS. This forced us to substantial manual work, which in turn was limited by time and budget constraints. Wherever possible, we downloaded the front page of each author's papers and collected their first names from there, if available. We then matched it to those included in the IBM GNR's library, and retrieved and

elaborated the associated information on gender. Where possible, ambiguous cases were solved manually, by inspecting the author's affiliation and its country location or, for prolific authors, their webpage.[5]

We thus managed to retrieve gender information for 6242 authors, of which 33% turned out to be female. Their distribution across technologies and discipline is rather uneven, with a larger presence in Pharma & Biotech, Chemicals and Materials, and Scientific Instruments.

4.2 Exclusion regressions: Variables and descriptive statistics

For our main regressions, we consider all PPSs from the 3p90 class (952 units, for a total of 1154 patents and 2299 publications). Our observations are author-publication couples. That is, each publication in a PPS generate as many observations as the number of its authors, for a total of >19000. We then exclude from the sample all the publications whose number of authors is smaller than the total number of inventors in the PPS, as well those whose authors are listed alphabetically (in which case we consider the order of the authors in the by line to be uninformative). We also exclude a certain number of authors with extremely common surnames, for whom the information on the stock of publication and seniority would certainly be unreliable. This leaves with little more than 14000 observations (xxxx patents and 1820 publications).

The binary dependent variable *exclusion* takes value one when the author of the publication does not appear as an inventor on any related patent.

The main explanatory variables refer either to the characteristics of the author, the publication on which s/he appears, or the patent from which s/he is (or is not) excluded, as follows:

- The position of the author in the publication's by-line, namely: *First*, *Middle*, , *Last*; where *Middle* indicates any position in between the first and the last author, and is the reference case.
- *Female*, which indicates the author's gender. In several specifications we interact it with the author's position in the by-line, with male middle authors (*Male Middle*) as the reference case
- The author's seniority, relative to the co-authors on the same publication, at the time of the publication. We measure it with two sets of variables. The first set considers the stock of past publications, with two dummy variables: *top_scholar* =1 for the author in the co-authorship team with the largest stock of past publications, and *bottom_scholar* =1 for the author with the smallest stock of past publications (all intermediate cases as references). Alternatively, we consider one continuous variable (*relative_scholarship*), which ranges from 0 (for the bottom scholar) to 1 (for the top scholar)[6].

[5]The *IBM Global Name Recognition* (IBM-GNR) system is a commercial product that performs various tasks, including the association of first names to gender, expressed as the probability *p* that the name is masculine (*1-p* feminine). Ambiguous cases are those in which a name's gender varies by country (e.g. Andrea, which is masculine in Italy and feminine in German and English-speaking countries) or the name is epicene (is both feminine and masculine in the same country, such as Dominique and Yannick in France). More details in Appendix 1

[6] In detail:

$$relative\ scholarship_{ij} = \frac{stock_{ij} - \min(stock_j)}{\max(stock_j) - \min(stock_j)}$$

The second set considers each co-author's date of first publication, with *most_senior*=1 for the co-author whose first publication is the oldest, and *most_junior*=1 for the one with the most recent (all intermediate cases as references). Alternatively, we consider one continuous variable (*relative_seniority*), which ranges from 0 (for the bottom scholar) to 1 (for the top scholar)[7].

The two measures capture different aspects of seniority: the former measures experience in publishing, and may be relatively high for a young researcher with an intense publishing activity; the latter is a better proxy for age , as with an aged co-author with few publications (for example a laboratory technician occasionally rewarded with authorship).

- *Proximity*, which is the inverse cosine distance measures between the patents and the publications in each PPS (with >1 patents or publications we consider the maximum distance)
- The time distance between the publication and the related patent, expressed as the publication year minus the patent's priority year (with >1 patents in the PPS we consider the maximum distance). We organize it in five dummies (-2 years, -1 year, 0 years, +1 year, +2 years), where negative values point at publications preceding the patents, and -2 years is the reference cases.
- The number of authors in the publication (*n_authors*) and the number of inventors of the related patent (*n_inventors*; with >1 patents in the PPS we consider the total number of distinct inventors in the patent set)
- The technological field, according to the following dummies: *Electrical engineering & Electronics*, *Instruments*, *Chemicals & Materials*, *Pharmaceuticals & Biotechnologies*, *Other technologies* (with *Electrical engineering & Electronics* as the reference case)
- The country of the academic inventor in the patent, with Austria as the reference case

Notice that while the same author enters several observations (as many as the publications on which she is listed), some of his/her characteristics may change across observations. In particular, the publication stock will increase when moving from less to more recent publications, and seniority may vary with both the date of the publication and, in relative terms, with respect to the seniority of the co-authors.

Based on our model, we can put forward a number of a priori on the sign of our explanatory variables of interest. In particular, we expect junior and female authors to have a higher probability of exclusion from the patents related to their publications.

As for our main controls, first and last authorship, they ought to be negatively correlated with the probability of exclusion, with the last author possibly exhibiting a lower exclusion probability than the first one.

---

where $stock_{ij}$ is the number of publications by co-author i at the time of paper j and $min(stock_j)$ [$max(stock_j)$] is the minimum (maximum) value for the same variable among those of all co-authors of paper j.

[7] In detail:

$$relative\ seniority_{ij} = 1 - \frac{first\ year_{ij} - \min(first\ year_j)}{\max(first\ year_j) - \min(first\ year_j)}$$

where $first\ year_{ij}$ is the year in which co-author i of paper j first published an article recorded by the Web of Science and $min(first\ year_j)$ [$max(first\ year_j)$] is the minimum (maximum) value for the same variable among those of all co-authors of paper j.

Other controls with expected negative coefficients are:

- the proximity between the text (contents) of the matched patent and publication: the closer the match, the least likely it is to be a false one, one in which the patent and the publication are indeed unrelated and the exclusion probability high
- the time distance between the patent and the publication: patent-publication matches in which the publication that follows the patent (time distance is positive) are less likely to be false ones, which lowers the probability of exclusion
- the number of inventors on the patent: the larger their number, the less likely it is that some authors of the related publications will be excluded (conversely, the larger the number of authors, the higher the probability of exclusion, due to dilution of the average author's contribution to the invention)

Table 3 reports the summary statistics. Notice the high number of missing observations for gender-related variables, and the dominance of observations related to *Pharmaceuticals & Biotechnologies*. This is due to:

- the disproportionate number of PPPs we observe in this field, as opposed to other science-based field such as *Electrical Engineering & Electronics* and *Chemicals & Materials*, as discussed in section 4.1 and appendix A
- the higher average number of authors per paper in the related disciplines.

Table 3. Exclusion regression: summary statistics

	Obs	Mean	Std. Dev.	Min	Max
exclusion	14261	0.68	0.47	0	1
Middle	14261	0.74	0.44	0	1
First	14261	0.13	0.34	0	1
Last	14261	0.13	0.34	0	1
Female	9148	0.35	0.48	0	1
Male Middle	*9148*	*0.45*	*0.50*	*0*	*1*
Female Middle	*9148*	*0.27*	*0.44*	*0*	*1*
Male first	*9148*	*0.08*	*0.27*	*0*	*1*
Female first	*9148*	*0.06*	*0.23*	*0*	*1*
Male last	*9148*	*0.12*	*0.32*	*0*	*1*
Female last	*9148*	*0.02*	*0.14*	*0*	*1*
relative_scholarship	14244	0.26	0.35	0	1
relative_seniority	14247	0.38	0.36	0	1
top_scholar	14261	0.13	0.34	0	1
bottom_scholar	14261	0.20	0.40	0	1
most_senior	14261	0.14	0.34	0	1
most_junior	14261	0.22	0.42	0	1
proximity	14261	0.38	0.09	0.28	0.80
-2 years	14261	0.11	0.32	0	1
-1 year	14261	0.13	0.34	0	1
0 years	14261	0.19	0.39	0	1
+1 year	14261	0.28	0.45	0	1
+2 years	14261	0.28	0.45	0	1
n_authors	14261	9.60	5.44	2	55
n_inventors	14261	4.01	2.82	1	40
Electrical eng; Electronics	14261	0.08	0.27	0	1
Instruments	14261	0.19	0.39	0	1
Chemicals; Materials	14261	0.07	0.26	0	1
Pharmaceuticals; Biotech	14261	0.63	0.48	0	1

Other technologies	14261	0.03	0.16	0	1
Austria	14261	0.08	0.28	0	1
Belgium	14261	0.05	0.23	0	1
Spain	14261	0.14	0.34	0	1
France	14261	0.12	0.32	0	1
Italy	14261	0.55	0.50	0	1
Sweden	14261	0.02	0.15	0	1
UK	14261	0.04	0.18	0	1

4.3 Renewed co-authorship regressions: variables and descriptive statistics

For our main regressions, we start from thee PPSs in the 3p90 class. More precisely, we consider all authors listed on publications included in such PPSs, with the exception of those with very common names and those with no further publications, for a total of 10149 observations (authors times PPSs, as a few authors appear in more than one PPS). For such authors we collect all publications from the PPS's date onward (future publications, for short).

Our dependent variable is the author's probability to stay in the PPS team. For each author, we then consider whether s/he publishes again with at least one contributor to the same PPS (PPS members, for short), namely a co-author of his/her publication(s) or the author of another publication in the same PPS. In this way, we try to capture all publications a research team may have produced in relationship to one or more patents, albeit possibly with different authors' configurations; and the relationships of the individual author with the entire team, and not just the specific co-authors of one or another publication.

As shown in table 4, almost all authors we retain for our regressions write at least one more paper with their previous PPS members (over 99%). This is possibly due to our definition of PPS, which is as conservative as possible and leaves out a few papers still related to the PPS. However, if we move from 1 to 2 or more papers the relevant figure drops dramatically to around 14%, We then consider as our dependent variable a binary one, which takes value one if the author writes at least two papers with one or more previous PPS members, and zero otherwise.

Table 4. Nr of renewed co-authorships per author, and % of first/last authorship – frequency distribution

Nr of co-authorships	Nr of authors	% of authors
none	9	0.09
1	8,722	85.94
>1	1,418	13.97
Total	*10,149*	*100*

The two explanatory variables of our interest concern the attribution rights s/he obtained by contributing to the PPSs of which s/he was a member, namely: his/her position in the PPS papers in which s/he appear as an author and the inventorship (inclusion in one or more patents).

As shown in table 5, most of the authors we consider have just one paper in the relevant PPS (around 74%) while very few have more than two (around 8%; with a maximum value of 25). As consequence, they most often enter our dataset with just one position in the author by-line, namely the one they got in the only paper they contributed to the PPS. When examining first authorship, around 84% of the authors never get it, 13.4% always get it (most often because they are first author of a single paper), and 3% get in on one or more, but not all their publications. The statistics for last authorship are very similar, while those for middle authorship are slightly more varied (respectively: 28%, 67%, and 5%). For ease of treatment, we consider the authors' positions in all the papers in the same PPS with three simple dummies: *First* (if they have at least one paper as first author), *Last* (if they have at least one paper as last authors, and none as first, which is almost always the case), and *Middle* (if they never get first or last authorship, which will be our reference case).

In addition, we also consider the number of papers each author has contributed to the PPS, which is at the same time an indicator of the seniority of the author within the PPS, and of the stability of his/her collaboration with the PPS team (*Nr of authorships in PPS*).

As for inventorship, we measure it with a dummy that takes value 1 if the author has never been listed as inventor on the patents included in the PPS (*Exclusion*).

An additional variable of interest is gender (*Female*), although its inclusion comes at the cost of losing about half the observations.

Table 5. Nr of PPS papers per author, and % of first/last authorship – frequency distribution

Nr papers per author	Nr of authors	% of authors
1	7,490	73.8
2	1,858	18.3
>2	801	7.9
% papers as last author	Nr of authors	% of authors
0	8,484	83.6
(0,100)	304	3.0
100	1,361	13.4
% papers as last author	Nr of authors	% of authors
0	8,371	82.5
(0,100)	320	3.2
100	1,458	14.4
% papers as middle author	Nr of authors	% of authors
0	2,868	28.3
(0,100)	498	4.9
100	6,783	66.8

Tot	*10,149*	*100*

Authors who appear in more than one PPS are counted more than once

We further control for :

- The author's seniority, as measured by the year of his/her first publication (*First publication year*) and the number of his/her publications before entering the PPS team (that is, at the time of the first publication included in the PPS; *Publication Stock*)
- *Proximity*, which is measured as in the exlcusion regressions and controls for the quality of our publication-patent matching
- The total number of authors in the PPS (*N_authors in PPS*) and the total number of inventors of the related patents (*N_inventors in PPS*)
- The year of the most recent publication included in the PPS (*Last year in PPS*), which enter the regressions as a set of calendar dummies

Table 6 reports the summary statistics.

Table 6. Renewed co-authorship regression: summary statistics

	Obs	Mean	Std. Dev.	Min	Max
Position:					
Medium	10149	0.67	0.471	0	1
First	10149	0.17	0.374	0	1
Last	10149	0.16	0.370	0	1
Exclusion	10149	0.69	0.463	0	1
Nr authorships in PPS	10149	1.43	1.107	1	25
Female	5886	0.32	0.467	0	1
First publication year	10021	1992.81	11.986	1950	2014
Publication stock	10021	26.59	57.457	0	1318
Proximity	10149	0.42	0.102	0.28	0.93
N_authors in PPS	10149	20.50	17.455	2	80
N_inventors in PPS	10133	4.68	3.352	1	17
Last year in PPS	10149	2004.87	2.454	1999	2009

5. Results

5.1 Probability of Exclusion from Inventorship

Table 7 reports the results of various specifications of the exclusion regressions, all estimated with OLS (Linear Probability Model - LPM). Though our dependent variable is binary, we followed a recent trend

and went for LPM for three reasons. First, it makes it easy to interpret estimated coefficients as changes in probabilities. Second, when it comes to interaction terms, they lend themselves to a straightforward interpretation as cross-partial derivatives. Third, when most regressors are discrete variables, as it is our case, it provides a reasonable linear approximation of non-linear marginal effects (Angrist and Pischke, 2009; ch. 3.4.2; Wooldridge, 2010; ch. 15.2). Still, we take care to validate our results by means of Logit estimates, which will also help us to further explore non-linearities.

