Paper to be presented at the

DRUID Society Conference 2014, CBS, Copenhagen, June 16-18

Lifting the veil on patents and inventions

Oliver Alexy
Technische Universität München
TUM School of Management
o.alexy@tum.de

Paola Criscuolo
Imperial College London
Business School
p.criscuolo@imperial.ac.uk

Ammon Salter
Imperial College London
Business School
a.salter@imperial.ac.uk

Dmitry Sharapov
Imperial College London
Business School
dmitry.sharapov@imperial.ac.uk

Abstract

Since patent data became accessible in the 1980s, we have known that research using this data?while providing tremendous opportunities?rests on important assumptions about how patents are actually generated by firms. It is well known that firm-level selection processes shape the likelihood that firms decide to patent or not an invention. What is unknown is to what extent these processes leave the results of work using patent data at risk of being distorted by sample selection bias. To understand the magnitude of this bias, we replicate two important prior studies using data from a novel, proprietary dataset, which contains more than 35,000 invention disclosures made by inventors within a single firm, only some of which went on to be patented. We find strong indications for the presence of significant selection bias in patent studies in examining the variance of creative outcome distributions and the impact of past experience in subsequent inventions. We highlight what the nature of this bias may mean for our current body of knowledge, and provide suggestions of how this issue should be addressed in future research.

LIFTING THE VEIL ON PATENTS AND INVENTIONS

ABSTRACT

Since patent data became accessible in the 1980s, we have known that research using this data—while providing tremendous opportunities—rests on important assumptions about how patents are actually generated by firms. It is well known that firm-level selection processes shape the likelihood that firms decide to patent or not an invention. What is unknown is to what extent these processes leave the results of work using patent data at risk of being distorted by *sample selection bias*. To understand the magnitude of this bias, we replicate two important prior studies using data from a novel, proprietary dataset, which contains more than 35,000 invention disclosures made by inventors within a single firm, only some of which went on to be patented. We find strong indications for the presence of significant selection bias in patent studies in examining the variance of creative outcome distributions and the impact of past experience in subsequent inventions. We highlight what the nature of this bias may mean for our current body of knowledge, and provide suggestions of how this issue should be addressed in future research.

KEYWORDS

Patent data; patenting propensity; replication; sample selection bias;

INTRODUCTION

In a recent paper, Gittelman (2008) states that to assess the value of patents as a useful indicator of innovative outputs of individuals and firms, one needs to gain a better understanding of the context surrounding firms' and inventors' decisions to use patents as a protection mechanism. In her words: "If we do not understand the institutional, organizational, and strategic context in which patents are created, we risk misusing the data, misinterpreting our results and in many cases attributing causality to covariance" (p. 21). Thus, only if we know that patent data is correctly interpreted can we believe in the validity of studies using patent data. Crucially, patent data, and its use as an indicator of innovative activity, suffers from an important limitation: *selection bias*—an issue that has been acknowledged for a long time (e.g., Griliches, 1984), but has not yet been comprehensively addressed. Simply put, we know that not all inventions are patented, but we use patent data as a reasonable proxy for invention itself.

Existing work has attempted to quantify and ameliorate these shortcomings by drawing on the information contained in patents. For example, prior work has demonstrated that there are significant differences between firms operating in complex-product industries, such as computers, semiconductors, and telecommunications, where hundreds of patents are needed to protect the IP contained in a single product, and those in discrete-product industries, such as pharmaceuticals and chemicals, where a smaller number of patents protect the IP embedded in a single product (Hall & Ziedonis, 2001). These contributions have enriched our understanding of the differences in the propensity to patent and in patenting practices between industries, which have also been addressed by work using survey-based evidence (Arundel & Kabla, 1998; Cohen et al., 2000) and by historians of technology, using databases of innovations (Moser, 2005, 2012).

At the same time, our knowledge of how patent applications are decided upon, and

then generated, inside organizations, remains relatively sparse. Notable survey-based work on individual inventors aside (e.g., Giuri et al., 2007), we know surprisingly little about what is really going on inside organizations *before* the patent application. This is particularly crucial because this work may challenge some of the data interpretations brought forward by scholars using patents. For example, co-patenting has frequently been proposed as a measure of inter-organization collaboration. Yet, in the PATVAL survey of inventors, significant collaboration with other organizations was reported in about 15% of cases, yet co-patenting only in 3% (Giuri et al., 2007). Thus, it is vital that we improve our understanding of what is going on inside organizations to be able to better sense the extent to which our current body of patent-based research is valid. This richer understanding will help us ensure such data provides a strong foundation for building theories and for advising practicing managers about the nature of invention and innovation.

Accordingly, we present empirical evidence analyzing and quantifying the magnitude of the selection bias that is characteristic of patent data (Jaffe & Trajtenberg, 2002), and that, if not appropriately accounted for, might significantly bias the results of studies based on patent data (Gittelman, 2008). Specifically, to assess the extent of this bias we exploit a unique dataset of *all* invention disclosures made by the employees of a company operating in a complex-product industry[1]. This dataset contains information on inventions which did not pass the novelty step, those that were novel but were not considered useful by the firm; and those that were eventually submitted to a patent office as a patent application.

Therefore, this dataset allows us to assess whether, and to what degree, *selection bias* may exist in patent data studies. Specifically, the way in which we will assess the presence of selection bias and the impact that this might have on the statistical inference drawn from

[1] To help mask the identity of our industrial partner, we do not report the sample period to which our patents and inventions data relate to and the exact number of inventions developed by this organization. Further, we report the total number of the firm's inventions and patents after reducing it by a random percentage value, while using the full population in our analysis.

patent based studies is by replicating two studies examining the sources of technological breakthroughs using our dataset: one by Singh and Fleming (2010) on the impact of collaboration on the variance of creative outcome distributions and the other by Audia and Goncalo (2007) the impact of inventors' past success in their future creativity performance. To do so, we will first estimate the models presented in these two studies using only the subset of inventions in our sample which have been granted a patent by the US Patent Office (USPTO). This first step aims to assess whether the findings obtained in these studies hold for the sample of patents granted to our firm. Second, we will estimate the same models controlling for selection bias derived from the inability of prior work to observe non-patented inventions.

Our results suggest that selection bias *does* matter. In short, we show how selection bias affects the impact of some important drivers of the heterogeneity of patented inventions. In particular, this bias leads to an overestimation of the likelihood of inventor teams generating low quality innovative outcomes, and of the influence of an inventor's prior innovative success on both the likelihood of patenting and the number of explorative ideas. However, we do not find evidence of selection bias in models explaining high quality patents. In our conclusions, we explore the implications of these findings for current research and suggestion potential corrective measures to help ensure more valid patent work in the future.

BACKGROUND

There is little doubt about the value of patent data; it has been essential to the progress of the field of innovation studies over the past 30 years. Given the ease at which it can be attained, it is unsurprising to find more than 17,000 papers when simply searching for "patent data" and "innovation" on Google Scholar. In recent years, there has been a surge of studies drawing on the NBER database and other online sources to understand invention and

innovation through the window of patenting behavior. Indeed, these studies using patent data have become the cornerstone of our understanding of how firms can support innovation. For example, the most cited paper in *Administrative Science Quarterly* on innovation since 2000 is the Ahuja (2000) study of collaboration and patenting. Over time, these patent studies have used larger datasets and increasingly more complex analytical approaches to examine the information contained in patents. They have also increasingly been used to link patents to a range of other managerial choices and behaviors, helping to unlock a wide range of insights about what firms know and can do.

At the same time, we have always known that patent data is far from perfect. When its computerization around the beginning of the 1980s sparked huge interest (see e.g., Griliches, 1984 for an overview), researchers were clear in stating potential issues of selectivity or differences between R&D, patenting, and innovation—in fact, those were some of the most crucial questions tackled. In addition, these early authors pointed out numerous opportunities for research on a never-before-attainable large-scale dataset, in whose creation and refinement they were eventually instrumental (Griliches, 1984; Hall et al., 2002).

Since then, although considerable progress has been made in innovation studies and related fields thanks to patent studies, the problems that come with using (exclusively) this data have been relegated to a few symbolic cites to those early works. In some respects, patent data has tended to increasingly be 'reified', treated as if it was an unambiguous, direct measure for invention *and* innovation. The reification of patent data may sow confusion between these imperfect measures and the reality of innovation itself. In doing so, there is a danger that researchers fall into what Alfred North Whitehead called the 'fallacy of misplaced correctness' (Whitehead, 1925, p. 51). In the following, we focus one aspect for why this approach may be problematic: selection bias.

