Assortative Matching in Patent Production

Giulia Lo Forte[*]

October 1, 2021

PRELIMINARY AND INCOMPLETE – PLEASE DO NOT CIRCULATE
Link to Most Recent Version

Abstract

This paper studies the collaboration process of inventors using the universe of patents filed at the European Patent Office between 1978 and 2017. I develop a one-factor static Roy model predicting that inventors sort into leaders and supporters based on an endogenous cut-off rule, and then match together to produce patents in teams of different size. Consistent with the model predictions, I find empirically that more productive leaders and supporters sort into bigger teams. Most importantly, leaders' productivity exhibits a strong positive correlation with the productivity of supporters working with them, but this correlation decreases in magnitude as team size increases. The extent of the assortative matching is significant: in the most conservative case, a one standard deviation increase in leaders' productivity is associated with doubling the average supporter productivity. The paper concludes with empirical evidence on the life-cycle of inventors, who move to bigger teams as they age and become leaders after gaining experience and learning from working on previous patents.

[*] Vancouver School of Economics, University of British Columbia

1 Introduction

Patents are a measure of the amount of innovation that is generated over time. Two agents are key when filing a patent: applicants, which are the legal entities sponsoring the invention and benefitting from its exclusive use, and inventors, which are the natural people responsible for the invention. Both the scientific (Wuchty et al. (2007)) and the economics literature (Jones (2009)) testify that papers and patents have increasingly been produced by larger teams of scientists and inventors in the last 50 years. Given the increasing dominance of teams in the production of knowledge, understanding how they are formed and whether bigger ones produce substantially different patents in terms of quality is key to foster innovation and growth.

This paper studies the collaboration process of inventors. Specifically, I investigate whether there is positive assortative matching between most productive inventors, whether bigger teams of inventors develop better patents, and if so, what are the channels that mediate the increase in quality.

To do so, I analyse the universe of patents filed at the European Patent Office (EPO) between 1978 and 2017. I find that, within applicant, the average patent quality increases in the size of teams. However, bigger teams do not seem to be associated with more productive inventors on average, whether measured by age, cumulative amount of past citations, or cumulative number of past patents. This result masks a great amount of heterogeneity across teams. Distinguishing inventors between leaders and supporters within each team based on the quality of their previous patents unveils another pattern. More productive team leaders sort into bigger teams, confirming the predictions developed by Akcigit et al. (2018). The productivity of supporting inventors increases with team size, but it is less statistically significant and less robust to alternative measurements of productivity. Most importantly, the productivity of leaders exhibits a strong positive correlation with the productivity of supporting inventors that work with them, hinting towards the existence of assortative matching between the two groups. Specifically, even in the most conservative case, a one standard deviation increase in leaders' productivity is associated, on average, to roughly doubling the average productivity of supporters.

From a theoretical point of view, these stylized facts can be rationalized by a simple one-factor version of the Roy model (Roy (1951)) with endogenous formation of groups. I introduce a static model where individuals with heterogeneous productivity choose whether to become leaders or supporters. Leaders and supporters join together to create a patent, whose quality is increasing in the productivity of the leader and the number of supporters he manages. If they become leaders, individuals choose the team size in order to maximize their utility, which is proportional to the quality of the patent produced, net of a disutility coming from monitoring costs. If they become supporters, individuals enter a frictionless market where they can optimally choose which leader to join. In doing so, supporters maximize their utility, which is increasing in the productivity of the leader they join, net of a disutility cost coming from sharing the leader with other teammates. Given these assumptions, inventors choose whether to become leaders or supporters of a team based on their productivity and an endogenous cut-off rule: if their productivity is bigger than a certain threshold level, they become leaders. Despite its simplicity, the model provides some testable predictions on the behaviour of individuals. Specifically, it predicts that more productive supporters match with more productive leaders, but the correlation between them decreases in magnitude as team size increases.

In order to validate the predictions of the model, I compile information on teams, characteristics of their members, as well as characteristics of the output produced by the teams. The Organisation for Economic Co-operation and Development (OECD) collects data on all patents filed at the European Patent Office between 1978 and 2017 into a patent database having all the required features. In this paper, I combine three different data sources. First, the REGPAT database, with information on both inventors and applicants of each patent. Second, the HAN database, which allows to harmonize the names of applicants over time. Third, the Quality Index database, which contains a wide set of quality indicators for each patent filed at the European Patent Office. Moreover, I rely on the algorithm described by Raffo and Lhuillery (2009) to disambiguate inventors based on homonymy or quasi-homonymy of their names and address of residence.