Overall, the results confirm previous findings by Lissoni et al. (2013) and extend them to countries other than Italy. The number of observations drops when moving from column (1) to column (2) and (3) due to the inclusion of gender among regression, for which we have many missing observations. As for columns (4) and (5) they reports results for specifications as in (3), but in absence of observations for, respectively, Pharmaceuticals & Biotechnology and Italy, namely the most represented technology and country in out dataset.

Column (1) shows confirms a number of our *a priori*, namely the role of seniority, as well as the validity of our controls. *Ceteris paribus*, the co-author with the strongest publication record (*relative_scholarship*=1) has around 40% less probability to be excluded from the patent than the one with the weakest record (*relative_scholarship*=0). On top of this, the most junior co-authors have 5% more probability to be excluded from the patent.

The size of these marginal effects is larger than the marginal effect of coefficients related to the authors' position in the paper by-line, which are around -21% for last authors and -14% for first authors. Notice that, as expected, the probability of exclusion for last authors is lower than that for first ones.

Table 7. Exclusion regression: OLS estimates (3p90 PPS class)

	(1)	(2)	(3)	(4)	(5)
First	-0.14***	-0.13***			
	(0.028)	(0.031)			
Last	-0.21***	-0.19***			
	(0.025)	(0.026)			
Female		0.06***			
		(0.016)			
Female Middle			0.06***	0.04	0.07***
			(0.014)	(0.027)	(0.025)
Male first			-0.16***	-0.17***	-0.21***
			(0.038)	(0.047)	(0.038)
Female first			-0.04	0.01	-0.05
			(0.031)	(0.053)	(0.044)
Male last			-0.19***	-0.14***	-0.23***
			(0.029)	(0.036)	(0.031)
Female last			-0.16***	-0.17***	-0.16**
			(0.041)	(0.059)	(0.061)
relative_scholarship	-0.41***	-0.40***	-0.40***	-0.36***	-0.32***
	(0.036)	(0.043)	(0.043)	(0.053)	(0.039)
most_junior	0.05***	0.05***	0.05***	0.03	0.06***
	(0.014)	(0.017)	(0.017)	(0.026)	(0.022)
proximity	-0.18***	-0.17**	-0.17**	-0.17	-0.11
	(0.058)	(0.078)	(0.078)	(0.138)	(0.119)
-1 year	0.01	0.02	0.02	0.02	-0.04*
	(0.019)	(0.020)	(0.020)	(0.028)	(0.024)
0 years	-0.04**	-0.02	-0.02	0.00	-0.11***
	(0.021)	(0.030)	(0.030)	(0.037)	(0.028)

+1 year	-0.05*	-0.03	-0.03	-0.03	-0.10***
	(0.028)	(0.031)	(0.031)	(0.037)	(0.023)
+2 years	-0.06*	-0.06**	-0.06**	-0.08**	-0.14***
	(0.030)	(0.031)	(0.031)	(0.039)	(0.027)
n_authors	0.01***	0.01***	0.01***	0.02***	0.02***
	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)
n_inventors	-0.03***	-0.03***	-0.03***	-0.04***	-0.03***
	(0.004)	(0.005)	(0.005)	(0.010)	(0.004)
Instruments	0.09**	0.15***	0.15***	0.13***	0.17**
	(0.041)	(0.055)	(0.054)	(0.049)	(0.071)
Chemicals; Materials	0.12***	0.20***	0.20***	0.21***	0.22***
	(0.044)	(0.056)	(0.056)	(0.052)	(0.072)
Pharmaceuticals; Biotech	0.13***	0.20***	0.20***		0.17**
	(0.040)	(0.054)	(0.054)		(0.071)
Other technologies	0.07	0.13*	0.13*	0.13**	0.05
	(0.054)	(0.070)	(0.070)	(0.067)	(0.086)
Belgium	-0.04	-0.00	-0.00	-0.03	0.02
	(0.023)	(0.034)	(0.034)	(0.066)	(0.034)
Spain	-0.07***	-0.01	-0.01	-0.09	-0.00
	(0.025)	(0.033)	(0.033)	(0.057)	(0.031)
France	-0.04*	-0.02	-0.02	-0.03	-0.02
	(0.020)	(0.027)	(0.027)	(0.045)	(0.026)
Italy	-0.08***	-0.05*	-0.05	-0.12***	
	(0.025)	(0.030)	(0.030)	(0.036)	
Sweden	0.07*	0.10*	0.10*	0.06	0.12***
	(0.042)	(0.051)	(0.051)	(0.075)	(0.043)
UK	-0.06	-0.12**	-0.12**	-0.18***	-0.11*
	(0.044)	(0.056)	(0.056)	(0.057)	(0.057)
Constant	0.89***	0.73***	0.73***	0.77***	0.73***
	(0.052)	(0.071)	(0.071)	(0.094)	(0.094)
Observations	14,244	9,141	9,141	2,943	3,872
R-squared	0.216	0.238	0.239	0.228	0.246
F-test	106.4	70.74	67.35	30.99	41.39

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

All controls work as expected, with a lower probability of exclusion when patents and papers in the PPS have more similar contents (the coefficient of *proximity* is negative and significant) and when the publication comes after the patent (negative coefficients for 0,+1, and +2 years of difference between the publication date and the patent date). This indicates that we control effectively for false patent-publication matches in the PPSs. In addition, the estimated coefficients for the number of authors and the number of inventors in the patent have the expected signs.

Several country and technology variables have significant estimated coefficients. In particular, those for *Instruments*, *Chemicals & Materials*, and *Pharma & Biotech* are large, positive and significant, which is in line with our expectations that our patent-publication matching techniques work best in these fields. As for *Other technologies*, the frequent lack of significance is also explained by the low number of observations.

The signs of country coefficients do not lend themselves to any straightforward interpretation. The positive sign for Sweden, and the negative sign of all other countries (which implies a positive one for the Austria, the reference case) could be interpreted as pointing in the direction of the importance of the professor privilege, which was retained by Austria until the early 2000s and still is in place in Sweden. According to this intuition, the professor privilege grant exclusive control of intellectual property rights to

faculty, as opposed to the university administration, which suggests that in these countries the senior researchers and chief of labs (most likely to be tenured) would have more power to exclude other team members from the patent. However, we tried to interact the country dummies with either the seniority variables and the variables indicating the position of authors in the by-line, but got no significant results (unreported, but available on request) .

As for the signs of technology coefficients, the results may point at different practices across scientific disciplines (to which technological fields are correlated), but they can also be biased by the different quality of our patent-publication matching, for the part not captured by other regressors. At the moment, we propend for this interpretation, once again because interactions of seniority and position variables with the technological ones does not produce any result of interest (unreported, but available on request).

In columns (2) we introduce gender, which turns out to be positive and significant as expected: *ceteris paribus*, female authors have 8% more probability to be excluded for the inventors. All other estimated coefficients remain more or less the same, which suggests little interaction of this variable with the others. We also notice a non-trivial increase of the goodness-of-fit (R2 from less than 0.20 to more than 0.22), which also points at the importance of gender.

In column (3) we interact gender and position. The reason for doing this is as follows. When considering authors in intermediate position in the article by-lines (*Middle*=1, which is the omitted reference case) we cannot measure their relative contribution to the research effort, being the exact position in the by-line (second rather than third or fourth, but in any case never last) is uninformative in this respect. To the extent that women may be relegated to less important tasks, their exclusion from the patent, when middle authors, could be motivated not by gender per se, but by lower contribution to the invention. Column (3), however, proves this not to be the case, as the estimated coefficients for male and female first authors significantly differ: in particular, male first authors have a lower probability of exclusion than female ones, while female first authors are as likely to be excluded as male middle ones (their estimated coefficients not being significantly different from zero).

When interacting gender with country variables we do not find significant results (unreported, available on results). We find some results for the interaction with technologies, in particular with Pharmaceuticals & Biotechnologies (unreported, available on results). Yet, this may be due to the relative low presence of women in other fields, or to the lower numerosity of observations in other fields, which limit the significance of any estimate.

When removing the technology or the country with largest number of observations (respectively: Pharma&Biotech, column 4; and Italy, column 5) our results do not change. Some coefficients, while maintaining their sign, become insignificant, but this is most likely due to the drop in the number of observations.

In appendix B, table B.1 report the equivalent logit estimates (odds ratios) of columns (1) to (3) of table 5.1.1. and columns (1) to (3) of table 5.1.2, along with some graphical representation of predicted probabilities and marginal effect as function of the main variables of interest. The latter appear coherent

with the LPM estimates. However, one important advantage of logit regression with respect to LPM consists in the inherent non-linearity of the estimation. This makes room for interaction effects even in the absence of interaction terms. In this respect, figure B.3 is rather interesting, as it shows that gender differences in the probability of exclusion decline with the increase of seniority (*relative_scholarship* → 1).

As for robustness checks, table 8 reports the results for a set of regressions (identical to those in table 7) for the more restrictive 2p95 sample. The estimated coefficients do not change much.

Table 8. Exclusion regression: OLS estimates (2p95 PPS class)

	(1)	(2)	(3)	(4)	(5)
First	-0.16***	-0.15***			
	(0.035)	(0.039)			
Last	-0.22***	-0.19***			
	(0.034)	(0.032)			
Female		0.07***			
		(0.019)			
Female Middle			0.07***	0.05	0.06*
			(0.019)	(0.033)	(0.035)
Male first			-0.17***	-0.21***	-0.27***
			(0.048)	(0.067)	(0.055)
Female first			-0.06	-0.02	-0.11**
			(0.040)	(0.075)	(0.057)
Male last			-0.19***	-0.12**	-0.25***
			(0.036)	(0.048)	(0.046)
Female last			-0.17***	-0.13	-0.19**
			(0.056)	(0.079)	(0.088)
relative_scholarship	-0.42***	-0.40***	-0.40***	-0.37***	-0.29***
	(0.043)	(0.051)	(0.051)	(0.075)	(0.048)
most_junior	0.05**	0.04**	0.04**	0.04	0.05*
	(0.020)	(0.022)	(0.021)	(0.035)	(0.029)
proximity	-0.15*	-0.17*	-0.17*	-0.15	-0.06
	(0.080)	(0.101)	(0.101)	(0.178)	(0.173)
-1 year	0.04	0.02	0.02	0.06*	-0.07*
	(0.026)	(0.028)	(0.028)	(0.030)	(0.040)
0 years	-0.03	-0.02	-0.02	0.06	-0.12***
	(0.029)	(0.035)	(0.035)	(0.036)	(0.037)
+1 year	-0.05	-0.04	-0.04	0.02	-0.13***
	(0.036)	(0.038)	(0.038)	(0.041)	(0.035)
+2 years	-0.04	-0.06	-0.06	-0.00	-0.15***
	(0.041)	(0.040)	(0.041)	(0.045)	(0.036)
n_authors	0.01***	0.01***	0.01***	0.03***	0.02***
	(0.003)	(0.003)	(0.003)	(0.005)	(0.004)
n_inventors	-0.05***	-0.05***	-0.05***	-0.05***	-0.05***
	(0.005)	(0.006)	(0.006)	(0.013)	(0.008)
Instruments	0.10**	0.12*	0.11*	0.09*	0.11
	(0.049)	(0.059)	(0.059)	(0.056)	(0.077)
Chemicals; Materials	0.16***	0.19***	0.19***	0.20***	0.20**
	(0.057)	(0.060)	(0.060)	(0.060)	(0.079)
Pharmaceuticals; Biotech	0.15***	0.19***	0.19***		0.13*
	(0.049)	(0.058)	(0.058)		(0.076)
Other technologies	0.07	0.04	0.04	0.06	0.01
	(0.058)	(0.083)	(0.083)	(0.079)	(0.105)
Belgium	-0.09**	-0.08	-0.08	-0.06	-0.04
	(0.037)	(0.049)	(0.050)	(0.080)	(0.044)
Spain	-0.10***	-0.05	-0.05	-0.03	-0.02
	(0.032)	(0.046)	(0.046)	(0.061)	(0.040)
France	-0.08***	-0.06	-0.06	0.03	-0.05
	(0.031)	(0.041)	(0.041)	(0.059)	(0.036)
Italy	-0.13***	-0.11**	-0.11**	-0.11**	
	(0.035)	(0.047)	(0.047)	(0.044)	
Sweden	0.03	0.05	0.05	0.16***	0.08*

	(0.048)	(0.061)	(0.061)	(0.045)	(0.047)
UK	-0.10	-0.16**	-0.16**	-0.12*	-0.14*
	(0.065)	(0.076)	(0.076)	(0.074)	(0.077)
Constant	0.97***	0.90***	0.90***	0.69***	0.87***
	(0.072)	(0.090)	(0.090)	(0.126)	(0.127)
Observations	8,070	5,075	5,075	1,381	2,109
R-squared	0.225	0.241	0.242	0.227	0.240
F-test	80.26	70.18	65.52	20.66	21.68

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

5.2 Renewed co-authorship regressions

Table 9 reports the results for three different specifications of the regression, with columns (1) to (3) being OLS estimates and columns (4) to (6) being Logit (estimated coefficients reported). We have chosen not to relegate Logit estimates to the Appendix because they serve well the purpose of illustrating some non-linearities in the marginal effects of the variables of interest, even without the insertion of specific interaction terms (too many of which would make the results less intelligible). Data refer to PPS class in the 3p90 percentile of patent-publication similarity.

Estimated coefficients in column (1) for *First* and *Exclusion*, and their interactions, go in the direction of confirming proposition 2 (the same holds for column (4)). Exclusion from the patent does not reduce the first author's probability to publish again with the PPS team: the sum of coefficients for *Exclusion* and *First*Exclusion* is not significantly different from zero. Notice also that both first and last authors have a higher probability than middle ones to renew co-authorship. As for proposition 3, we observe that the estimated coefficient for *Exclusion*, which indicates the marginal effect of the exclusion from the patent for a middle authors, is significantly different from zero. This goes in the direction of confirming the proposition, although not completely, as we do not yet control for the first and middles authors' contribution to the PPS.