Selection Bias: Issues with Assessing the Quality of Inventions

Although scholars have been able to considerably increase our understanding of the institutional, technological, and legal contexts that shape firms' strategic reasons for patenting, we still know relatively little about what affects the organization's decision to patent an invention. In particular, we do not know how the selection process inside organizations concerning the decision to patent an invention affects any statistical inference drawn from studies using patent data. First, not all inventions are patentable due to explicit legal exclusion. Second, not all inventions are patented because firms may decide to protect their innovations by alternative appropriability methods, for example by keeping the invention secret. Third, and most importantly, many inventions are not patented because they do not pass an (internal-defined) novelty step and/or are not deemed useful to the inventor or the firm. As a result, when using patent data we are only able to observe a reduced sample of inventions. If the main purpose of a study is to examine the inventive activity of either an individual or a firm, then patent data suffer from an important form of selection bias.

As pointed out by Gittelman (2008), the presence of selection bias will affect in particular those studies which examine the quality of inventions as the researcher will be dealing with a significant level of unobserved heterogeneity. It might be of pressing concern when analyzing the variance in innovative outcomes (Girotra et al., 2010; Singh & Fleming, 2010; Taylor & Greve, 2006), as the researcher may be unable to observe the low end of the quality distribution. At the same time, it will also affect those studies which examine how inventors' past experience shapes their innovative performance as less successful events are not taken into account (Audia & Goncalo, 2007; Conti et al., 2014)—such information would need to be gathered through additional sources different from patent data (Giuri et al., 2007). Finally, firms' or individuals' propensity to patent cannot be reliably observed from public data alone (de Rassenfosse & van Pottelsberghe de la Potterie, 2009; Fontana et al., 2013). Although using past patent experience might be a useful proxy for inventor experience, it is

an imperfect one, as it assumes a close match between the number of inventions and patents of an individual. It could be that some individuals are prolific inventors, but patent only rarely due to their preference for quality and/or they lack of resources to pay the costs of patenting. There may also be individuals who patent all their inventions, regardless of their quality, simply because they have additional resources to funding their patenting efforts.

More formally, based on the exposition in Stolzenberg and Relles (1997), suppose that Y_1 is the dependent variable and Y_2 a binary indicator for whether or not the invention is patented. Y_1 is only observed for those inventions which have been patented (selected cases) while is missing for other cases (censored cases). The *outcome regression model* with only one independent variable can then be written as

$$Y_1 = \beta_0 + X\beta_1 + \sigma\varepsilon \quad (1)$$

where X is the independent variable explaining the outcome variable Y_1, and $\sigma\varepsilon$ is the regression error terms, where σ is a scalar and ε is ~Normal(0,1).

The *selection equation* for the same data can then be defined as:

$$Y_2 = \alpha\boldsymbol{Z} + \delta \quad (2)$$

where $\boldsymbol{Z}$ is a vector of independent variables which explain the likelihood of patenting an invention and δ is the error term which is ~Normal(0,1).

Y_1 is observed only if Y_2 is greater than T (the selection threshold). For a given value of $\boldsymbol{Z}$, the probability of selection depends on the value of T, α, and δ. If α is equal to zero, then selection is random and, as a result, the sample used to estimate equation 1 is smaller. However if the selection is non-random, by estimating equation 1 one would introduce a bias in the coefficient estimate of X. In particular, Heckman (1976) derived the conditional expectation of Y_1 *given that* Y_1 *is observed*, as:

$$\mathrm{E}(Y_1|Y_2 > T) = \beta_0 + X\beta_1 + \sigma\rho_{\varepsilon\delta}\lambda(T - \alpha\boldsymbol{Z}) \quad (3)$$

Where $\rho_{\varepsilon\delta}$ is the correlation between ε and δ, and λ is the inverse of the Mills' Ratio

which is equal: $\lambda(T - \alpha \boldsymbol{Z}) = \phi(T - \alpha \boldsymbol{Z})/[1 - \Phi(T - \alpha \boldsymbol{Z})]$ with $\phi(\cdot)$ and $\Phi(\cdot)$ being the standard normal probability density and standard normal cumulative density functions. Therefore if selection is random and $\sigma\rho_{\varepsilon\delta} = 0$, then β_1 can be consistently estimated using the sample of patents. However if selection in not random and $\sigma\rho_{\varepsilon\delta} \neq 0$, estimating equation (1) without including the inverse Mills' ratio will produce biased estimates because the model will suffer from omitted variable bias.

In the presence of selection, equation 3 can be estimated using the Heckman sample selection model, in which the estimate of the inverse Mills' ratio from a probit regression explaining the likelihood of an invention being protected with a patent is then used in the model explaining invention quality. While β_1 is identified in the Heckman procedure even if $\boldsymbol{Z} = X$, due to the nonlinearity of the inverse Mills' ratio, for more precise estimates of coefficients in β_1, it is useful to include in $\boldsymbol{Z}$ exogenous variables which affect the likelihood of an invention being protected with a patent but do not affect the outcome variable Y_1.

DATA AND METHODS

Sample

In this study, we exploit a unique dataset of invention disclosures made by all employees working for a large multinational company operating in a complex-product industry which we will call Venus for reasons of confidentiality. A total of 35,144 inventions were submitted by inventors during this sample period. As is common in many large technology-based companies, all employees in Venus, whether working with external parties or not, are requested to document their inventions and to store this information in an IT system so that these can be subsequently evaluated by a team of patent engineers and experts. The main objective of the evaluation team is to decide whether the invention contains a novelty step and whether it is useful to the firm either by potentially being incorporated into a

product or service or as a means of production or service provision. The evaluation process can result in four different outcomes:

1. The invention is not novel or does not contain an inventive step; therefore Venus does not acquire the rights to this invention—the invention is thus "given" to the inventor.
2. The invention contains an inventive step but is not currently considered to be useful for the company, so Venus decides not to seek patent protection but keeps the rights to this invention as it might be patented in the future.
3. The invention has been judged to be novel and useful and Venus proceeds with applying for patent protection in one or several patent offices.
4. The invention is considered novel and useful but Venus decides to keep the invention secret.

Almost half of the inventions in our sample are considered as not new or obvious (category 1) and are thus of low quality from the perspective of the firm. Categories 2 and 4 are interesting as they might represent high quality inventions which the company decides not to protect with a patent. Thus, if there is a selection bias in current studies examining valuable innovations, then this might also stem from these types of inventions. However, only a very small proportion of inventions are kept secret in Venus (less than 1%), but almost 10% of inventions fall under category 2 above, i.e. they do contain an inventive step but they are not yet deemed useful for Venus. Eventually, only 15% of these inventions are patented by the firm, so the vast majority of these inventions, although potentially of high quality in terms of novelty, will never appear in patent databases. The remaining inventions (category 3) are patented.

Data Preparation

To be able to test for the presence of a selection bias in the studies, we decided to replicate, we adopt a two-step approach à la Heckman described above where first we

estimate the probability that an invention has been protected with a granted patent using the entire sample of invention disclosures and then we estimate the model explaining the main outcome variable using only the subset of inventions which have been patented taking into account the estimated probability that an invention is protected with a granted patent, as captured by the inverse Mills' ratio. As mentioned above, to precisely estimate these two models, however, we need a variable which explains the selection process, i.e. why the evaluation team has decided to patent the invention, but which does not influence the outcome variable used in the outcome regression model.

To better understand the evaluation process, we carried out 20 exploratory interviews with inventors, patent engineers, experts, and managers of legal and IP departments, as well as taking part in numerous formal and informal meetings. Further, one of the authors spent 40 days observing a team of patent engineers dealing with the evaluation of new invention disclosures and the maintenance of Venus' patent portfolios.

Through these interviews and observations, we found out that one of the main reasons explaining why novel inventions are not considered useful and thus are not patented is because they are originated from outside a formal project. As in many other organizations, inventors in Venus work on pet or bootleg projects (Criscuolo et al., 2014), i.e. projects which are non-programmed innovation efforts and not officially authorized by the organization. Engagement in these activities can often result in inventions which are then disclosed to and evaluated by the organization. Although novel, these inventions do not always fit with the main strategic and technological priorities of the firm and may not be easily or directly incorporated in the company products. As a result, inventions resulting from creative efforts outside formal projects are often novel, but not useful (category 2).

Therefore, to improve identification in the two-equation system, we used a dummy variable which is equal to one if the invention under evaluation originated from an official

R&D project. Unfortunately, this information is available only for a subset (28%) of our invention disclosures, which resulted in 1,910 USPTO granted patents. For the first-step selection model, we further control for the technology area of the invention using Venus's internal technology classification. Each invention disclosure is classified by the responsible patent engineer in one or multiple 4-digit technology classes.

ANALYSIS

To assess the extent of the selection bias introduced by the unobservability of inventions which are not patented, we replicated the results of two recent studies that used patent data to examine what drives the emergence of technological breakthroughs by focusing on the impact of collaboration among teams of inventors and of past creative outcomes.