We try to do so in column (2), by inserting the number of papers each author has contributed to in the PPS, and its square term. We first observe that this variable has the largest predictive power on the probability of renewed co-authorship (the R2 of the regression jumps from around 8% to 50%). Both the coefficients for *First* and *Last* go to zero, as it is clearly the case that first and last authors are those most committed to the team and have the lion's share of its efforts and publications. Still, the previous results hold, the sum of coefficients for *Exclusion* and *First*Exclusion* being once again zero. We observe however two counter-intuitive results. First, exclusion from the patent(s) appears to reinforce the last author's probability to renew co-authorship, as the sum of coefficients for *Exclusion* and *Last*Exclusion* is positive and significant (albeit only at 90%). Second, the sign for *Proximity* turns to negative, which apparently suggests that the closer the papers and the publications in the PPS, the lower the probability to observe renewed co-authorship. As we intend *Proximity* as a control for the quality of our patent-publication matching, we expected the contrary (as it is the case in column (1)). Removing potential outliers, such as authors with a very large number of publications (over 10), does not change the results (unreported, but available on request).

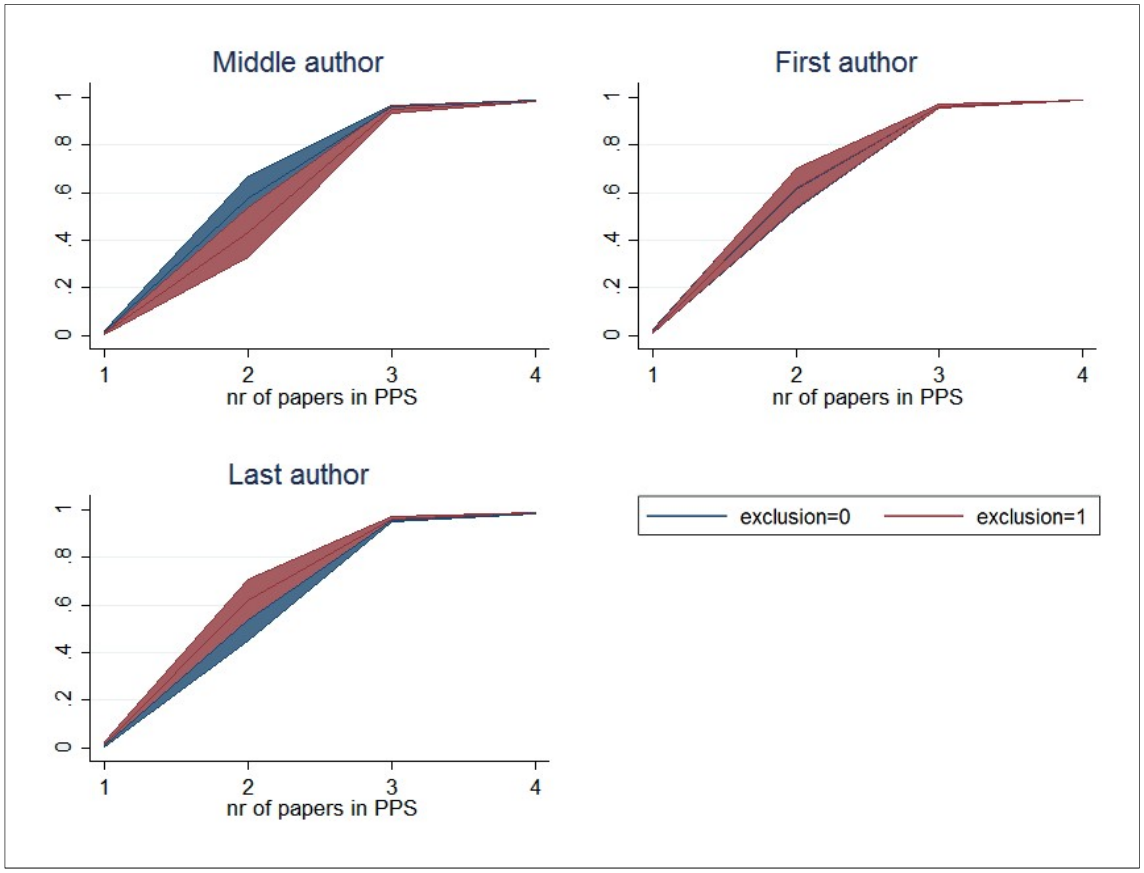
Table 9. Renewed co-authorship regressions: OLS & Logit estimates (3p90 PPS)

	(1)	(2)	(3)	(4)	(5)	(6)
First	0.07***	0.02	0.01	0.48***	0.18	0.09
	(0.016)	(0.012)	(0.016)	(0.110)	(0.167)	(0.190)
Last	0.11***	-0.01	-0.01	0.63***	-0.21	-0.22
	(0.016)	(0.013)	(0.018)	(0.099)	(0.181)	(0.204)
Exclusion	-0.08***	-0.04***	-0.04***	-0.79***	-0.70***	-0.61***
	(0.011)	(0.009)	(0.012)	(0.103)	(0.162)	(0.203)
First*Exclusion	0.05**	0.04**	0.05**	0.56***	0.74***	0.76***
	(0.022)	(0.015)	(0.021)	(0.154)	(0.229)	(0.264)
Last*Exclusion	-0.02	0.07***	0.09***	0.19	1.15***	1.34***
	(0.022)	(0.017)	(0.022)	(0.160)	(0.258)	(0.264)
Female			-0.00			0.00
			(0.009)			(0.136)
Nr authorships in PPS		0.35***	0.35***		5.01***	3.79***
		(0.022)	(0.024)		(0.363)	(0.372)
Nr authorships in PPS (sq)		-0.01***	-0.02***		-0.43***	-0.14***
		(0.002)	(0.002)		(0.037)	(0.015)
First publication year	-0.00**	-0.00	0.00	-0.01***	-0.00	0.00
	(0.000)	(0.000)	(0.000)	(0.003)	(0.005)	(0.006)
Publication stock	0.00***	0.00	0.00	0.00***	0.00	0.00
	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)
Proximity	0.24***	-0.16**	-0.16**	2.05***	-3.68***	-3.60**
	(0.076)	(0.067)	(0.082)	(0.613)	(1.352)	(1.464)
N_authors in PPS	0.00***	0.00	-0.00	0.02***	0.01	0.00
	(0.001)	(0.001)	(0.001)	(0.004)	(0.010)	(0.012)
N_inventors in PPS	0.01**	0.00	0.00	0.05**	0.00	0.03
	(0.003)	(0.003)	(0.004)	(0.024)	(0.053)	(0.066)
Last year in PPS (dummies)	yes	yes	yes	yes	yes	yes
Constant	1.61**	-0.13	-0.93	14.75**	-1.29	-15.23
	(0.776)	(0.607)	(0.814)	(6.325)	(10.683)	(12.495)
Observations	10,005	10,005	5,819	10,005	10,005	5,819
R2/Pseudo-R2	0.077	0.499	0.515	0.0934	0.568	0.565
F-test	21.67	63.46	53.71			
Model chi-square				428.6	485.5	326.3

Columns (1) to (3): OLS ; Columns (4) to (6) Logit (estimated coefficients reported)
Robust standard errors in parentheses ; *** p<0.01, ** p<0.05, * p<0.1

The logit version of the same specification (column (4)) suggests an important qualification of our results, which we illustrate in figure 4. The figure reports the predicted probabilities of renewed co-authorship for middle, first, and last authors, as a function of the number of publications they contributed to the PPS and the exclusion from patents. For all types of authors, the probability to renew co-authorship appears to be close to zero if they contributed only one publication to the PPS (occasional PPS members or authors therein), and close to one if they contributed three articles or more (stable PPS members and authors), irrespective of their inclusion or exclusion from the patent(s). Having been excluded makes a difference only for authors with two papers in the PPS, which we could describe as neither occasional nor stable, and only for middle authors.

Figure 4. Predicted probability of renewed co-authorship, by author's position, inclusion in patent, and ex-ante co-authorships (Logit estimates - 3p90 PPS)



As from regression (3), table 5.2.1 (95% confidence intervals) / Other variables at means

Going back to table 5.2.1 we comment upon column (3), in which we control also for the authors' gender, but do not get any significant results for the specific coefficients. As for the others, the sign and size do not change much, despite the sample being halved due to missing observation for the new regressor.

Table 10 reports the results for a robustness checks based on replacing the 3p90 sample with the 2p95 one, which contains closer patent-publication matches. Overall we obtain the same results as in the previous table, with some improvement. In particular, when controlling for *Nr authorships in PPS*, the sign of Proximity is no more negative; and the renewed co-authorship probability for the last author no more appears to be (positively) influenced by the exclusion from the patent (the sum of coefficients for *Exclusion* and *Last*Exclusion* is zero).

Table 10. Renewed co-authorship regressions: OLS & Logit estimates (2p95 PPS)

	(1)	(2)	(3)	(4)	(5)	(6)
First	0.07***	0.02	0.04*	0.59***	0.10	0.12
	(0.021)	(0.017)	(0.021)	(0.174)	(0.282)	(0.301)
Last	0.08***	0.02	0.03	0.60***	-0.36	-0.20
	(0.021)	(0.017)	(0.023)	(0.162)	(0.279)	(0.330)
Exclusion	-0.07***	-0.04***	-0.04**	-0.75***	-0.80***	-0.71***
	(0.016)	(0.012)	(0.016)	(0.159)	(0.201)	(0.251)
First*Exclusion	0.05*	0.05**	0.06**	0.49***	0.71**	0.67*
	(0.024)	(0.023)	(0.030)	(0.191)	(0.324)	(0.374)

Last*Exclusion	-0.01	0.03	0.03	0.09	1.11***	0.87**
	(0.032)	(0.023)	(0.030)	(0.238)	(0.345)	(0.396)
Female			0.01			0.04
			(0.010)			(0.165)
Nr authorships in PPS		0.14***	0.14***		4.08***	4.06***
		(0.024)	(0.022)		(0.395)	(0.434)
Nr authorships in PPS (sq)		-0.00***	-0.00***		-0.29***	-0.28***
		(0.000)	(0.000)		(0.051)	(0.049)
First publication year	-0.00***	-0.00***	-0.00**	-0.02***	-0.01	-0.01
	(0.000)	(0.000)	(0.001)	(0.004)	(0.007)	(0.008)
Publication stock	0.00**	0.00	0.00	0.00***	0.00	0.00
	(0.000)	(0.000)	(0.000)	(0.001)	(0.002)	(0.002)
Proximity	0.48***	0.27***	0.31***	4.52***	0.50	0.53
	(0.096)	(0.089)	(0.106)	(0.887)	(1.700)	(1.975)
N_authors in PPS	0.00***	0.00	0.00	0.01***	0.01	0.01
	(0.000)	(0.000)	(0.000)	(0.002)	(0.007)	(0.007)
N_inventors in PPS	-0.01**	-0.01***	-0.01***	-0.05*	-0.23***	-0.24***
	(0.003)	(0.003)	(0.003)	(0.026)	(0.074)	(0.079)
Last year in PPS (dummies)	yes	yes	yes	yes	yes	yes
Constant	2.54***	2.39***	2.58**	28.13***	14.52	3.54
	(0.952)	(0.747)	(1.157)	(8.419)	(14.724)	(16.763)
Observations	6,241	6,241	3,581	6,241	6,202	3,552
R2/Pseudo-R2	0.089	0.349	0.359	0.118	0.538	0.553
F-test	13.36	51.26	34.51			
Model chi-square				315.1	455.0	394.1

Columns (1) to (3): OLS ; Columns (4) to (6) Logit (estimated coefficients reported)
Robust standard errors in parentheses ; *** p<0.01, ** p<0.05, * p<0.1

6. Conclusions

Scientific credit is increasingly earned by individual scientists by means of contributions to collective, team-based research, whose results are indivisible across team members. In this paper we have argued that scientists in teams negotiate over the distribution of attribution rights in order to influence how third parties will bestow scientific credit to each of them. Since they attach different marginal values to authorship (and, in several fields, first authorship in particular) and inventorship, they can by and large reach privately optimal solutions, which will preserve the stability of the team, conditional on the bargaining power each team member has, according to his/her seniority and gender. However, nothing guarantees that this solution will also be socially optimal. In fact, interested third parties and society at large will be induced to under(over)-estimate the contribution to technology transfer by junior and female (senior and male) scientists, due to manipulation of inventorship attribution.

Our paper contributes to what we described (in section 2) as the prequel literature on scientific credit distribution and team formation. It assumes that society applies a specific heuristic to distribute scientific credit across research team members (it values first over middle authorship; it reads last authorship as typical of senior authors; it splits credit for invention equally across inventors), and it infers its influence on team formation (in our case, team stability). It also contributes to what we described as the sequel literature, as it assumes that research is conducted by teams, and studies how different individuals within

the team may earn more or less scientific credit according not just to their contribution, but to their status (bargaining power).

With respect to our own previous work (Lissoni et al., 2013) we prove the robustness of the original results by extending them to several European countries besides Italy, and extend the implications of our "negotiated" view of attribution rights to the case in which junior scientists have an exit option (and not just a voice, aka litigation, one).

From the policy and managerial viewpoint, the key implication of our results is that scientists in a team have an economic interest in keeping the information on individual contribution as private as possible, since this gives them latitude to negotiate over attribution rights. As originally suggested by Robert Merton, the team leader (the Senior scientist of our model) may even have little interest in establishing clearly among themselves who did what and what he/she deserves in terms of public recognition. Such lack of clarity may give him/her more latitude in distributing the attribution rights according not to objective criteria, but subjective needs (returns from right), and for the sake of the team's stability.

From the policy viewpoint, this suggests that the national evaluation agencies (such as the English HEFCE, the Italian ANVUR, or the French HCERS) should be wary of treating bibliographic information as an objective measure of contribution to research and invention. And of how the economic value they attach to authorship and inventorship (in the form of rewards attached to it, for either the individuals or their institutions) will inevitably affect the negotiations within the team.

These implications are even stronger from a managerial viewpoint, to the extent that university administrators are even more likely than evaluation agencies to distribute rewards for authorship and inventorship not just to teams or collections of teams (such as departments or faculties), but individuals. For example, an excessive emphasis on first authorship may put team stability at risk (as hopeful junior scientists who get middle authorship instead, are more likely to quit); and at the same time it gives leverage to senior scientists for staking their exclusive (or less inclusive) claims on inventorship.

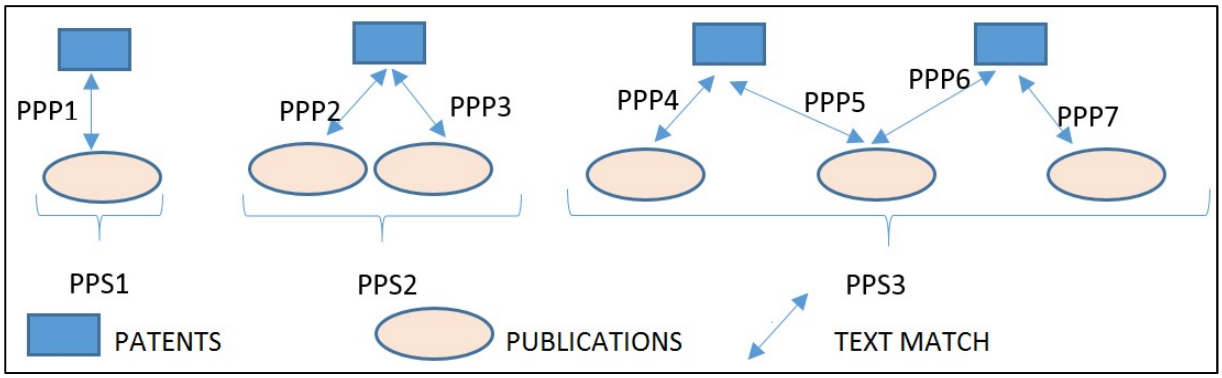
Our future research will go in the direction of exploring more in detail how such policy and managerial practices may affect, in theory, or have affected, historically, the scientists' negotiation practices.

APPENDIX A - Patent-publication data collection and linkage

A.1 Patent-publication pairs and sets: definition and collection methodology

Observations in our regression analysis are authors of papers included in patent-publication "pairs" and/or "sets". Patent-publication pairs (PPPs) are the empirical equivalent of "double disclosures" in our theoretical model. They consist of patents and publications whose lists of inventors and authors share at least one surname-and-initial and, according to text analysis of their titles and abstracts, are very likely to deal with the same research result. To the extent that the research result may be described by more than one patent and/or publication, PPPs may combine to form what we call patent-publication sets (PPSs), in which we have either one patent matched to N publications, or (less often) vice versa, or N patents matched to M publications (most often with N<M). In figure A.1, lines represent PPPs (dyadic relationship between one patent and one publications, for a total of 7 instances), while PPSs are indicated with horizontal brackets, for a total of 3).

Figure A.1 Patent-publication pairs (PPPs) and patent-publication sets (PPSs): definition



Our main source for patent data is the APE-INV programme, whose database contains information on patent applications over inventions by academic scientists from several European countries (in short: "academic patents"), as filed at the European Patent Office (EPO) from its opening in 1978 to around 2010. The countries we selected for this paper are Austria, Belgium, Italy, Spain, Sweden, and the UK. To these we added France, whose data were collected independently from APE-INV, but with a similar methodology. All data, but the French ones, are available on the APE-INV website (http://www.esf-ape-inv.eu/index.php?page=3#acadpat). We selected, where available, only patents with priority dates comprised between 1997 and 2007.

According to the APE-INV definition, which emphasizes the origin of the invention as opposed to its property, academic patents were identified through name matching of inventors and academic scientists. Whether assigned to a university, a firm, or an individuals, all patents in our dataset contain at least one "academic inventor", that is an inventor who conducted her patent-related research within a university.

For all countries, inventors' names come from EPO patent documents and were disambiguated, before matching, either by means of the Massacrator 2.0 algorithm (Pezzoni et al., 2014) or by similar algorithms (all discussed in a series of APE-INV "Name Game" workshops). As for scientists' names, sources were less homogeneous, varying from administrative records of central governments (Italy) and individual universities (Austria, Belgium, France, Sweden, and the UK) to bibliometric sources (Spain). This implies that in some countries only tenured faculty were considered for name-matching with inventors, while in others untenured faculty, and even post-docs may have been included. In addition, for some countries data were provided for all universities and disciplines, while for others only selected ones were available (most notably, for Belgium only French-speaking universities were considered, while for the UK we have only

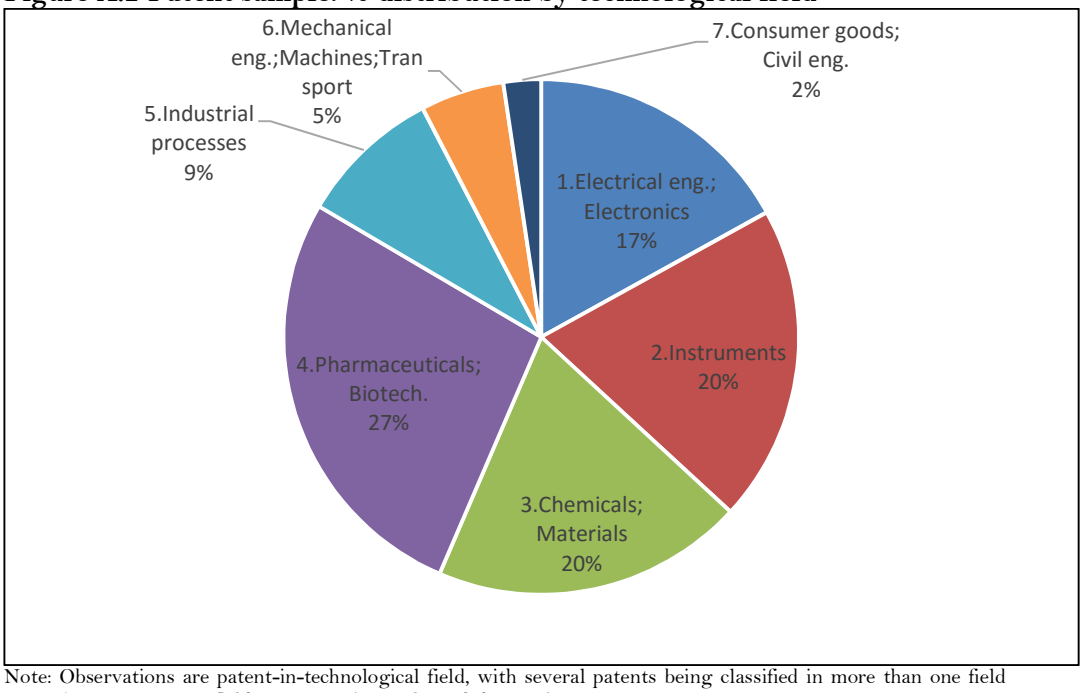
engineering faculties; as for France, small universities were ignored). Nor time coverage was uniform (we miss 1997-98 patents for Spain, and 2007 patents for the UK; see table A.1). As a result, our sample cannot be considered as representative of academic patenting in Europe, due especially to the under-representation of Belgium, France, and the UK, and the over-representation of Italy and Spain. More details on the sampling scheme for each country can be found in the papers also listed in table A.1.

Table A.1 Patent sample and methodological reference, by country

	patents		years		reference paper
	nr	%	from	to	
Austria	572	6.4	1997	2007	Stummer et al. (2013)
Belgium	457	5.1	1997	2007	Mejer (2013)
Spain	1820	20.4	1999	2007	Maraut and Martinez (2014)
Italy	3217	36.0	1997	2007	Pezzoni et al. (2014)
Sweden	579	6.5	1997	2007	Ljungberg et al. (2014)
UK	917	10.3	1997	2006	Banal-Estanol et al. (2010)
France	1371	15.3	1997	2007	Cassi (xxx)
Total	8933	100			

Figure A.2 reports the breakdown of our patent dataset by technological field. We notice that the first four fields, which are the most science-based ones, make up for over 80% of the observations, with Pharmaceuticals & Biotechnology accounting alone for over one quarter of the observations, despite the absence of patents from UK in this field. This is in line with the literature on academic patenting, as surveyed by Lissoni (2013).

Figure A.2 Patent sample: % distribution by technological field



Note: Observations are patent-in-technological field, with several patents being classified in more than one field (on average, 1.33 fields per patent). Total nr of observations = 12187

In order to identify the publications to associate to the academic patents in our sample we searched the Web of Science © (WoS, by Thomson Reuters), as follows. For each patent in our sample, we searched retrieved a large number of publications by authors with the same name and initials of the patent inventors. We restricted the search by producing a table of "incompatible" technological fields (of patents) and

disciplines (of journals)[8]. We also considered only the publications within a time range of two years before and two years after the patent's priority year.[9]

We then filtered all matches by technological fields. For each field, we conducted a separate text analysis of all matches, based on a weighted bag-of-words method. Roughly speaking, the method consists first in eliminating stop words from the titles and abstracts (we used the stop word lists available in Python; https://pypi.python.org/pypi/stop-words/2014.5.26). Second, a weight is assigned to all the remaining words, which is inversely proportional to the frequency of the word in the technological field under examination. In this way, the title and abstract of each document in a match are transformed in a vector of weights (varying from 0, for words that are present in the vocabulary of the technological field, but not on the document; and ~1, for very rare words). Finally, on the basis of the vectors of each matched patent and publication, we computed a "similarity index" equal to the cosine distance between the vectors, as in Lissoni et al. (2013)[10].

At the present stage of our research, we had access only to the publications dated from 1999 onward. This returned over 4.6 million matches with non-zero values of the similarity index, based on around 140000 publications for around 9000 patents. Of these we retained for further analysis only the matches with similarity index greater than 0.2, which amount to less than half of all observations in the first percentile of the frequency distribution of the index values[11]. In total, we are dealing with 4359 publications and 2260 patents. These form 6997 PPPs, which in turn form 1761 PPSs. In the following, we indicate this class of PPSs with the label p4Q3

Figure A.3 reports the distributions of PPPs by technological field of the patents. When compared with figure A.1 it shows immediately that "Pharmacauticals & Biotechnology" is the technological field in which patents are most likely to correspond to a publication. The PPP share of this field is 40%, that is 13% more than that of patents. Among the other science-based technological field, "Electrical engineering & Electronics" and "Chemicals & Materials" see their share going down of, respectively, 4% and 3%. This is possibly due to an inferior propensity of scientists in these fields to "double disclose" their inventions, but also to the fact that they privilege conference proceedings to publications, with the former being underrepresented in WoS and not included in our search strategy. As for non-science based fields, they altoghether drop from 16% of the original patent sample to less than 10% of the PPPs.

If we further restrict our analysis to higher values of the similarity index (top 10% of the first percentile) the number of PPPs falls to 952 (label 3p90). One further restriction (to the top 5% of the first percentile) further reduces to PPS to 561 (label 2p95).

[8] For technological fields, we reclassified the original IPC codes of patents into 30 classes, in accordance with the OST-INPI/FHG-ISI reclassification methodology, as updated by Coffano and Tarasconi (2014). As for disciplines, we simply adopted the classification scheme of the Journal Citation Reports © (JCR, also by Thomson Reuters).

[9] In principle, no scientist should publish a paper before filing a patent at the EPO, in order to avoid killing the novelty requirement (contrary to the USPTO, the EPO does not allow for any grace period; Franzoni and Scellato, 2010). At the same time, it is unlikely that the same scientist, once filed the patent, will wait too long before submitting the paper, in order to avoid compromising her race to priority in academic recognition (in principle, it could make sense sending the paper to a journal right after having filed the patent application). This suggested us to limit the search for publications to a short time (that we set in two years) after the priority date of the patent. At the same time, though, it is possible that the scientist will try to publish a paper before filing the patent application, either by mistake (due to scant knowledge of patent laws) or by limiting the disclosure to parts of the research results that will not form the subject of claims. (This risks invalidating the patent, and in fact, a sizable proportion of patents on our database have not been granted, most often due to withdrawal after the publication of the search report.) For this reason, we also consider publications that appeared on journals up to two years before the priority date.

[10] Notice that when a patent was classified under more than one technological fields, we computed more than one similarity index for each one of its publication matches, and then retained the maximum value. This is because the various technological fields have different vocabularies, from which we obtain different weights for the same word, and possibly different similarity indexes between pairs of documents.

[11] More precisely, the observations we retain are those falling into the top quarter of the first percentile.

Figure A.3 Selected PPPs: % distribution by technological field

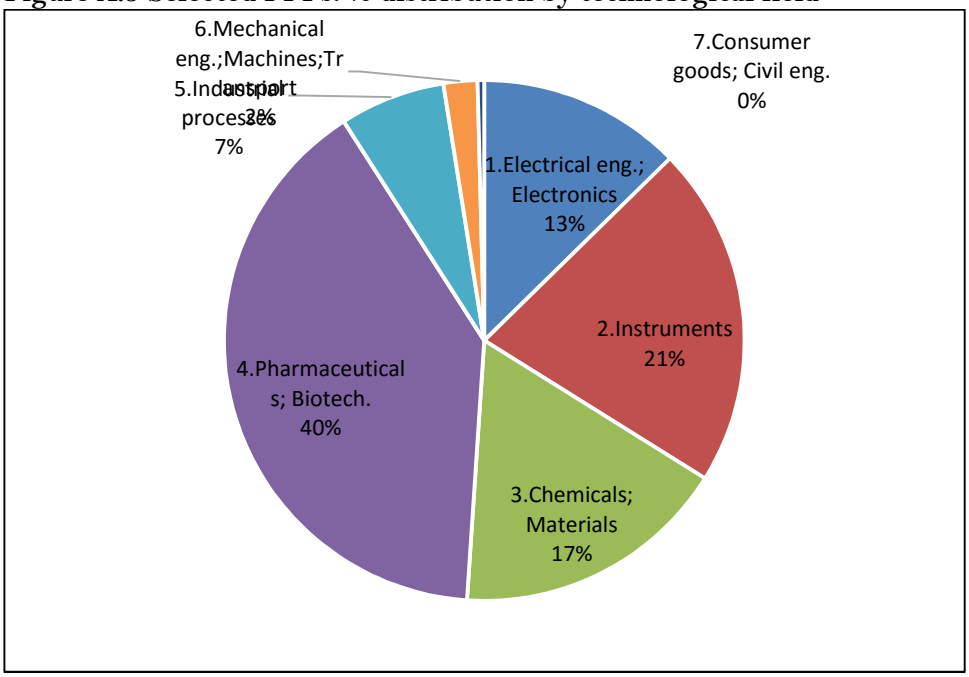


Table A.2 shows the distribution of similarity index across the PPSs within each possible sample: the more restrictive the sample (moving from p4Q3 to 2p95), the higher the average and median values (for PPSs including more than one paper and/or publication, the maximum value is considered).