The first study, by Singh and Fleming (2010), focuses on team size and the resulting collaboration among inventors as a source of inventions with extremely high quality, and also assesses the effect of this factor in explaining the occurrence of inventions of extremely low quality. Their main argument is that teams of inventors are more likely to discover breakthroughs because of their greater diversity of knowledge which in turn leads to higher combinatorial opportunity. But teams of inventors are also less likely to produce low quality inventions because of the greater and more rigorous process of ideas selection. As the main mechanism through which team size affects the generation of technological breakthrough is the diversity of the team background, the authors postulate that the technological experience of the team of inventors and the size of their network of indirect collaborators mediate the quality of their innovative efforts. One of key features of this study is that it tries to "examine the entire distribution of creative outcomes" (p. 41), but by using patent data it cannot fully capture the entire distribution as low quality inventions, those that are not patented, are missing from the lower tail of the distribution. In other words, the distribution must be left

truncated.[2]

In particular, the sample selection problem which might bias the estimates of this study can be formally described as follows. In the *outcome regression model* (equation 1) Y_1 is invention quality and *X* is *team affiliation* and in the *selection equation* Y_2 (equation 2) is the likelihood of patenting an invention. Note that under the assumptions made by these authors α_1 and β_1 expected to be positive. Thus we can rewrite equations 1 and 2 as:

$$Quality = \beta_0 + \beta_1 Team + \sigma\varepsilon$$

$$Patenting = \alpha_1 Team + \delta$$

If it is true, as it is often assumed in literature, that unpatented inventions did not pass the threshold of novelty necessary to deserve to be patented, then the sample of selected cases is under-representing single inventor inventions and the coefficient of *team affiliation* will be biased downward. In other words, if we observe an invention by a single inventor among the patented inventions, there might be other reasons − captured in the error term δ – which could explain why this invention was patented which could also explain its quality. As a result the *cov*(*Team*,δ)<0 and the error terms δ and $\sigma\varepsilon$ are likely be correlated meaning that also *cov*(*Team*, $\sigma\varepsilon$)<0. If we re-write the quality equation to include the error term from the selection equation we get:

$$Quality = \beta_0 + \beta_1 Team + \beta_2\delta + \sigma\varepsilon$$

If we assume that β_2 is positive, then ignoring δ and attributing all its impact to the *Team* variable will have a negative effect on the magnitude of β_1, i.e. β_1 will be downward biased. If, instead, we consider that not all novel inventions are patented (for example, some

[2] Additionally, the variables for the size of the inventor's networks of direct and indirect collaborators and their technological experience are likely to be measured with error if only patented inventions are considered. In results available on request from the authors, we compare the effects on the likelihood of producing high and low-quality patents of these variable when they are calculated using the subset of inventions which are patented versus the whole set of inventions. The results show that the effect of average experience on the likelihood of generating high-quality patents is significantly under-estimated, while the effects of both average and joint experience on the likelihood of generating low-quality patents is significantly over-estimated when only patent data is used to construct these variables. This implies that the measurement errors in these variables are 'non-classical' and lead to biased and inconsistent estimates.

novel inventions may not help the company pursue its objectives), then the sample of selected cases is under-representing inventions discovered by team of inventors and the coefficient of *team affiliation* will also be biased downward.

A similar problem affects the second study, by Audia and Goncalo (2007). In this paper, the authors focus on an inventor's past experience in successful creative efforts as a driver for the subsequent generation of explorative ideas. Audia and Goncalo posit that inventors with a strong track record of producing inventions might be better able to develop more inventions because they become faster and more efficient at generating new ideas. However, past experience becomes an obstacle for the generation of explorative ideas because successful inventors tend to apply the same heuristics used in the past and to draw from familiar knowledge sets. The negative impact of past success on the development of divergent ideas is, however, moderated by the presence of other inventors involved in the creative efforts. By relying only on those creative ideas which were patented, this study is unable to fully measure an inventor's past and current experience, as it disregards the creative endeavors which did not result in a patent. One could assume, as the authors do, that the selection is random (i.e. α in equation 2 is equal to 0) and the only consequence is that one would estimate equation 1 with a smaller sample. However, this might not be true as there might be a relatively high level of unobserved heterogeneity at the inventor level in their ability to generate inventions which are then patented as well as in their ability to develop inventions which are divergent from past innovative efforts. In this case, it is difficult to predict the direction of the bias we will assume instead that we expect not to find any selection bias as predicted by Audia and Goncalo.

Replicating Singh and Fleming's study

We start by reporting the results of our estimations of the Singh and Fleming's models using the subset of USPTO patents for which we have information on whether the invention

is linked to an official project. Results for the sample of USPTO patents are included in the appendix (see Table A1). To enable the reader to compare our results to the ones reported in the Singh and Fleming's paper, we have kept the same variable labels. The main independent and control variables were computed following the description provided in the paper and using only the USPTO patent applications which were granted. However, as our invention disclosures dataset uniquely identifies inventors, we included also the inventions and corresponding patents granted made by inventors residing outside the US.

Regarding the dependent variable, to determine whether a patent is in the top 5% in terms of frequency of forward citations, we compared the citations received by the focal patent with those received by patents applied in the same year and in the same primary 3-digit IPC technology classification. To derive the citation frequency and the frequency distributions per year and technology class, we used citations made by patents applied for in all patent offices, rather than only within USPTO citations.[3] Although our models use data from a more recent period, all the variables capturing different aspects of the team of inventors were derived using invention disclosures since Venus' inception. We have, however, not used the data for this longer time period in our regressions as there are relatively few observations during this earlier period.

Before reporting the estimation results, we looked at whether inventions by single inventors tend to be of lower quality than those with multiple inventors using the sample of 35,144. More than half of the inventions in our sample are generated by individual inventors and 62% of these fall in our lowest quality category, which seems to confirm Singh and Fleming's main expectation. As the number of inventors in the team increases, the proportion of inventions considered prior art decreases, while the number of patented inventions increases. However, we cannot observe a systematic trend as it seems that teams with more

[3] We have estimated models using the within-USPTO citation frequencies, but we decided not to report them here as the results were not consistent with the ones obtained by Fleming and Singh as the team size variable was not always significant.

than four inventors produce marginally more inventions of relatively lower quality. The proportion of novel but not useful inventions, however, does not change dramatically as the size of the team increases.

Table 1 reports the summary statistics for the variables used in our main models. As in the sample used in the Singh and Fleming's paper, the experience and network size variables display high skew with network size displaying a maximum value of over several hundred inventors and standard deviation of 63. The percentage of granted patents which received zero forward citations is equal to 9% in our sample, which is consistent we what found in the sample used by Singh and Fleming (7%). However, the proportion of patents in the top 5% of the distribution of forward citations is much higher in our sample (18%) than in the sample used by Singh and Fleming (5%). In Table 2, we report the correlation matrix among the variables used in the regressions.

--- INSERT TABLES 1 and 2 HERE---

Table 3 contains the coefficient estimates of the main models of the Singh and Fleming's paper (see their Table 6). We estimated also the negative binomial models which regress the two mediation variables (experience diversity and network size) and found that team's size had a positive and significant impact in explaining both variables (see Tables 2A in the appendix). According to the estimates for Model 1, patents with more than one inventor are 11.3% more likely to be in the 95^{th} percentile of the citations than patents with only one inventor. This effect is significantly smaller than what was found by Singh and Fleming (28%), but it is significant at the 1%-level. We also found that team affiliation affects the likelihood of poor outcome patents. The estimate of the *team* variable in Model 6 indicates that patents granted to teams of inventors are 6% less likely to receive zero citations

than those granted to single inventors.[4] Models 2-5 and Models 7-10 include the mediator variables experience diversity and network size and their interactions. In Model 2, we found that experience diversity, although significant, has the opposite sign than what predicted by Singh and Fleming. Network size is only significant and positive in Models 4 and 5. Therefore, we could not replicate the mediation effects of these two variables in our sample of breakthrough innovations. Similarly, experience diversity is not significant in Model 2 and network size is only significant and negative in Model 10.

--- INSERT TABLE 3 HERE ---

In Table 4, we report the estimates of the same models but controlling for the possible selection bias introduced by not including inventions which were not patented. To this end, we first estimated a probit model to predict the likelihood that an invention was protected with a patent granted by the USPTO using the entire sample of inventions for which we have information on whether the invention was stemming from an R&D project or not. We included as explanatory variables in this first stage model a dummy variable equal to one if the invention was generated in an R&D project, another dummy variable equal to one if the invention was the result of collaboration among multiple inventors, the logarithm transformation of the average number of previous inventions for the team of inventors, the logarithm transformation of the number of past inventions invented by the same team, and 15 one-digit technology class dummies. It is interesting to report that the coefficient of the R&D project variable was positive and significant at one percent confirming our expectations. From this first stage model, we derived the inverse Mills' ratio which we included in the second stage logit models reported in Table 4.[5]

[4] Similar results in terms of significance of the coefficient estimates are found for the larger sample of patents reported in Table 1A in the appendix. However effect sizes are much smaller: 5.6% and 4.1% for high and poor quality outcomes, respectively.