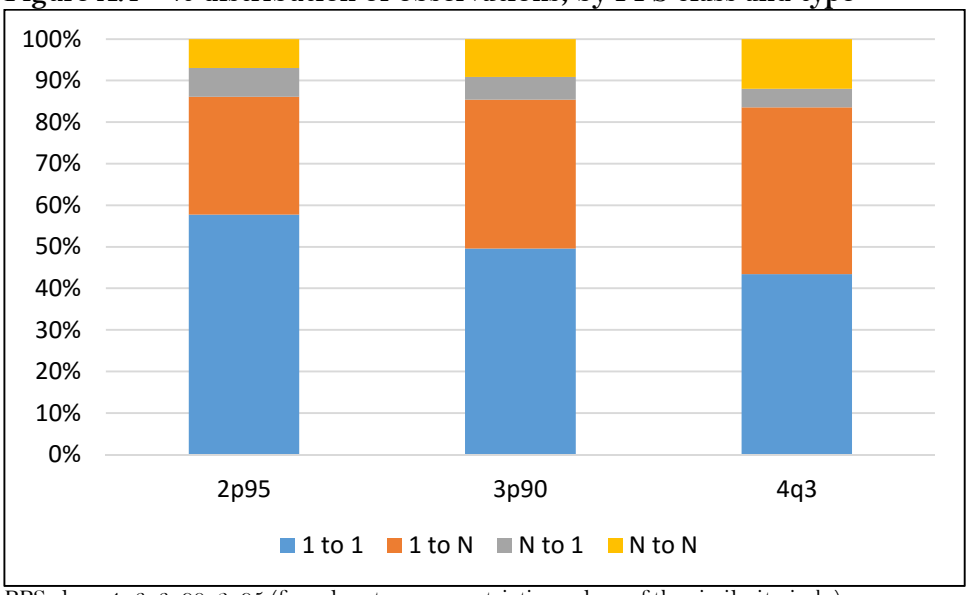
Table A.2 – Patent-publication sets (PPSs): % distribution of observations, by sample (from less to more restricted) and values of similarity index*

	Similarity index (value intervals)							
	0.2-0.3	*0.3-'0.4*	*0.4-0.5*	*0.5-0.6*	*0.6-0.7*	*0.7-0.8*	*>0.8*	*TOT*
4q3	58.1	24.4	11.0	4.6	1.2	0.5	0.2	100
3p90	19.0	47.6	21.3	8.6	2.3	0.8	0.3	100
2p95	--	42.1	37.3	14.8	3.9	1.4	0.5	100

* similarity index = inverse of cosine distance between title/abstract of each patent and each publication in the PPS

Figure A.4 shows the distribution of PPSs by type, where types are defined according to the number of patents and publications included in the PPS, by sample. We notice that the more restrictive the sample, the higher the percentage of 1-to-1 matches (i.e. the percentage of PPSs that coincides with straightforward PPPs). This percentage goes from around 40% in 4Q3 to almost 60% in 2p95. We also notice that 1-to-N (1 patent to N publications) is the second most frequent type of PPS, which ranges from 40% in 4Q3 to 28% in 2p95. N-to-1 and N-to-N PPSs are residual category, which together account in between 14% and 16% of all PPSs.

Figure A.4 – % distribution of observations, by PPS class and type



PPS class: 4q3, 3p90, 2p95 (from less to more restrictive values of the similarity indx)
PPS type: 1-to-1, 1-to-N, N-to-1, N-to-N (nr patents –to- nr of publications in the PPS)

Tables A.3 and A.4 show the distribution of the number of patents and publications in the various PPS classes. We see once again that the largest majority of the PPSs in the non-restrictive 4q3 class are of the 1-to-N type, as they include just one patent. But when we move to the more restrictive 3p90 and 2p95 classes, which we use for our analysis, the 1-to-1 matches prevail. We also notice that PPPs of the N-to-1 or N-to-N type mostly include no more than 2 patents, for the most restrictive classes (2P95 and 3P90), or 3 patents, for the least restrictive (4Q3).

As for the publications, the 1-to-N and N-to-N PPS types rarely include more than 5 publications in the most restrictive class (2p95) and 10 publications in the second most restrictive one (3P90), while in the case of the least restrictive (4Q3) we observe a substantial number of PPSs with more than 10 publications.

Table A.3 Number of patents in the PPS - Frequency distribution

	Nr of observations (in PPS class)		
nr of patents in the PPS	*pps_2p95*	*pps_3p90*	*pps_4q3*
1	483	813	1471
2	62	109	198
3	13	23	56
4		2	15
5	1	2	8
6	1		3
7			3
8			1
9		1	2
10			
11	1		
12			
13		1	
14		1	1
15			2
16			1
Total PPSs	561	952	1761
Total nr of patents	668	1154	2260

Table A.4 Number of publications in the PPS - Frequency distribution

	Nr of observations (in PPS class)		
nr of publications in the PPS	*pps_2p95*	*pps_3p90*	*pps_4q3*
1	363	524	843
2	96	187	332
3	44	96	205
4	24	48	118
5	10	30	53
6	6	19	40
7	4	16	33
8	3	8	25
9		2	14
10	3	1	14
11-20	5	16	64
21-50	3	3	13
51-100		2	6
100-200			1
Total PPS	561	952	1761
Total nr of publications	1142	2299	5570

The high percentage of 1-to-1 PPSs, especially in the more restrictive samples, suggests that our algorithm can be quite selective when it comes to match publications to patents. At the same time, it suggests that including in the sample many PPPs with rather low values of the similiarity index can lead to inflating the number of 1-to-N, N-to-1, and N-to-N PPS types, which may hide many false positive matches (matches between patents and publications that, in reality, are not instances of double disclosure).

To further check the quality of data we calculate how often our PPSs end up including patents by academic inventors from different countries. Based on the reasonable assumption that such patents most likely unrelated one to another, any PPS with more than one countries is very likely to include one or several false positive patent-publication matches. The results are reassuring. When considering PPS in the 3p90 class we have only 4 cases with two countries out of 952 (and no cases with more than two countries); and again two cases with two countries out of 561 in the 2p95 class.

We finally notice that, concerning the technological fields of patents, the PPPs we obtained through our text-matching techniques are disproportionately concentrated in Pharmaceuticals & Biotechnology, well over and above the distribution of the original patent sample.

A.2 Number of authors and inventors in PPSs: descriptive analysis

We calculate the difference between the number of inventors and the number of authors in each PPP included in the least restrictive PPS class (4Q3), as well as in each PPS, for each class (from the least to the most restrictive). In the case of PPPs (and of 1-to-1 PPSs) the difference is simply the difference between the number of inventors on the only patent and the number of authors on the only publication in the pair. In the case of 1-to-N, N-to-1, and N-to-N PPSs (that is, of PPSs that include more than one PPP) the difference is that between the total number of distinct inventors of the patents and the total number of distinct authors of the publications in the same PPS. Negative values of the inventors-authors difference indicate that, in the PPP or PPS, we have more authors than inventors, as expected. Figure A.5 shows the

frequency distribution of values. In the large graph, for readability reasons, we have truncated the values of the inventors-authors difference to -100 (no rightward truncation). The smaller graph is the same as the large one, but with no truncation. Table A.5 reports the same data, in a more compact way.

Figure A.5 - % distribution of observations, by value of the difference between number of inventors and authors in PPPs and three classes of PPSs

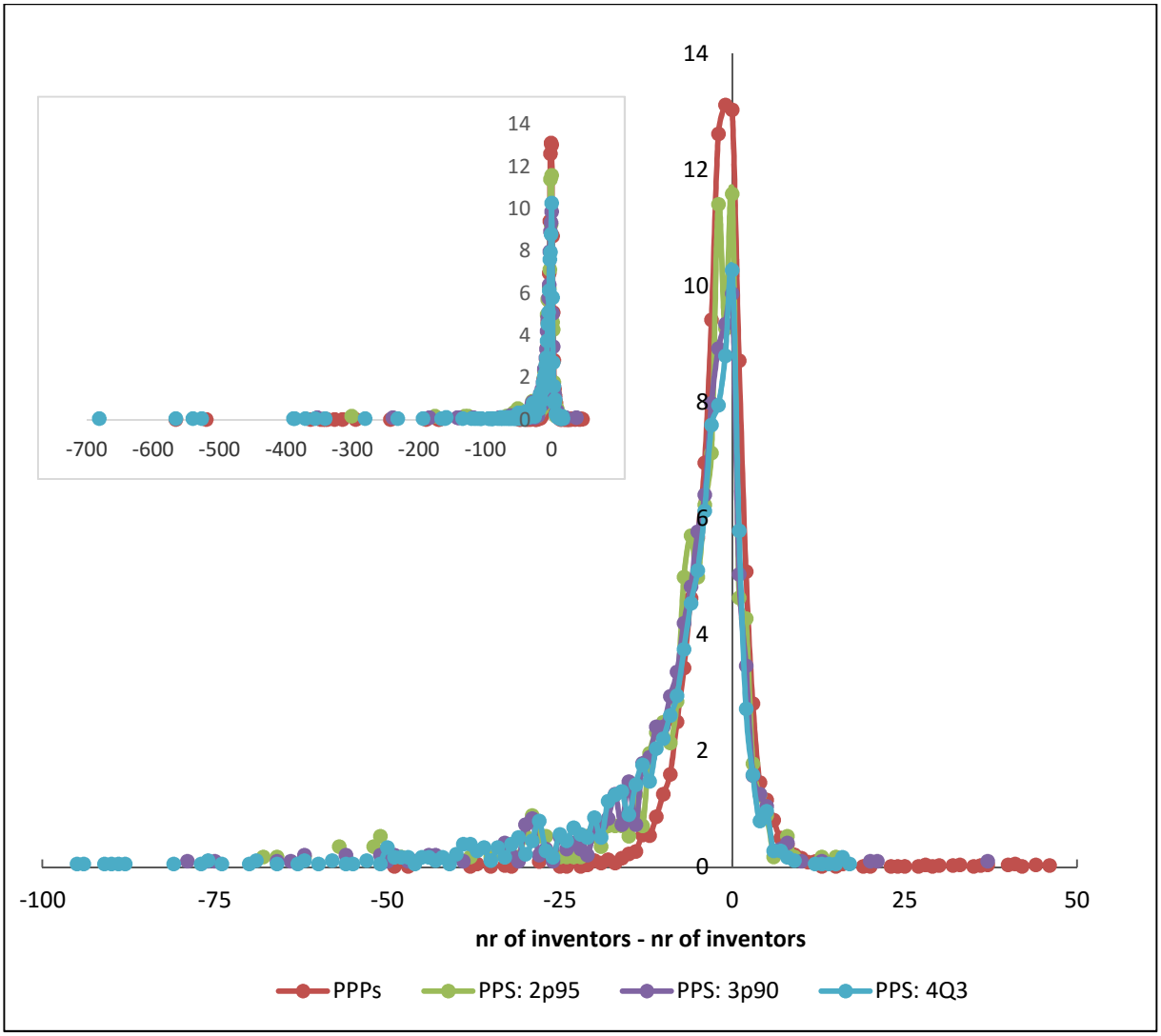


Table A.5 - % distribution of observations, by value of the difference between number of inventors and authors in PPPs and three classes of PPSs

	PPPs	PPS: 2p95	PPS: 3p90	PPS: 4Q3
-700,-100	0.2	0.7	0.7	1.2
-99,-50	0.0	1.8	1.3	1.8
-49,-10	4.6	17.1	20.7	24.0
-9,-5	17.8	20.7	21.1	19.0
-4	7.0	6.2	6.4	6.1
-3	9.4	7.1	8.0	7.6
-2	12.6	11.4	8.9	8.0
-1	13.1	9.3	9.3	8.8
0	13.0	11.6	9.9	10.3
+1	8.7	4.6	5.0	5.8

+2	5.1	4.3	3.5	2.7
+3	2.8	1.8	1.6	1.6
+4	1.5	1.2	1.3	0.8
+5,+9	3.1	1.8	1.8	1.8
+10,+49	1.0	0.4	0.5	0.5

We see that the distribution of values leans to the left: there are many more negative cases than positive ones, as expected. However, the mode value is zero, even though its frequency comes close to that of the lowest negative values (in absolute terms, that is : -1 and -2). For PPPs, we also observe a long tail to the left, which is due to the 1-to-N cases in which we have either many publications (N is large) and/or many authors per paper (which, in a few limited cases, may be over 50 or even over 100).

When it comes to PPPs, around 10% of the cases we retained exhibit the same number of inventors and authors; around 50% of the cases have more authors than inventors, up to a difference of 9 in absolute value; and 13% of the cases have more inventors than authors,up to a difference of 9 in absolute value. The remaining 17% of cases are in the tails (absolute difference of 10 or greater), mostly in the left hand one (authors>inventors). This distribution is in line with our expectations.

When considering PPS, and moving from the least to the most restrictive class (which include an increasing share of 1-to-N and N-to-N types), we find an increasing percentage of observations in the left tail : a little less than 20% in 2p95, more than 23% in 3p90, and 27% in 4Q3. (In the last case we have several some observations with difference between the number of inventors and authors well over -200).

Again, this preliminary evidence suggests to limit the sampling to one of the two most restrictive cases (2p90 and 3p90), as the least restrictive one (4Q3) may include too many false matches, as indicated by the disproportion between the number of authors and the number of inventors in the set.