[5] We also estimated the same models using the heckprob command in STATA which fits maximum-likelihood probit models with sample selection and produces the correct standard errors that control for the two-step

--- INSERT TABLE 4 HERE ---

The inverse Mills' ratio is never significant in Models 1-5, but it is significant in Models 6-10 of Table 4. This suggests that the selection bias affects only the estimations of the poor quality outcomes and not for the high quality ones. This implies that by not including all the inventions generated by scientists and engineers in Venus, researchers would not underestimate the effect of team affiliation on the likelihood of generating technology breakthroughs, but they will overestimate the effect of team affiliation on the likelihood of generating poor innovative outcomes. Indeed, the effect of team affiliation is now equal to 3.7% according to the estimates of Model 6, instead of 6% and this difference is statistically significant at the 1% level.

However, the only way to correctly estimate the real effect of team affiliation on the likelihood of generating poor innovative outcomes is by considering all inventions and estimating a logit model with a dependent variable equal to 1 if the invention was not patented (*Notpat*), i.e. it was neither novel nor useful. Similarly, one could estimate the likelihood of an invention being prior art (*Priorart*), i.e. not novel, or of an invention being novel but not useful (*Notuseful*). These regressions are reported in Table 5. Being part of a team has a negative and significant effect on the likelihood of an invention not being patented and also of an invention not being novel. However, it has a positive and significant impact on the likelihood of producing an invention which is novel but not useful (see Model 9), which suggests that this type of invention might be more similar to patented inventions than those in the other categories. The effect sizes are quite large, especially if compared with what obtained by Singh and Fleming. According to the coefficient estimates of Model 4, inventors working in a team are 22.8% less likely to generate inventions which are patented than lone

estimation approach. However, we decided to report the results obtained by including the inverse Mills' ratio in the second stage logit model to allow comparison with Fleming and Singh's paper.

inventors. The mediation effects of network size appear to have the expected sign and significance, but the one for experience diversity goes in the opposite direction than what found by Singh and Fleming for the non-patenting and prior art outcomes.

--- INSERT TABLE 5 HERE ---

Replicating Audia and Goncalo's study

While Singh and Fleming clearly stated that the regressions were estimated using the patent as unit of analysis, in the Audia and Goncalo paper, it was unclear whether this was the case. As the authors state that "when a patent has multiple inventors, we attribute it to each inventor listed as co-author" (p. 7), we have assumed that the unit of analysis is a patent-inventor dyad. Also, the authors do not explicitly state whether they use patent applications or granted patents. As the source of the data is the USPTO, we have assumed that the authors have used granted patents.

One of key variables in this study is the past success of an individual inventor in his/her creative endeavor. Audia and Goncalo assume that inventors will compare their past performance with that of other inventors specialized in the same technology area. As a result, they measure this variable by calculating the number of patents developed by each inventor in the preceding two years minus the average number of patents generated by other inventors in the same technology area during the same period. We followed a similar procedure to compute this variable using our sample of inventors. First, we extracted all patents granted to each inventor since Venus's inception and identified the main area of specialization as the IPC technology class with the highest share of patents. Second, we derived the relative success measure for each inventor by applying Audia and Goncalo's formula. As we have information on all inventions developed by an individual, we calculated the success measure as well as all the other measures both using invention data as well as using the sample of

those patents granted at the USPTO. We also controlled for when the inventor first received a granted patent from the USPTO by deriving three cohort dummy variables: one for inventors who obtained their first patent from this patent office during the early years of operation of Venus (*Cohort with first patent first period*), one for inventors who received their first patent in the more recent years, and one for the inventors who were granted their first patent in between these two periods (*Cohort with first patent second period*). When we estimate our models using the sample of inventions we computed these cohort dummy variables using the date of an inventor first invention.

Before presenting our results, we wanted to evaluate whether it is plausible that the selection problem - although present - is random. We derived the proportion of patented inventions over the total number of inventions generated by all the inventors in our sample with more than two inventions. We did not take into account at which patent office the invention was protected and we considered patent applications instead of only granted patents. As shown in Figure 1, there is a lot of variation across inventors in their proportion of patented inventions, which suggests that there might a systematic bias introduced when one only considered the subset of patented inventions.

--- INSERT FIGURE 1 HERE ---

Table 6 presents the summary statistics of the variables used to replicate Audia and Goncalo's paper. It is worth noting that the maximum number of subclasses and new subclasses in our sample is smaller than what was found in the sample analyzed by Audia and Goncalo. This can be the result of our use primary and secondary IPC classification instead of the USPTO subclasses. The success measure derived using our sample of USPTO granted patents displays a wider range (min=-2, max=17.5) than the one used by Audia and Goncalo (min=-1.3 and max=0.995). This may be due to the fact that our sample of inventors seems to be more prolific than the ones in their sample: the yearly average number of patents in our

sample varies between 1 and 1.78; and among our inventors there are some very productive ones: 2% of inventors are granted more than five patents a year.

--- INSERT TABLE 6 HERE ---

In Table 7, we report the coefficients of the Cox proportional hazard model estimations. Due to the high number of inventors in our sample, we could not estimate the inventor-specific fixed effect model. Models 1 and 2 are estimated using the sample of patents for which we have information on whether the invention was generated within the scope of a R&D project or not. Model 1 does not produce consistent estimates for the control variables to the ones reported in Table 2 of Audia and Goncalo's paper. In particular *the inventor proportion of solo-inventor's patents* is not significant and the *cumulative inventor patents* variable is negative and significant. Model 2, however, confirms what is predicted and found by Audia and Goncalo, namely that the prior success of an inventor has a positive impact on the likelihood of patenting; the magnitude of this effect is relatively consistent with that found in their study. Using the minimum and maximum value of their success variable (min=-1.31, max= 0.99), we computed that the least successful inventors are 19% less likely (instead of 31%) to patent than the most successful inventors.

Models 3 and 4 include the inverse Mills' ratio which was generated from a first-stage probit model where we predicted the likelihood of an invention being granted a patent from the USPTO. As explanatory variables in this model, we included whether the invention was linked to an R&D project or not, whether it was generated by a team of inventors, and the cumulative number of inventions that the inventors have developed so far. Finally, we control for technology fixed-effects by including 15 one-digit technology class dummies. The inverse Mills' ratio is negative and significant in both Models suggesting that the estimates in Models 1 and 2 are affected by a selection bias. However, the inclusion of the inverse Mills' ratio does not seem to affect substantially the magnitude of the coefficient estimate for the

inventor's success variable. We do notice a substantial change in the effect of this variable when we consider all inventions that an inventor has generated and not only those protected by a USPTO granted patent. These results are reported in Models 5 and 6 of Table 7. Using the estimates of Model 6, we calculated that the difference in the likelihood of generating an invention between the least and most successful inventors in Audia and Goncalo's study is 8%, i.e. more than three times smaller than what was found in their study and more than two times smaller than what was found in Model 2.[6] Thus, by measuring the success of an inventor using only the number of granted patents, the researcher may be overestimating the impact of past success on the likelihood of generating new ideas.

--- INSERT TABLE 7 HERE ---

In Table 8, we report the estimates of the random-effects Poisson models predicting the number of new subclasses (3-digit IPC technology classes) in which a patent is classified which are new to the inventor. Model 1 shows results consistent to the ones obtained by Audia and Goncalo: the proportion of sole-inventor patents has a negative and significant impact on the probability that an inventor generates a patent in new technology areas, while the number of IPC classes in which a patent is classified increases this probability. Model 2 includes the variable capturing the innovative performance of the focal inventor which is negative and significant, confirming what was found by Audia and Goncalo and providing support for their Hypothesis 2, which stated that inventors who have been successful in the past are less likely to generate divergent ideas. To compare the magnitude of our coefficient estimate with that found by Audia and Goncalo, we calculate the difference in the probability of developing patents in new subclasses between the most (*inventor's success*=0.99) and least (*inventor's success*=-1.131) successful inventors in their sample. According to the coefficient

[6] We could not test whether these two coefficients are statistically different because the seemingly unrelated estimation procedure cannot be implemented with a Cox hazard model.

estimate in Model 2, this difference is equal to -64%, which corresponds to more than double the size of the effect found by Audia and Goncalo. Model 3 includes the interaction term between the *inventor's success* and the *proportion of sole-inventor patents*. This interaction effect is positive and significant, instead of being negative and significant. Thus, we do not find support for their Hypothesis 3.

Of most interest are the results reported in Models 4-6 where we introduce the inverse Mills' ratio to control for the selection bias which could be present in the previous models due to the fact that we only observe inventions which have been protected with a granted patent. The inverse Mills' ratio is always positive and significant in these models suggesting that the estimates of the previous models might be biased. To assess the extent of this bias we calculate the difference between the most and least successful inventors using the coefficient estimate obtained in Model 5 and found that this difference is now -65.8%. Thus, although there is a selection bias, this does not seem to affect the impact of the main independent variable. The last three models in Table 8 reports the results obtained by using the entire sample of inventions. The sign and significance of the coefficients are consistent to what found using the subset of patented inventions, however the magnitude of the coefficient for the success variable produces a difference in the probability of generating divergent ideas between the most and least success inventors significantly smaller than what found in Model 2 (-12% vs. -64%).[7] Thus, also in this case, by taking into consideration all innovative efforts of an individual and not only those leading to a granted patent, we found that a successful track record does affect the ability of an inventor to produce ideas which deviate from the past, but this impact is relatively small.

--- INSERT TABLE 8 HERE ---

[7] We could not test whether these two coefficients are statistically different because the seemingly unrelated estimation procedure cannot be implemented with the random effects poisson model.

DISCUSSION

In this study, we have tried to shed new light on the use and usefulness of patent data. In particular, by looking at what happens before a patent application is filed our goal was to understand whether the fact that some actions regularly occurring as part of the inventive process cannot be observed in patent data would introduce a significant bias in works building on patent data. To do so, we focused on the pivotal issue of selection bias and built on a company-internal dataset to replicate selected studies.

While it is clear that working with the internal data of just one company limits the generalizability of our findings, they are an important first step in *showing* that the above problems *do indeed exist*. And to do so, valuable intra-company data is needed, which we present, to the best of our knowledge, at a quality and depth rarely seen before in the literature. In addition, the methods that we employed made clear that comparability indeed exists. For our replication studies focusing on selection, we always provided a "baseline" effect for our sample, and then focused on the changes that occurred when introducing the additional information we had available—and importantly, the baseline models always looked similar to the results of the original studies. In short, while our focus lies merely on providing some initial evidence on the potential magnitude of the problem we highlight, we would go as far as claim the values we uncover should not differ much from the general population of firms, and definitely not from those similar to Venus.

In turn, these results indeed cast a shadow of doubt on some elements of the work building on patent data. First, focusing on potential selection issues, our study demonstrates that patent data is subject to significant left-side truncation, as many of the low-quality inventions are simply not observed. Although this has been commented upon previously, ours is the first study that we know of to actually demonstrate the scale of this selection bias. Our results suggest that patent-based studies of performance are liable to underestimate the

amount of inventive failure and therefore effort made in large organizations and by inventors. As a result, the studies of inventive success may overplay the salience of many variables in shaping innovative success, as many of our expectations about the value of patents and the underpinning drivers of creative success are subject to a significant degree of systematic measurement error. This would suggest that authors give greater attention to the dangers of directly attributing some behaviors and experiences to positive creative outcomes, holding back from drawing strong inferences about causality with respect to the nature of invention from incomplete patent data. At the same time, though, we find some corroboration for some of the critical variables in recent work on patents, such as the effect of team affiliation of high quality inventions and the importance of prior experience on future inventive success.

CONCLUSION

There is an increasing understanding that many aspects of managerial and corporate behavior are subject to hard to observe selection processes (e.g., Berk, 1983; Kovács & Denrell, 2008). These selection processes may hide from researchers some of the key drivers of performance outcomes or lead researcher to suggest that some behaviors drive performance when they do not. This is because in management research, we usually only see the part of a population – the firms, individuals and inventions - that appears on public records and repositories. Failure, partial attempts, and incomplete efforts are often unrecorded, hidden away behind organizational walls or obscured by poor information.

Patent data has always been known to be subject to sample selection. Yet, to date, few studies have attempted to account for this fact. Our study provides an attempt to grapple with this challenge. Our results raise a clear note of caution on the interpretation of this data. We urge our fellow researchers not to be lured by patent data's availability and expected potency, but they instead should make substantial efforts at either focusing on studying research

questions that can be covered with patent data, or on complementing it with other sources of data to overcome the issues we have identified. To move the field forward in this spirit, we see several opportunities.

First, patent-based work may benefit from purposefully limiting itself to those bits of the innovation process it can actually describe—those parts after the firm has made a decision to file an invention—and drop inferences about what might have happened before. For example, comparing the application-over-patent-granted rate as a proxy for organizational skill or innovative capability *across* firms may be misleading if one acknowledges that organizations' selection decisions with regards to patenting inventions vary: of course, if a firm only files its best inventions, it will get more patents granted when compared to a firm that files pretty much everything. However, *within-firm* comparisons may still be appropriate. For example, a firm's increase in patenting success following specific managerial decisions (given these do not affect the propensity to patent an invention) may allow to derive insights about their efficacy. At the same, one could also study factors that would change the propensity to patent (given these do not change the average quality of the firm's inventions).

Second, the most promising approach we see is the complementing of patent data with primary data collected from inventors or companies. While it may indeed be difficult and costly to attain such data, it is not impossible—for example, beyond our study, many examples exist where researchers attained access to a limited set of inventions, but were able to attain their complete history, including for example inventor surveys (Giuri et al., 2007) or the study of university inventors (Kotha et al., 2013).

Patent data is and will remain a key tool for research on the innovation process, as it provides a powerful lens on the nature of the people and technologies that underpin change in the economic system. At the same time, research using this data needs to be tempered with a strong degree of modesty about what is and what is not observed. This acknowledgement

could also be a spur to capture new and related data and information on the innovation process.

TABLES AND FIGURES

Figure 1 Distribution of the proportion of patented inventions by inventors in our sample

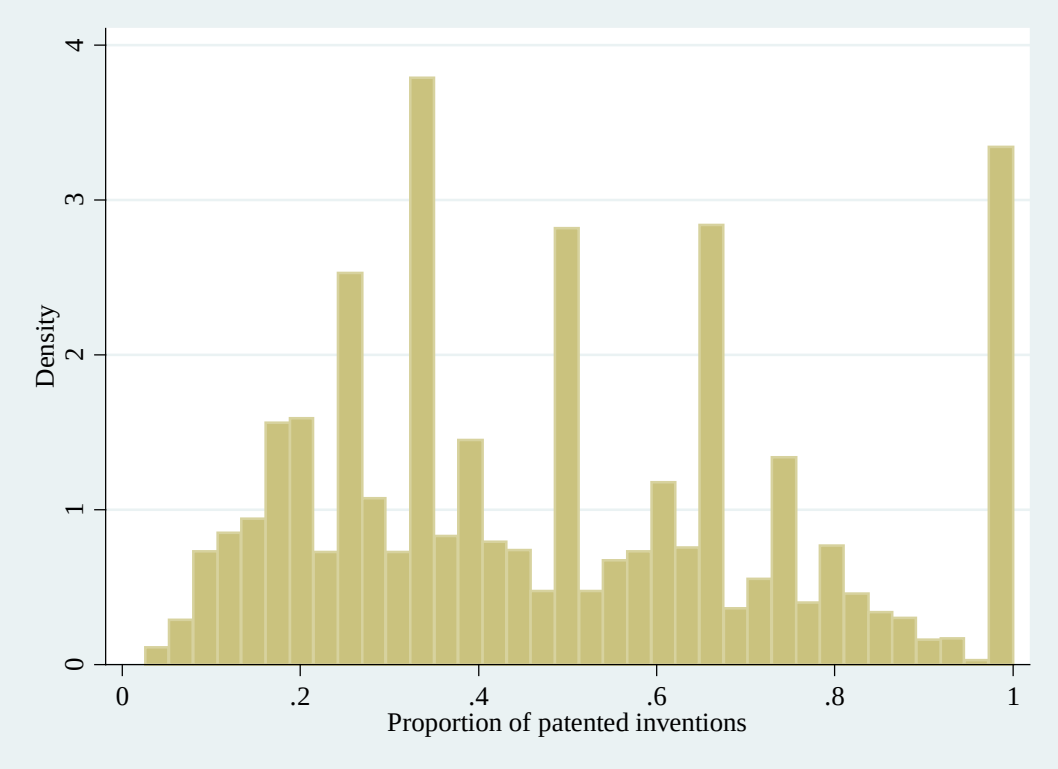

Table 1 Summary statistics (N=1,910)

	Mean	S.D.	Min	Max
Cites_p95	0.182	0.386	0	1
CitesEQ0	0.088	0.284	0	1
Team	0.632	0.482	0	1
Claims	26.094	17.516	1	330
Patent_references	14.686	14.715	0	134
Nonpatent_references	5.616	10.266	0	128
Average_experience	7.413	9.126	0	85
Joint_experience	1.025	1.457	0	18
Experience_diversity	2.155	1.666	0	11
Network_size	34.531	63.16	0	666