Figure A.6 examines the distribution of values for the inventor-author differences according to the technological field of the patents in the PPPs we retained for analysis. Shades of red indicate negative values (more authors than inventors), shades of blues positive ones (more inventors than authors), with white for null values (same number of authors and inventors). We notice a clear prevalence of negative values in three out of four science-based fields: "Instruments", "Chemicals & Materials", and "Pharma & Biotech" (which exhibits the largest differences, as also found by Fehder et al., 2014). As for the fields with fewer connections with science and a marginal presence of academic patents ("Industrial process" and, especially, "Mechanical engineering & Transport" and "Consumer Goods"), the prevalence of negative values is much less clear, with the extreme case of "Consumer Goods", in which it is the positive values that prevail. This evidence in is line with our expectation to find many more instances of "double disclosures" in the science-based technologies, where the inventive activity of academic scientists may be a straightforward consequence of their research activity. On the contrary, in more traditional fields, we expect to find several academic inventions to stem out of targeted applied research or consultancy, or even extra-academic activities. This implies that, in fields where we suspect to have fewer instances of double disclosure, our methodology will produce a higher rate of false positives, which may explain the many cases in which the number of authors in the PPP is higher than that of inventors. A slightly puzzling result is that for "Electrical engineering & Electronics", which is also science-based, but does not exhibit a clear prevalence of negative values. One possible explanation may reside in the different publication strategy followed by academics in this field, which target the proceedings of important conferences rather than journals, the former being under-represented in WoS and therefore excluded from our search.

Figure A.7 examines the distribution of values for the inventor-author differences according to the priority year of the patents in the PPPs (all PPPs considered). We do not observe any time trend.

Figure A.6 % frequency distribution of PPPs by difference between number of inventors and authors, and technological fields of patents

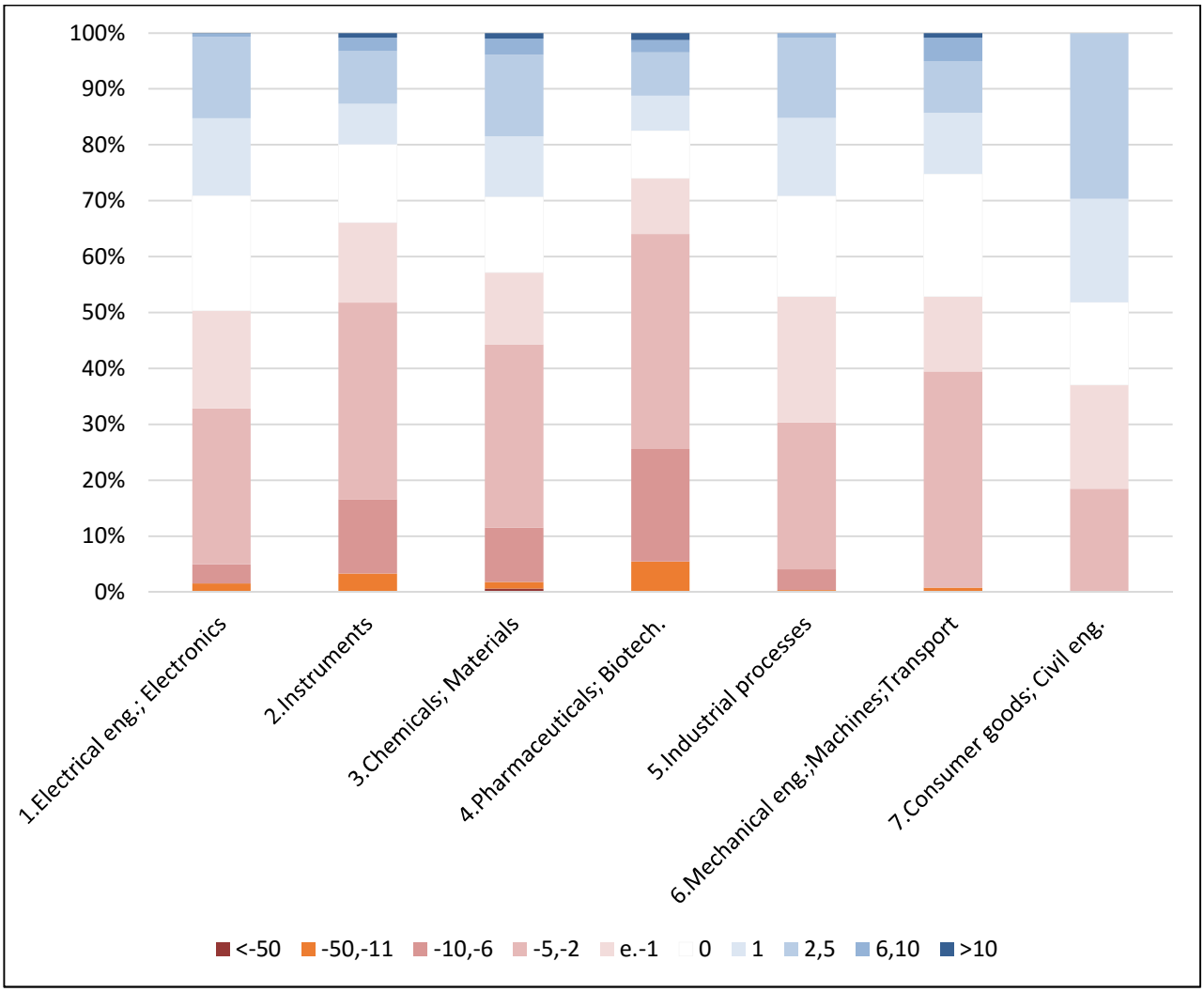


Figure A.7 % frequency distribution of PPPs by difference between number of inventors and authors, and priority year of the patents

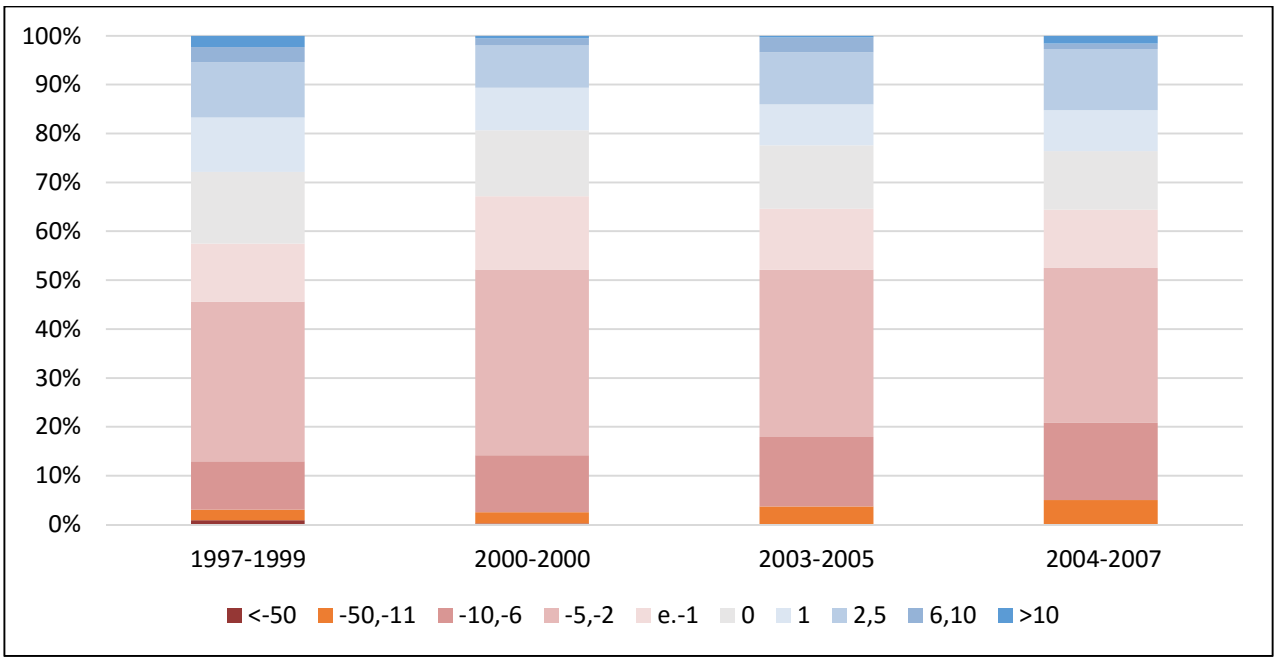
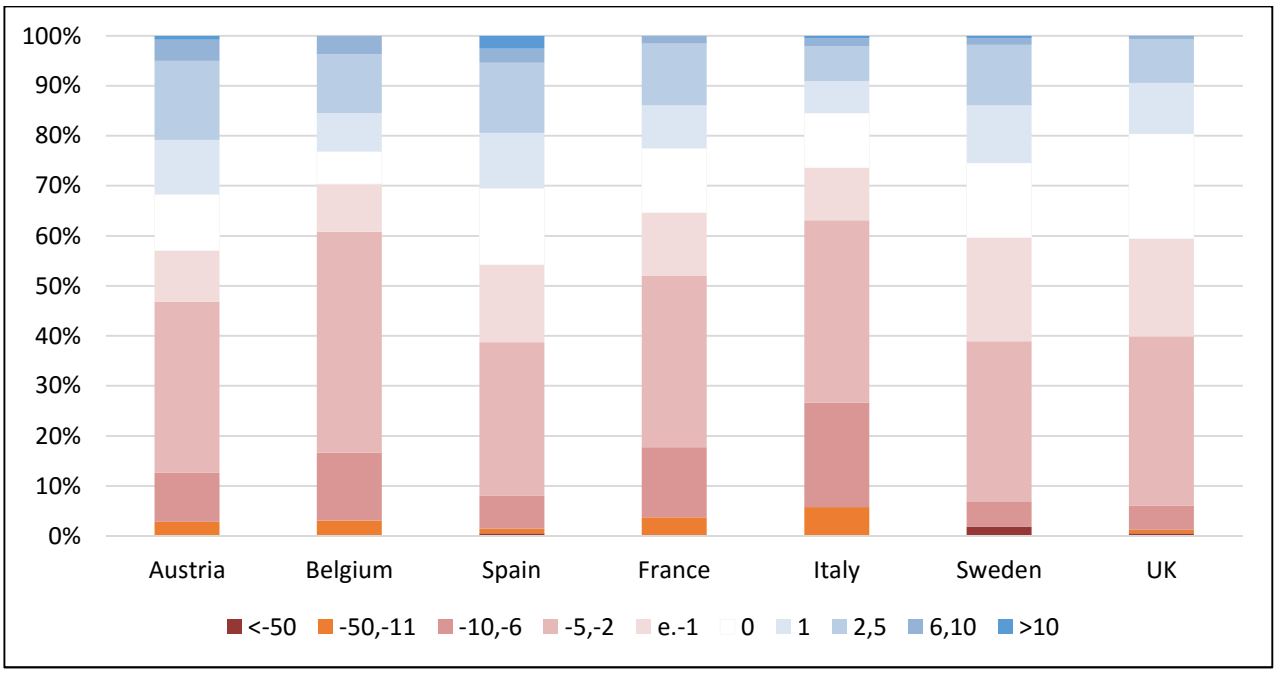


Figure A.8 reports the frequency distribution of inventor-author differences per PPP varies by country. The inventor-author difference is negative in more than 60% of the cases for Belgium, France and Italy, 60% for the UK and Sweden, and less than 60% for Austria and Spain. But in all cases, we never observe the share of negative values to go under 55% (Spain) or over 75% (Italy). While these differences may suggest scross-country differences in the way attribution rights are negotiated, they certainly hide a composition effect, as each country's portfolio of academic patents differ in the mix of technologies. Most notably, the result for the UK appears influenced by the absence, in our data for that countries, of Pharma&Biotech patents, while that for Spain reflects the high share of patents from technology fields other than the science-based ones.

Figure A.8 % frequency distribution of PPPs by difference between number of inventors and authors, and country of the academic inventors



A.3 Seniority and gender data: collection and descriptive statistics

For all authors of papers in the selected PPPs we went back to the WoS and collected all the publications by authors with the same surname and initials. We limited the search to the journals relevant for the technological fields of the patent in the PPP. We excluded from the search all authors with extremely common surnames (such as Smith, Muller, or Park). For all the other authors, we than calculated the year of their first publication and their stock of publication in each following year.

Information on author's gender could not retrieved as easily. For authors with at least one patent, the full first name could be obtained from the patent; for authors without patents we had access only to the initials, as derived from WoS. This forced us to substantial manual work, which was limited by the time and budget at our disposal. For this reason we decided to limit it to the most reliable PPPs, namely those where the paper and the publication exhibit a higher similarity index, mostly falling in the 3p90 and 2p95 PPS classes. Even in these sets, however, technical difficulties made it impossible to produce gender information for all observations.

We proceeded as follows. For each author in a PPS we downloaded at least one full paper from WoS and collected manually the first name. We matched the latter to the IBM GNR's library, which provided for each of them information on gender. Where possible, ambiguous cases were solved manually, either inspecting the publication or patent (if available) associated to the author or, for prolific authors, his/her webpage.[12]

[12]The *IBM Global Name Recognition* (IBM-GNR) *system* is a commercial product that performs various tasks, including the association of first names and surnames to one or (more often) several "countries of association" (CoAs). When fed with either a name or a surname or both, IBM-GNR returns a list of CoAs and two scores of interest, :

- "frequency", which indicates to which percentile of the frequency distribution of names or surnames the name or surname belongs to, for each CoA;
- "significance", which approximates the frequency distribution of the name or surname across all CoA.

More importantly for us, IBM-GNR also associates each first name to gender, expressed as the probability p that the name is masculine ($1-p$ feminine). In most cases, p is either equal to or higher than 90%, or equal/lower than 10%. We treated all these instances as unambiguous masculine or feminine first names. In all other cases, we proceeded to further collection of information and manual inspection of records. Typically, we associated the author to a country, based on affiliation information from his/her publications and the CoAs provided by IBM for the author's name and surname. We then looked on web resources on the

In sum, we retrieved gender information for 6242 authors, of which 33% turned out to be female. Their distribution across technologies and discipline is rather uneven, as reported in table A.6. Observations are the couples author-publications in the selected PPPs (for which gender was available); technologies are those of the patent in the PPP (all non science-based technologies are grouped under the "Other" label, due to low figures).

We notice that PPSs in Pharmaceuticals & Biotechnologies, besides being the most represented technology, is also the one with the largest share of female authors, followed closely by Chemicals & Materials. Engineering-based technologies (Electrical Engineering & Electronics as well as the Other technologies) all lag behind. As for Instruments, which is science-based but also includes many contributions from engineering, is in the middle. The overall distribution mirrors the one for the original set of patents, which implies that missing observations are distributed randomly.

male/female distribution of first names in the country (the classic example is Andrea, masculine in Italy and Spain and feminine elsewhere). In case of epicene names (such as Yannick in France or Terry in English) we left the information missing.