Table 2 Correlation matrix among variables

		1	2	3	4	5	6	7	8	9
1	*Cites_p95*									
2	*CitesEQ0*	-0.147								
3	*Team*	0.109	-0.038							
4	*ln_claims*	0.067	-0.058	0.109						
5	*ln_patent_references*	0.004	-0.053	0.050	0.090					
6	*ln_nonpatent_references*	0.018	-0.032	0.103	0.161	0.213				
7	*ln_average_experience*	0.032	0.034	0.158	0.137	0.044	0.162			
8	*ln_joint_experience*	0.079	-0.009	0.833	0.096	0.080	0.130	0.286		
9	*ln_experience_diversity*	-0.001	0.042	0.328	0.175	0.106	0.122	0.785	0.343	
10	*ln_network_size*	0.070	0.000	0.341	0.131	0.063	0.148	0.634	0.317	0.604

Correlations>|0.045| significant at 5%

Table 3 Regression Analyses of Extreme Outcomes upon Lone Invention

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Cites_p95	*Cites_p95*	*Cites_p95*	*Cites_p95*	*Cites_p95*	*CitesEQ0*	*CitesEQ0*	*CitesEQ0*	*CitesEQ0*	*CitesEQ0*
Regression Model	Logistic	Logistic	Logistic	Logistic	Logistic	Logistic	Logistic	Logistic	Logistic	Logistic
Team	0.911***	1.107***	0.795***	0.982***	0.939***	-0.832***	-0.943***	-0.743**	-0.856**	-0.788**
	(0.235)	(0.247)	(0.243)	(0.254)	(0.257)	(0.308)	(0.321)	(0.328)	(0.337)	(0.343)
ln_experience_diversity		-0.530**		-0.573***	-0.429*		0.284		0.319	0.0586
		(0.211)		(0.211)	(0.256)		(0.281)		(0.280)	(0.341)
ln_network_size			0.0684	0.0850*	0.170*			-0.0701	-0.0777	-0.266*
			(0.0469)	(0.0473)	(0.0889)			(0.0737)	(0.0736)	(0.152)
ln_expdiversity_ln_netsizeo					-0.0731					0.145
					(0.0676)					(0.105)
ln_claims	0.484***	0.495***	0.485***	0.497***	0.495***	-0.637***	-0.644***	-0.636***	-0.644***	-0.637***
	(0.121)	(0.119)	(0.120)	(0.118)	(0.119)	(0.152)	(0.152)	(0.152)	(0.152)	(0.152)
ln_patent_references	0.169**	0.374***	0.0942	0.298**	0.273**	0.00261	-0.114	0.0763	-0.0466	0.0121
	(0.0718)	(0.106)	(0.0878)	(0.116)	(0.119)	(0.0948)	(0.156)	(0.127)	(0.171)	(0.177)
ln_nonpatent_references	-0.363	-0.447**	-0.302	-0.379*	-0.350	0.581*	0.634**	0.547*	0.602**	0.556*
	(0.221)	(0.223)	(0.220)	(0.222)	(0.223)	(0.302)	(0.306)	(0.302)	(0.305)	(0.307)
ln_average_experience	-0.0799	-0.0456	-0.0879	-0.0514	-0.0465	-0.192	-0.208*	-0.187	-0.203*	-0.218*
	(0.112)	(0.114)	(0.112)	(0.114)	(0.114)	(0.120)	(0.121)	(0.121)	(0.121)	(0.123)
ln_joint_experience	0.0897	0.0787	0.0890	0.0769	0.0808	-0.150*	-0.139*	-0.146*	-0.134	-0.152*
	(0.0631)	(0.0622)	(0.0635)	(0.0628)	(0.0633)	(0.0848)	(0.0845)	(0.0843)	(0.0840)	(0.0829)
Year fixed effects	Included	Included	Included	Included	Included	Included	Included	Included	Included	Included
Technology fixed effects	Included	Included	Included	Included	Included	Included	Included	Included	Included	Included
Chi square test	107.674	113.597	109.288	116.670	120.229	105.145	106.629	105.163	107.095	107.010
Log-Likelihood	-843.541	-840.098	-842.507	-838.545	-838.013	-513.661	-513.126	-513.074	-512.410	-511.202
Observations	1,910	1,910	1,910	1,910	1,910	1,910	1,910	1,910	1,910	1,910

Robust standard errors for two-tailed tests clustered by the first inventor. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 4 Regression Analyses of Extreme Outcomes upon Lone Invention accounting for selection bias

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Cites_p95	*Cites_p95*	*Cites_p95*	*Cites_p95*	*Cites_p95*	*CitesEQ0*	*CitesEQ0*	*CitesEQ0*	*CitesEQ0*	*CitesEQ0*
Regression Model	Logistic	Logistic	Logistic	Logistic	Logistic	Logistic	Logistic	Logistic	Logistic	Logistic
Team	0.995***	1.205***	0.880***	1.080***	1.037***	-0.516	-0.641**	-0.465	-0.587*	-0.496
	(0.245)	(0.254)	(0.252)	(0.261)	(0.264)	(0.314)	(0.323)	(0.330)	(0.336)	(0.344)
ln_experience_diversity		-0.543***		-0.585***	-0.448*		0.300		0.320	-0.0261
		(0.208)		(0.208)	(0.253)		(0.269)		(0.269)	(0.333)
ln_network_size			0.0680	0.0847*	0.166*			-0.0388	-0.0466	-0.296**
			(0.0469)	(0.0473)	(0.0888)			(0.0712)	(0.0711)	(0.150)
ln_expdiversity X lnetsize					-0.0696					0.190*
					(0.0675)					(0.101)
ln_claims	0.485***	0.495***	0.486***	0.497***	0.495***	-0.528***	-0.537***	-0.527***	-0.536***	-0.529***
	(0.121)	(0.119)	(0.120)	(0.119)	(0.119)	(0.155)	(0.156)	(0.155)	(0.156)	(0.155)
ln_patent_references	0.142*	0.350***	0.0684	0.274**	0.251**	0.0413	-0.0779	0.0824	-0.0368	0.0398
	(0.0736)	(0.107)	(0.0896)	(0.118)	(0.120)	(0.0981)	(0.150)	(0.124)	(0.162)	(0.171)
ln_nonpatent_references	-0.361	-0.446**	-0.301	-0.379*	-0.351	0.449	0.503*	0.429	0.482	0.421
	(0.219)	(0.221)	(0.219)	(0.221)	(0.221)	(0.294)	(0.296)	(0.294)	(0.296)	(0.300)
ln_average_experience	-0.0917	-0.0571	-0.0997	-0.0629	-0.0581	-0.284**	-0.300**	-0.282**	-0.298**	-0.312**
	(0.113)	(0.115)	(0.113)	(0.115)	(0.115)	(0.123)	(0.123)	(0.123)	(0.124)	(0.125)
ln_joint_experience	0.108*	0.0982	0.107*	0.0962	0.0994	-0.0526	-0.0443	-0.0502	-0.0407	-0.0628
	(0.0624)	(0.0617)	(0.0628)	(0.0623)	(0.0627)	(0.0827)	(0.0823)	(0.0823)	(0.0819)	(0.0804)
Inverse Mills' Ratio	0.539	0.592	0.537	0.591	0.579	0.844*	0.791*	0.841*	0.783*	0.803*
	(0.369)	(0.368)	(0.371)	(0.371)	(0.371)	(0.461)	(0.461)	(0.461)	(0.462)	(0.466)
Year fixed effects	Included	Included	Included	Included	Included	Included	Included	Included	Included	Included
Technology fixed effects	Included	Included	Included	Included	Included	Included	Included	Included	Included	Included
Chi square test	109.574	117.082	111.116	120.197	123.548	86.458	87.667	86.826	88.192	92.534
Log-Likelihood	-842.542	-838.906	-841.521	-837.368	-836.884	-527.881	-527.277	-527.698	-527.015	-524.844
Observations	1,910	1,910	1,910	1,910	1,910	1,910	1,910	1,910	1,910	1,910

Robust standard errors for two-tailed tests clustered by the first inventor. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 5 Regressions Analyses of Poor Innovative Outcomes (N=35,144)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Dependent Variable	*Notpat*	*Notpat*	*Notpat*	*Notpat*	*Priorart*	*Priorart*	*Priorart*	*Priorart*	*Notuseful*	*Notuseful*	*Notuseful*	*Notuseful*
Regression Model	Logistic	Logistic	Logistic	Logistic	Logistic	Logistic	Logistic	Logistic	Logistic	Logistic	Logistic	Logistic
Team	-1.063***	-1.211***	-0.855***	-1.021***	-0.775***	-1.044***	-0.743***	-0.967***	0.168***	0.171**	0.0197	0.0550
	(0.0402)	(0.0488)	(0.0437)	(0.0492)	(0.0452)	(0.0501)	(0.0469)	(0.0502)	(0.0572)	(0.0687)	(0.0624)	(0.0699)
ln_experience_diversity		0.236***		0.351***		0.418***		0.420***		-0.00607		-0.204***
		(0.0519)		(0.0662)		(0.0519)		(0.0680)		(0.0555)		(0.0764)
ln_network_size			-0.119***	-0.165***			-0.0176	-0.0791***			0.0827***	0.0398
			(0.0174)	(0.0205)			(0.0192)	(0.0236)			(0.0210)	(0.0255)
ln_expdiversity X ln_netsize				0.00535				0.0143				0.0340***
				(0.00975)				(0.0112)				(0.0110)
ln_average_experience	-0.186***	-0.342***	-0.0438	-0.241***	-0.0494**	-0.333***	-0.0285	-0.295***	0.101***	0.105**	0.00205	0.0508
	(0.0208)	(0.0455)	(0.0335)	(0.0476)	(0.0235)	(0.0483)	(0.0391)	(0.0519)	(0.0228)	(0.0441)	(0.0387)	(0.0479)
ln_joint_experience	0.333***	0.374***	0.276***	0.318***	0.188***	0.260***	0.179***	0.230***	-0.0523	-0.0533	-0.00964	-0.0344
	(0.0282)	(0.0298)	(0.0302)	(0.0316)	(0.0379)	(0.0360)	(0.0383)	(0.0338)	(0.0356)	(0.0372)	(0.0374)	(0.0398)
Year fixed effects	Included	Included	Included	Included	Included	Included	Included	Included	Included	Included	Included	Included
Technology fixed effects	Included	Included	Included	Included	Included	Included	Included	Included	Included	Included	Included	Included
Chi square test	109.574	117.082	111.116	120.197	123.548	86.458	87.667	86.826	88.192	92.534	109.574	117.082
Log-Likelihood	-842.542	-838.906	-841.521	-837.368	-836.884	-527.881	-527.277	-527.698	-527.015	-524.844	-842.542	-838.906

Robust standard errors for two-tailed tests clustered by the first inventor. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 6 Descriptive statistics and correlations (N=4,183)

	Mean	S.D.	Min	Max	1	2	3	4	5	6
1 *Inventor success using patents*	1.512	3.047	-2	17.583						
2 *Proportion sole-inventor patents*	0.120	0.259	0	1	0.187					
3 *New subclasses*	0.522	0.743	0	5	-0.355	-0.198				
4 *Cumulative inventor patents*	6.241	7.711	1	-	0.634	0.144	-0.368			
5 *Cohort with first patent first period*	0.108	0.310	0	1	0.175	0.080	-0.155	0.387		
6 *Cohort with first patent second period*	0.642	0.480	0	1	0.109	0.044	-0.137	0.013	-0.465	
7 *Number of subclasses*	4.065	2.501	1	17	0.032	0.007	0.151	0.017	-0.005	0.041

Correlations>|0.018| significant at 5%.

^ The maximum of these two variables is not reported to avoid identification of Venus

Table 7 Cox Models of the probability of Patenting and Inventing

	(1)	(2)	(3)	(4)	(5)	(6)
Cohort with first patent first period	2.084***	2.230***	1.962***	2.114***		
	(0.156)	(0.154)	(0.163)	(0.155)		
Cohort with first patent second period	1.280***	1.295***	1.311***	1.325***		
	(0.0653)	(0.0755)	(0.0641)	(0.0735)		
Proportion sole-inventor patents	0.0860	0.0214	0.242***	0.184**		
	(0.0791)	(0.0805)	(0.0744)	(0.0763)		
Cumulative inventor patents	-0.0571***	-0.0844***	-0.0198**	-0.0464***		
	(0.00907)	(0.0149)	(0.00818)	(0.0136)		
Inventor success using patents		0.0921***		0.0922***		
		(0.0169)		(0.0148)		
Inverse Mills' ratio			-1.825***	-1.847***		
			(0.114)	(0.117)		
Cohort with first invention first period					1.101***	1.411***
					(0.117)	(0.144)
Cohort with first invention second period					0.996***	1.143***
					(0.0550)	(0.0631)
Proportion sole-inventor inventions					0.00546	0.0104
					(0.0399)	(0.0431)
Cumulative inventor inventions					-0.0128***	-0.0309***
					(0.00389)	(0.00589)
Inventor success using inventions						0.0370***
						(0.00517)
Log-Likelihood	-30020.122	-29957.237	-29705.677	-29639.333	-904408.495	-901518.698
Observations	4,183	4,183	4,183	4,183	69,621^	69,621^

^ The number of observations is been reduced by a randomly drawn percentage figure.
Robust standard errors for two-tailed tests clustered by the first inventor. * significant at 10%; ** significant at 5%; *** significant at 1%

Table 8 Poisson Regressions of Patent Characteristics Indicative of Diverging Creative Efforts

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Cohort with first patent first period	-1.369***	-0.954***	-0.952***	-1.358***	-0.951***	-0.949***			
	(0.122)	(0.114)	(0.114)	(0.121)	(0.114)	(0.114)			
Cohort with first patent second period	-0.590***	-0.371***	-0.362***	-0.598***	-0.358***	-0.349***			
	(0.0492)	(0.0449)	(0.0449)	(0.0490)	(0.0452)	(0.0452)			
Proportion sole-inventor patents	-1.483***	-0.794***	-1.102***	-1.436***	-0.800***	-1.111***			
	(0.141)	(0.123)	(0.149)	(0.140)	(0.123)	(0.149)			
Number of subclasses	0.0842***	0.0821***	0.0829***	0.0820***	0.0838***	0.0845***	0.524***	0.508***	0.508***
	(0.00799)	(0.00756)	(0.00756)	(0.00800)	(0.00759)	(0.00759)	(0.00382)	(0.00362)	(0.00361)
Inventor success using patents		-0.290***	-0.328***		-0.299***	-0.337***			
		(0.0153)	(0.0174)		(0.0158)	(0.0179)			
Success X Proportion of sole-inventor patents			0.357***			0.359***			
			(0.0667)			(0.0672)			
Inverse Mills' ratio				-0.344***	0.210**	0.210**			
				(0.0915)	(0.0874)	(0.0871)			
Cohort with first invention first period							-0.375***	-0.342***	-0.338***
							(0.0500)	(0.0405)	(0.0402)
Cohort with first invention second period							-0.193***	-0.130***	-0.129***
							(0.0114)	(0.00942)	(0.00937)
Proportion sole-inventor inventions							-0.340***	-0.216***	-0.243***
							(0.0150)	(0.0136)	(0.0140)
Inventor success using inventions								-0.0559***	-0.0609***
								(0.000752)	(0.000926)
Success X Proportion of sole-inventor inventions									0.0235***
									(0.00238)
Constant	-0.438***	-0.492***	-0.513***	0.0468	-0.793***	-0.814***	-0.838***	-0.820***	-0.820***
	(0.0510)	(0.0481)	(0.0485)	(0.138)	(0.135)	(0.135)	(0.0104)	(0.00942)	(0.00939)
Log-Likelihood	-3677.924	-3424.514	-3412.758	-3670.706	-3421.686	-3409.891	-88324.665	-85125.374	-85079.680
Observations	4,183	4,183	4,183	4,183	4,183	4,183	69,621^	69,621^	69,621^

^ The number of observations is been reduced by a randomly drawn percentage figure.

REFERENCES

Ahuja, G. 2000. Collaboration networks, structural holes, and innovation: A longitudinal study. ***Administrative Science Quarterly***, 45(3): 425-455.

Arundel, A. & Kabla, I. 1998. What percentage of innovations are patented? Empirical estimates for European firms. ***Research Policy***, 27: 127–141.

Audia, P. G. & Goncalo, J. A. 2007. Past success and creativity over time: A study of inventors in the hard disk drive industry. ***Management Science***, 53(1): 1-15.

Berk, R. A. 1983. An introduction to sample selection bias in sociological data. ***American Sociological Review***: 386-398.

Cohen, W. M., Nelson, R. R., & Walsh, J. 2000. Protecting their intellectual assets: Appropriability conditions and why U.S. Manufacturing firms patent (or not), ***NBER Working Paper Series***. Massachusetts: National Bureau of Economic Research.

Conti, R., Gambardella, A., & Mariani, M. 2014. Learning to be Edison: Inventors, organizations, and breakthrough inventions. ***Organization Science***, forthcoming.

Criscuolo, P., Salter, A., & Ter Wal, A. L. J. 2014. Going underground: Bootlegging and individual innovative performance. ***Organization Science***, forthcoming.

de Rassenfosse, G. & van Pottelsberghe de la Potterie, B. 2009. A policy insight into the R&D-patent relationship. ***Research Policy***, 38(5): 779-792.

Fontana, R., Nuvolari, A., Shimizu, H., & Vezzulli, A. 2013. Reassessing patent propensity: Evidence from a dataset of R&D awards, 1977–2004. ***Research Policy***, 42(10): 1780-1792.

Girotra, K., Terwiesch, C., & Ulrich, K. T. 2010. Idea generation and the quality of the best idea. ***Management Science***, 56(4): 591-605.

Gittelman, M. 2008. A note on the value of patents as indicators of innovation: Implications for management research. ***The Academy of Management Perspectives***, 22(3): 21-27.