Table A.6 - % distribution of authors/publication couples, by technology of the patent in the PPP and author's gender (missing observation excluded)

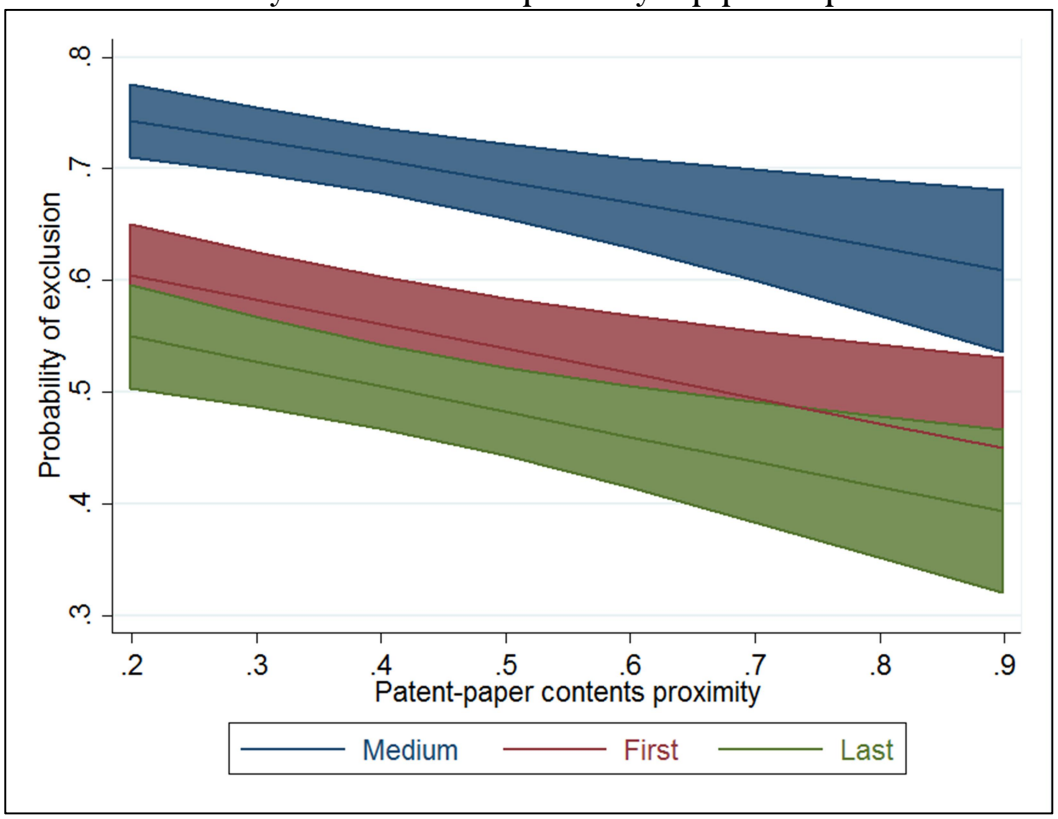
	Electrical eng. & Electronics	Instruments	Chemicals & Materials	Pharma & Biotech.	Other techn.	Total
Male	1,441	3,080	1,271	6,833	385	13,010
%	*85.06*	*74.22*	*68.89*	*62.11*	*84.06*	*67.94*
Female	253	1,070	574	4,169	73	6,139
%	*14.94*	*25.78*	*31.11*	*37.89*	*15.94*	*32.06*
Total	1,694	4,150	1,845	11,002	458	19,149
%	*100*	*100*	*100*	*100*	*100*	*100*

APPENDIX B – Exclusion regressions: logit estimates

Table B.1 – Exclusion regression: Logit estimates (3p90 and 2p95 PPS class) – Odds Ratios

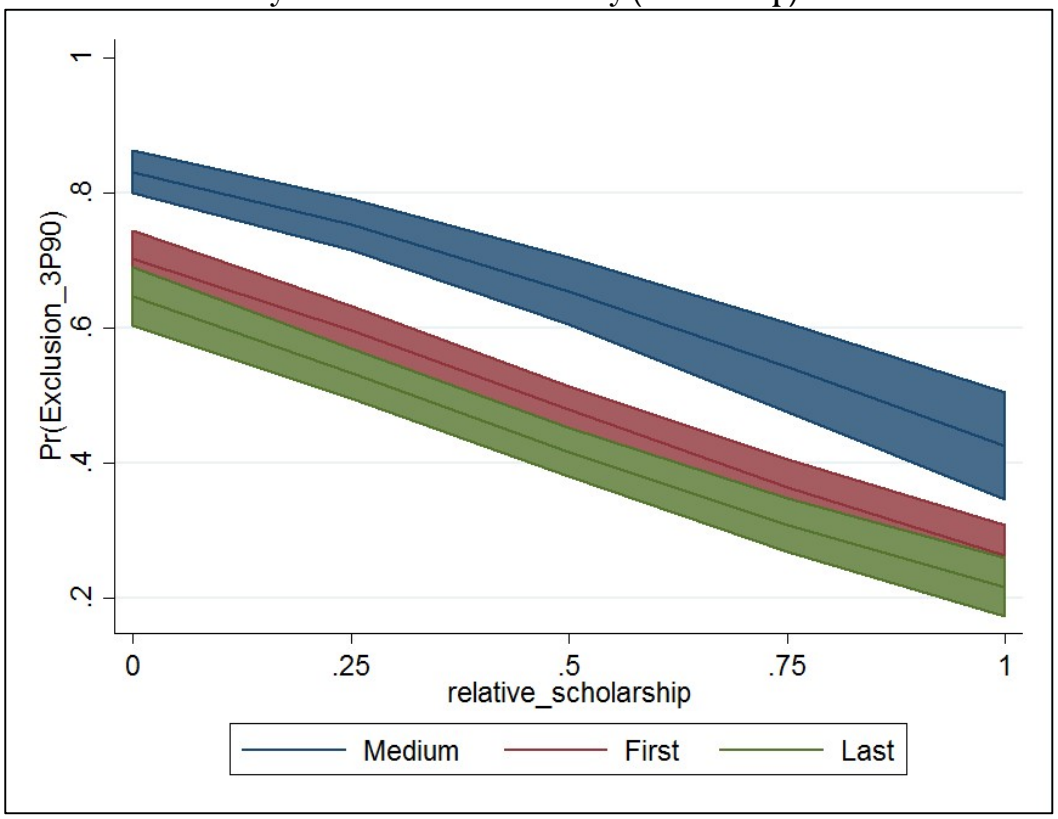
	3p90			2p95		
	(1)	(2)	(3)	(4)	(5)	(6)
First	0.46***	0.49***		0.41***	0.43***	
	(0.066)	(0.079)		(0.075)	(0.090)	
Last	0.35***	0.39***		0.34***	0.38***	
	(0.042)	(0.049)		(0.058)	(0.060)	
Female		1.47***			1.55***	
		(0.156)			(0.210)	
Female Middle			1.57***			1.55*
			(0.256)			(0.355)
Male first			0.37***			0.26***
			(0.067)			(0.069)
Female first			0.75			0.52**
			(0.178)			(0.152)
Male last			0.34***			0.31***
			(0.053)			(0.071)
Female last			0.47***			0.38**
			(0.138)			(0.160)
relative_scholarship	0.14***	0.15***	0.22***	0.14***	0.15***	0.24***
	(0.023)	(0.030)	(0.042)	(0.026)	(0.035)	(0.059)
most_junior	1.48***	1.53***	1.59***	1.55***	1.50***	1.51**
	(0.122)	(0.152)	(0.214)	(0.191)	(0.189)	(0.283)
proximity	0.34***	0.36**	0.57	0.41**	0.36*	0.76
	(0.110)	(0.154)	(0.374)	(0.185)	(0.210)	(0.715)
-1 year	1.07	1.14	0.80	1.25	1.14	0.65*
	(0.129)	(0.139)	(0.114)	(0.202)	(0.194)	(0.158)
0 years	0.77**	0.91	0.53***	0.83	0.90	0.50***
	(0.101)	(0.165)	(0.087)	(0.158)	(0.197)	(0.109)
+1 year	0.77	0.86	0.55***	0.73	0.77	0.45***
	(0.134)	(0.161)	(0.076)	(0.174)	(0.187)	(0.097)
+2 years	0.73*	0.71*	0.45***	0.78	0.72	0.42***
	(0.135)	(0.133)	(0.071)	(0.207)	(0.180)	(0.089)
n_authors	1.08***	1.09***	1.12***	1.08***	1.07***	1.11***
	(0.025)	(0.026)	(0.020)	(0.025)	(0.028)	(0.030)
n_inventors	0.84***	0.83***	0.82***	0.76***	0.75***	0.76***
	(0.025)	(0.027)	(0.021)	(0.022)	(0.029)	(0.032)
Instruments	1.53**	2.06**	2.39**	1.53*	1.70*	1.70
	(0.310)	(0.585)	(0.883)	(0.386)	(0.551)	(0.673)
Chemicals; Materials	1.85***	2.82***	3.12***	2.25***	2.65***	2.81**
	(0.412)	(0.838)	(1.181)	(0.679)	(0.896)	(1.174)
Pharmaceuticals; Biotech	1.96***	2.83***	2.29**	2.14***	2.60***	1.85
	(0.394)	(0.796)	(0.837)	(0.543)	(0.826)	(0.731)
Other technologies	1.41	2.05*	1.28	1.36	1.25	1.01
	(0.383)	(0.748)	(0.551)	(0.407)	(0.569)	(0.523)
Belgium	0.81	0.98	1.09	0.61**	0.64	0.79
	(0.115)	(0.191)	(0.218)	(0.138)	(0.184)	(0.200)
Spain	0.70**	0.98	1.02	0.59***	0.79	0.92
	(0.098)	(0.182)	(0.175)	(0.116)	(0.208)	(0.201)
France	0.82	0.93	0.90	0.63**	0.71	0.76
	(0.100)	(0.147)	(0.136)	(0.121)	(0.173)	(0.156)
Italy	0.62***	0.76		0.48***	0.55**	
	(0.089)	(0.131)		(0.104)	(0.154)	
Sweden	1.50	1.76*	1.91***	1.22	1.34	1.57*
	(0.383)	(0.519)	(0.464)	(0.369)	(0.492)	(0.425)
UK	0.74	0.56*	0.59*	0.61	0.43**	0.48*
	(0.170)	(0.172)	(0.171)	(0.221)	(0.183)	(0.188)
Constant	6.79***	2.76***	2.87**	11.03***	7.53***	6.54***
	(1.840)	(1.044)	(1.379)	(4.400)	(3.801)	(4.503)
Observations	14,244	9,141	3,872	8,070	5,075	2,109
Pseudo-R2	0.178	0.196	0.204	0.188	0.200	0.200
logL	-7323	-4734	-2036	-4062	-2589	-1105

Figure B.1 – Predicted probability of exclusion from patents, by author's position in the by-line and contents proximity of paper and patent



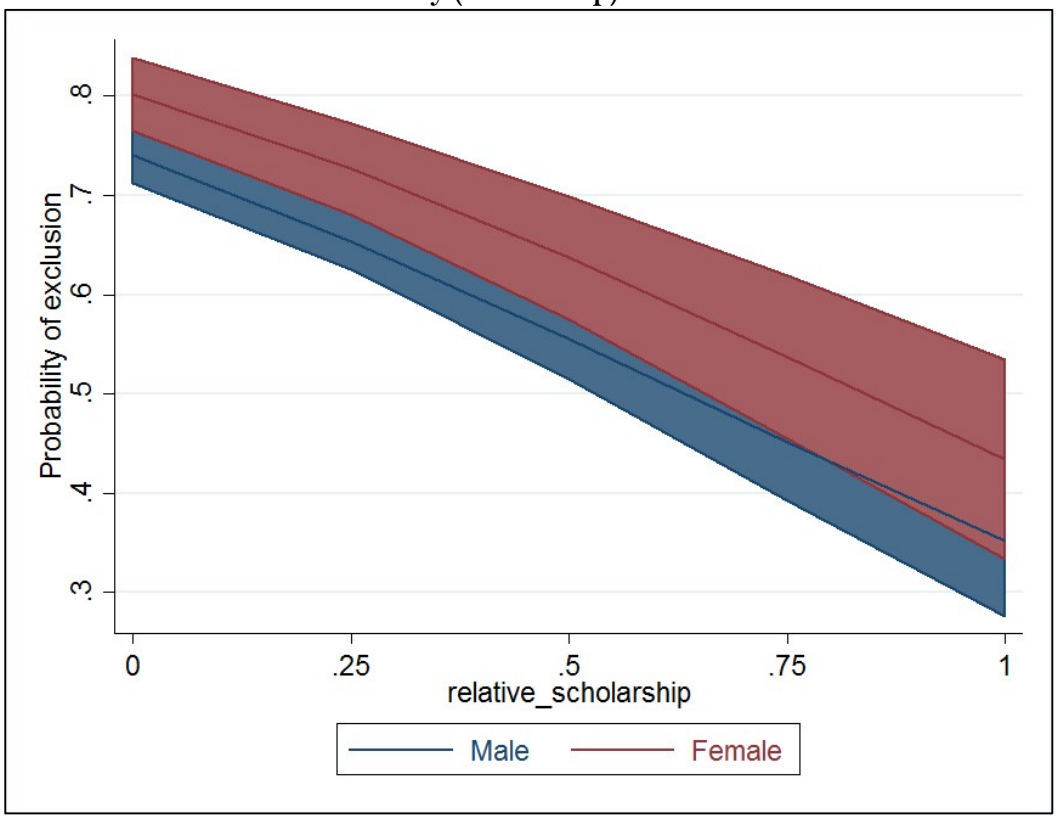
As from regression (1), table B.1 (95% confidence intervals) / Other variables at means

Figure B.2 – Predicted probability of exclusion from patents, by author's position in the by-line and relative seniority (scholarship)