Giuri, P., Mariani, M., Brusoni, S., Crespi, G., Francoz, D., Gambardella, A., Garcia-Fontes, W., Geuna, A., Gonzales, R., Harhoff, D., Hoisl, K., Le Bas, C., Luzzi, A., Magazzini, L., Nesta, L., Nomaler, Ö., Palomeras, N., Patel, P., Romanelli, M., & Verspagen, B. 2007. Inventors and invention processes in Europe: Results from the patval-eu survey. ***Research Policy***, 36(8): 1107-1127.

Griliches, Z. 1984. ***R&D, patents and productivity***: The University of Chicago Press.

Hall, B. H. & Ziedonis, R. H. 2001. The patent paradox revisited: An empirical study of patenting in the US semiconductor industry, 1979-1995. ***RAND Journal of Economics***: 101-128.

Hall, B. H., Jaffe, A. B., & Trajtenberg, M. 2002. The NBER patent citations data file: Lessons, insights and methodological tools. In A. B. Jaffe & M. Trajtenberg (Eds.), ***Patents, citations and innovations: A window on the knowledge economy***: 403-460. Cambridge, MA: MIT Press.

Heckman, J. J. 1976. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models, ***Annals of economic and social measurement, volume 5, number 4***: 475-492: NBER.

Jaffe, A. B. & Trajtenberg, M. 2002. Introduction to "patents, citations and innovations: A window on the knowledge economy". In A. B. Jaffe & M. Trajtenberg (Eds.), ***Patents, citations and innovations: A window on the knowledge economy***: 1-24. Cambridge, MA: MIT Press.

Kotha, R., George, G., & Srikanth, K. 2013. Bridging the mutual knowledge gap: Coordination and the commercialization of university science. ***Academy of Management Journal***, 56(2): 498-524.

Kovács, B. & Denrell, J. 2008. Selective sampling of empirical settings in organizational studies. ***Administrative Science Quarterly***, 53(1): 109-144.

Moser, P. 2005. How do patent laws influence innovation? Evidence from nineteenth-century world's fairs. ***American Economic Review***, 95(4): 1214-1236.

Moser, P. 2012. Innovation without patents: Evidence from world's fairs. ***Journal of Law and Economics***, 55(1): 43-74.

Singh, J. & Fleming, L. 2010. Lone inventors as sources of breakthroughs: Myth or reality? ***Management Science***, 56(1): 41-56.

Stolzenberg, R. M. & Relles, D. A. 1997. Tools for intuition about sample selection bias and its correction. ***American Sociological Review***: 494-507.

Taylor, A. & Greve, H. R. 2006. Superman or the fantastic four? Knowledge combination and experience in innovative teams. ***Academy of Management Journal***, 49(4): 723-740.

Whitehead, A. N. 1925. ***Science and the modern world***. New York: Macmillan.

Appendix

Table 1A Regression Analyses of Extreme Outcomes upon Lone Invention using USPTO patents (N=5,077)^

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Cites_p95	*Cites_p95*	*Cites_p95*	*Cites_p95*	*Cites_p95*	*CitesEQ0*	*CitesEQ0*	*CitesEQ0*	*CitesEQ0*	*CitesEQ0*
Regression Model	Logistic	Logistic	Logistic	Logistic	Logistic	Logistic	Logistic	Logistic	Logistic	Logistic
Team	0.394***	0.388***	0.268*	0.272*	0.274*	-0.629***	-0.682***	-0.545***	-0.601***	-0.559***
	(0.137)	(0.146)	(0.140)	(0.147)	(0.148)	(0.179)	(0.190)	(0.187)	(0.196)	(0.196)
ln_experience_diversity		0.0133		-0.0114	-0.0162		0.129		0.146	0.0131
		(0.118)		(0.120)	(0.134)		(0.164)		(0.164)	(0.177)
ln_network_size			0.0944***	0.0946***	0.0906			-0.0715*	-0.0738*	-0.202**
			(0.0300)	(0.0304)	(0.0567)			(0.0427)	(0.0425)	(0.0896)
ln_expdiversity_ln_netsize					0.00346					0.101
					(0.0417)					(0.0638)
ln_claims	0.411***	0.411***	0.405***	0.405***	0.405***	-0.468***	-0.468***	-0.465***	-0.464***	-0.460***
	(0.0658)	(0.0658)	(0.0657)	(0.0657)	(0.0657)	(0.0851)	(0.0852)	(0.0853)	(0.0854)	(0.0853)
ln_patent_references	0.0379	0.0319	-0.0649	-0.0600	-0.0590	0.000760	-0.0585	0.0752	0.0104	0.0419
	(0.0434)	(0.0649)	(0.0579)	(0.0712)	(0.0728)	(0.0580)	(0.0994)	(0.0725)	(0.108)	(0.109)
ln_nonpatent_references	-0.131	-0.128	-0.0642	-0.0664	-0.0679	0.447**	0.473***	0.410**	0.439**	0.404**
	(0.141)	(0.143)	(0.143)	(0.145)	(0.145)	(0.177)	(0.181)	(0.178)	(0.182)	(0.183)
ln_average_experience	0.0288	0.0285	0.0238	0.0240	0.0239	-0.0837	-0.0860	-0.0819	-0.0844	-0.0895
	(0.0623)	(0.0623)	(0.0625)	(0.0625)	(0.0627)	(0.0693)	(0.0691)	(0.0695)	(0.0693)	(0.0694)
ln_joint_experience	0.0819**	0.0822**	0.0802**	0.0799**	0.0799**	-0.0671	-0.0615	-0.0627	-0.0562	-0.0609
	(0.0344)	(0.0344)	(0.0345)	(0.0345)	(0.0346)	(0.0468)	(0.0475)	(0.0467)	(0.0474)	(0.0476)
Year fixed effects	Included	Included	Included	Included	Included	Included	Included	Included	Included	Included
Technology fixed effects	Included	Included	Included	Included	Included	Included	Included	Included	Included	Included
Chi square test	207.120	207.416	211.992	212.489	213.213	302.898	307.966	304.643	310.757	309.390
Log-Likelihood	-2972.309	-2972.301	-2966.557	-2966.550	-2966.547	-1717.433	-1717.045	-1715.789	-1715.295	-1713.626

Robust standard errors for two-tailed tests clustered by the first inventor. * significant at 10%; ** significant at 5%; *** significant at 1%.

^ The number of observations is been reduced by a randomly drawn percentage figure.

Table 2A Regressions of Experience Diversity and Network Size as Potential Moderators using the USPTO sample

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable:	Experience_diversity	Network_size	Experience_diversity	Network_size	Experience_diversity	Network_size
Regression model:	Negative binomial	Negative binomial	Negative binomial	Negative binomial	Negative binomial	Negative binomial
Team	0.607***	2.104***	0.673***	2.117***	0.665***	2.160***
	(0.0501)	(0.149)	(0.0529)	(0.155)	(0.0341)	(0.111)
ln_claims	0.0345	-0.0402	0.0340	-0.0392	-0.00424	0.142***
	(0.0241)	(0.0703)	(0.0241)	(0.0700)	(0.0147)	(0.0455)
ln_average_experience	0.556***	1.415***	0.541***	1.413***	0.636***	1.661***
	(0.0209)	(0.0549)	(0.0208)	(0.0547)	(0.0184)	(0.0436)
ln_joint_experience	-0.231***	-1.025***	-0.235***	-1.028***	-0.284***	-1.158***
	(0.0380)	(0.143)	(0.0370)	(0.142)	(0.0316)	(0.113)
ln_patent_references	0.125***	0.125**	0.117***	0.122**	0.0555***	0.0784*
	(0.0250)	(0.0531)	(0.0244)	(0.0537)	(0.0163)	(0.0415)
ln_nonpatent_references	-0.0496***	-0.0307	-0.0360**	-0.0283	-0.0724***	0.0144
	(0.0143)	(0.0354)	(0.0141)	(0.0358)	(0.0112)	(0.0254)
Inverse Mills' ratio			0.386***	0.0600		
			(0.0974)	(0.262)		
Year fixed effects	Included	Included	Included	Included	Included	Included
Technology fixed effects	Included	Included	Included	Included	Included	Included
Chi square test	1681.093	1350.932	1742.482	1352.244	3023.134	2765.857
Log-Likelihood	-2779.427	-6952.866	-2770.247	-6952.822	-8847.910	-19156.528
Observations	1,910	1,910	1,910	1,910	5,077^	5,077^

Robust standard errors for two-tailed tests clustered by the first inventor. * significant at 10%; ** significant at 5%; *** significant at 1%.
Estimates of Models 5 and 6 are obtained using the full sample of USPTO granted patents while estimates of Models 1-4 using the subset of patents for which we have information on the R&D project.
^ The number of observations is been reduced by a randomly drawn percentage figure.