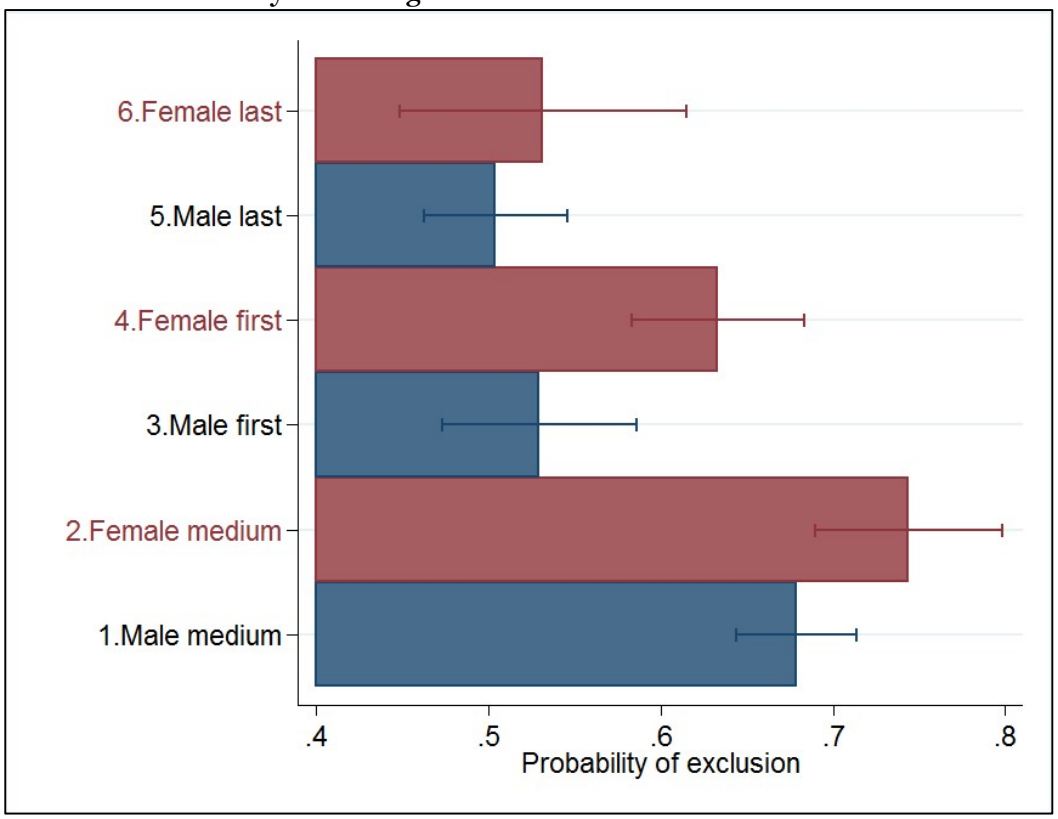
As from regression (1), table B.1 (95% confidence intervals) / Other variables at means

Figure B.3 – Predicted probability of exclusion from patents, by author's gender and relative seniority (scholarship)



As from regression (2), table B.1 (95% confidence intervals) / Other variables at means

Figure B.4 – Predicted probability of exclusion from patents, by author's position in the by-line and gender



As from regression (3), table B.1 (95% confidence intervals) / Other variables at means

APPENDIX C - Proof of Proposition 1

Consider norm 1. For *J*, the expected discounted payoff in the candidate equilibrium is:

$$\mu\frac{\left(R_1^J-I\right)}{1-\delta}+(1-\mu)(-I) \tag{C1}$$

For *J*, the only deviation which is possibly profitable is not joining the team at t=0 (nor in any subsequent period) thus getting a payoff 0 (notice that for *J* staying in the team after a deviation of S is never profitable, since he would get a negative payoff instead of 0). *J* will prefer to join the team iff:

$$\mu\frac{\left(R_1^J-I\right)}{1-\delta}+(1-\mu)(-I)\geq 0 \tag{C2}$$

i.e.

$$\delta\geq\frac{I-\mu R_1^J}{I-\mu I} \tag{C3}$$

For *S*, the expected discounted payoff in the candidate equilibrium is:

$$\frac{R_{N1}^S+v}{1-\delta} \tag{C4}$$

For *S* the profitable deviation is to play (*S*,*S*) in period 0, leading to the dissolution of the team from $t>1$ onwards. In this case, *S* gets R_1^S+v. S will prefer not to deviate iff:

$$\frac{R_{N1}^S+v}{1-\delta}\geq R_1^S+v \tag{C5}$$

i.e.

$$\delta\geq\frac{R_1^S-R_{N1}^S}{R_1^S+v} \tag{C6}$$

This strategy profile is thus an equilibrium if:

$$\delta \geq max\left\{\frac{I-\mu R_1^J}{I-\mu I}, \frac{R_1^S-R_{N1}^S}{R_1^S+v}\right\} \quad \text{(C7)}$$

Consider norm 2. For *J*, the expected discounted payoff in the candidate equilibrium is:

$$\mu\frac{\left(R_1^J+\frac{v}{2}-I\right)}{1-\delta}+(1-\mu)(-I) \quad \text{(C8)}$$

For *J*, the only deviation which is possibly profitable is not joining the team at *t*=0 (nor in any subsequent period) thus getting a payoff 0. *J* will prefer to join the team iff:

$$\mu\frac{\left(R_1^J+\frac{v}{2}-I\right)}{1-\delta}+(1-\mu)(-I) \geq 0 \quad \text{(C9)}$$

i.e.

$$\delta \geq \frac{I-\mu(R_1^J+\frac{v}{2})}{I-\mu} \quad \text{(C10)}$$

For *S*, the expected discounted payoff in the candidate equilibrium is:

$$\frac{R_{N1}^S+\frac{v}{2}}{1-\delta} \quad \text{(C11)}$$

For *S*, the possibly profitable deviation is to play *(S,S)* in period 0. In this case, *S* gets R_1^S+v. S will prefer not to deviate iff:

$$\frac{R_{N1}^S+\frac{v}{2}}{1-\delta} \geq R_1^S+v \quad \text{(C12)}$$

i.e.

$$\delta \geq \frac{R_1^S-R_{N1}^S+\frac{v}{2}}{R_1^S+v} \quad \text{(C13)}$$

This strategy profile is thus an equilibrium if:

$$\delta \geq max\left\{\frac{I-\mu(R_1^J+\frac{v}{2})}{I-\mu I}, \frac{R_1^S-R_{N1}^S+\frac{v}{2}}{R_1^S+v}\right\} \quad \text{(C14)}$$

Finally, consider norm 3. For *J*, the expected discounted payoff in the candidate equilibrium is:

$$\mu\frac{\left(\frac{v}{2}-I\right)}{1-\delta}+(1-\mu)(-I) \quad \text{(C15)}$$

For *J*, the only deviation which is possibly profitable is not joining the team at t=0 (nor in any subsequent period) thus getting a payoff 0. *J* will prefer to join the team iff:

$$\mu\frac{\left(\frac{v}{2}-I\right)}{1-\delta}+(1-\mu)(-I) \geq 0 \quad \text{(C16)}$$

i.e.

$$\delta \geq \frac{I-\mu\frac{v}{2}}{I-\mu I} \quad \text{(C17)}$$

For *S*, the expected discounted payoff in the candidate equilibrium is $\frac{R_1^S+\frac{v}{2}}{1-\delta}$. For *S* the profitable deviation is to play (S,S) in period 0. In this case, S gets R_1^S+v. S will prefer not to deviate iff:

$$\frac{R_1^S+\frac{v}{2}}{1-\delta} \geq R_1^S+v \quad \text{(C18)}$$

i.e.

$$\delta \geq \frac{\frac{v}{2}}{R_1^S+v} \quad \text{(C19)}$$

This strategy profile is an equilibrium if:

$$\delta \geq max\left\{\frac{I-\mu\frac{v}{2}}{I-\mu I}, \frac{\frac{v}{2}}{R_1^S+v}\right\} \quad \text{(C20)}$$

References

Agrawal, A., McHale, J., Oettl, A., 2013. Collaboration, stars, and the changing organization of science: Evidence from evolutionary biology, National Bureau of Economic Research.

Angrist, J.D., Pischke, J.-S., 2009, Mostly harmless econometrics: An empiricist's companion. Princeton University Press.

Azoulay, P., Stuart, T., Wang, Y., 2013. Matthew: Effect or fable? Management Science, 60(1), 92-109.

Banal-Estanol, A., Jofre-Bonet, M., Meissner, C., 2008. The impact of industry collaboration on research: Evidence from engineering academics in the UK, UPF Working Paper no1190, Barcelona, Spain.

Becker, G.S., Murphy, K.M., 1992. The Division of Labor, Coordination Costs, and Knowledge. The Quarterly Journal of Economics, 107(4), 1137-1160.

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/resource-library/editorial-policies/cse-policies/retreat-and-task-force-papers/authorship-task-force/cse-task-force-on-authorship/Council of Science Editors - last accessed: January 2015).

Bikard, M., Murray, F.E., Gans, J., 2015. Exploring tradeoffs in the organization of scientific work: Collaboration and scientific reward. Management Science, 61(7), 1473-1495.

Checchi, D., 1999. Tenure. An appraisal of a national selection process for associate professorship. Giornale degli economisti e Annali di economia, 137-181.

de Solla Price, D.J., 1963, Little science, big science... and beyond. Columbia University Press New York.

Ding, W.W., Murray, F., Stuart, T.E., 2013. From bench to board: gender differences in university scientists' participation in corporate scientific advisory boards. Academy of Management Journal, 56(5), 1443-1464.

Engers, M., Gans, J.S., Grant, S., King, S.P., 1999. First-author conditions. Journal of Political Economy, 107(4), 859-883.

Fehder, D.C., Murray, F., Stern, S., 2014. Intellectual property rights and the evolution of scientific journals as knowledge platforms. International Journal of Industrial Organization, 36, 83-94.

Frische, S., 2012. It is time for full disclosure of author contributions. Nature, 489(7417), 475-475.

Galison, P., Hevly, B.W., 1992, Big science: The growth of large-scale research. Stanford University Press.

Gans, J., Murray, F., 2014. Markets for Scientific Attribution, National Bureau of Economic Research.

Gans, J.S., Murray, F., 2013. Credit history: The changing nature of scientific credit, National Bureau of Economic Research.

Gans, J.S., Murray, F.E., Stern, S., 2013. Contracting over the disclosure of scientific knowledge: Intellectual property and academic publication, National Bureau of Economic Research.

Haeussler, C., Sauermann, H., 2013. Credit where credit is due? The impact of project contributions and social factors on authorship and inventorship. Research Policy, 42(3), 688-703.

Häussler, C., Sauermann, H., 2014, The Anatomy of Teams: Division of Labor and Allocation of Credit in Collaborative Knowledge Production. Available at SSRN.

Häussler, C., Sauermann, H., 2015. The Anatomy of Teams: Division of Labor and Allocation of Credit in Collaborative Knowledge Production. paper presented at DRUID conference - June 15-17, Rome.

Jin, G.Z., Jones, B., Lu, S.F., Uzzi, B., 2013. The reverse Matthew effect: catastrophe and consequence in scientific teams, National Bureau of Economic Research.

Jones, B.F., 2009. The burden of knowledge and the “death of the Renaissance man”: is innovation getting harder? The Review of Economic Studies, 76(1), 283-317.

Jones, B.F., Wuchty, S., Uzzi, B., 2008. Multi-university research teams: Shifting impact, geography, and stratification in science. science, 322(5905), 1259-1262.

Lissoni, F., 2013. Academic patenting in Europe: a reassessment of evidence and research practices. Industry and Innovation, 20(5), 379-384.

Lissoni, F., Montobbio, F., 2015. Guest authors or ghost inventors? Inventorship and authorship attribution in academic science. Evaluation review (forthcoming), 0193841X13517234.

Lissoni, F., Montobbio, F., Zirulia, L., 2013. Inventorship and authorship as attribution rights: An enquiry into the economics of scientific credit. Journal of Economic Behavior & Organization, 95(0), 49-69.

Ljungberg, D., Bourelos, E., McKelvey, M., 2013. Academic inventors, technological profiles and patent value: an analysis of academic patents owned by Swedish-based firms. Industry and Innovation, 20(5), 473-487.

Maraut, S., Martínez, C., 2014. Identifying author–inventors from Spain: methods and a first insight into results. Scientometrics, 101(1), 445-476.

Mejer, M., 2013. Academic Patenting in Belgium: Methodology and Evidence, iCite Working Paper 2013 - 003, Solvay Brussels School of Economics and Management, Universite Libre de Bruxelles.

Merton, R.K., 1968. The Matthew Effect in Science. Science, 159(3810), 56-63.

Merton, R.K., 1988. The Matthew Effect in Science, II: Cumulative Advantage and the Symbolism of Intellectual Property. Isis, 79(4), 606-623.

Mowatt, G., Shirran, L., Grimshaw, J.M., Rennie, D., Flanagin, A., Yank, V., MacLennan, G., Gøtzsche, P.C., Bero, L.A., 2002. Prevalence of honorary and ghost authorship in Cochrane reviews. Jama, 287(21), 2769-2771.

Pezzoni, M., Lissoni, F., Tarasconi, G., 2014. How to kill inventors: testing the Massacrator© algorithm for inventor disambiguation. Scientometrics, 1-28.

Rennie, D., Flanagin, A., Yank, V., 2000. The contributions of authors. Jama, 284(1), 89-91.

Simcoe, T.S., Waguespack, D.M., 2011. Status, quality, and attention: What's in a (missing) name? Management Science, 57(2), 274-290.

Stephan, P., 2012, How economics shapes science. Harvard University Press.

Stummer, C., Günther, M., Backs, S., 2013. A survey on academic patents at Austrian universities: Methodology and initial results, WP 2013-01, Chair of Innovation and Technology Management, Bielefeld University.

Thursby, J.G., Thursby, M.C., 2005. Gender patterns of research and licensing activity of science and engineering faculty. The Journal of Technology Transfer, 30(4), 343-353.

Whittington, K.B., Smith-Doerr, L., 2005. Gender and commercial science: Women's patenting in the life sciences. The Journal of Technology Transfer, 30(4), 355-370.

Wooldridge, J.M., 2010, Econometric analysis of cross section and panel data. MIT press.

Wuchty, S., Jones, B.F., Uzzi, B., 2007. The increasing dominance of teams in production of knowledge. Science, 316(5827), 1036-1039.