The result is a rich dataset of 3,490,866 patents filed by 4,147,765 inventors and 508,087 applicants. For each patent, the data comprises information on applicants' and inventors' name, their address, the technology field of the patent as well as several measures of its quality. The baseline measure of quality used here is the number of times the patent has been cited by other patents in a 5-year window after it has been filed. This simple count is normalized by filing year and technology field in order to account for potential differences in citations due to time trends or intrinsic characteristics of the field of the patent.

This dataset allows me to track the same inventor over time and use common patents between inventors to define teams. I rely on the technological field and quality of past patents produced by an inventor to construct a measure of his specialization and productivity, respectively. Both measures are used to identify leader inventors within each team as inventors with the highest productivity among those specialized in the same field as the patent's. The panel feature of the data is exploited in the empirical section, together with a set of fixed effects accounting for unobservable characteristics of inventors and applicants that could otherwise bias the results.

I first show that, within applicant, technology field, and filing year, the quality of a patent is increasing in the size of the team of inventors and in their average productivity. However, the team average productivity does not seem to vary by size. Only distinguishing between leaders and supporters within each team shows that more productive leaders sort into bigger teams, as in Akcigit et al. (2018), as well as more productive supporters, even though this result is less statistically significant and is less robust to alternative definitions of productivity. The main novelty of this paper, however, consists in the empirical evidence of positive assortative matching between leaders and supporters: more productive leaders work with more productive supporters. This match seems to be more pronounced in smaller teams. A number of robustness checks confirms that this is not mechanically driven. First, I exclude the most productive supporter within each team and repeat the analysis. I find that the productivity of the leader is positively correlated with the average productivity of his remaining supporters. This check directly addresses the threat that the most productive supporter could mechanically increase the productivity of the leader based on its definition. Second, I use a placebo dataset to estimate the same specifications, given by reassigning randomly inventors into teams of the same size as in the original data, and then identifying again leaders and supporters within each team. I still find a positive and significant correlation between leaders' and supporters' productivities, but its magnitude is at most only 45% of my original findings. This suggests that the definition of leaders drives mechanically only half of the effect I find, leaving out a still sizeable positive relationship: in the most conservative case, on average, a one standard deviation in the leaders' productivity is associated to roughly doubling the average supporter productivity. Moreover, I show that supporters' productivity increases in leaders' productivity at a decreasing rate in team size. This is consistent with more productive leaders monitoring bigger teams and the existence of an endogenous cut-off productivity determining whether inventors are leaders or supporters. In fact, very productive inventors would sort themselves into being leaders of their own team rather than being supporters for someone else who is only marginally more productive. Finally, I explore whether inventors play different roles over time, finding that 22% of inventors are always supporters throughout their life, while 34% of inventors are always leaders. For inventors that are leaders at least once, experience seems to be important for leadership: on average, inventors play the role of supporters for 4 years, or in 2 patents, before becoming leaders for the first time. In addition, I show that all inventors, regardless of their role, move to bigger patents as they age. These result provide a microfoundation for a dynamic version of the model, which I leave to the next extension of this paper.

This paper relates to the literature on patents and production in teams. The relevance of teams in the research process has been documented in the scientific literature by Wuchty et al. (2007), and Jones et al. (2008), and in the economics literature by Jones (2009). Specifically, Jones (2009) rationalizes the surge in larger teams with a general equilibrium model where the total amount of knowledge in the world increases over time, thus forcing inventors to specialize in a field and collaborate to develop patents. However, the relationship between team size and the quality itself of the output remains unexplored.

While the economics literature does not deal with this aspect, the management literature provides some insights. Lee et al. (2015) use survey data on active US academics in various fields to show that team size has an inverted-U shape relationship with novelty and an increasing relationship with the impact of the paper. Diversity in terms of field of experience or task does not seem to have an effect on the impact of the paper, net of novelty. Singh and Fleming (2010) use data on patents filed at the US Patents and Trademarks Office to show that those developed by teams are more likely to be breakthrough patents while simultaneously being less likely to be of low value. Moreover, the effect of team affiliation seems to be mediated by experience diversity and network size. However, their study only compares solo inventors with inventors working in teams, regardless of their size.

Since the production of a patent probably entails a learning process in the team, this paper also relates to the literature on learning between individuals. While there is a large body of work in macroeconomics focused on learning from random meetings within the population or a sub-sample of the population (among others, see Lucas (2009) and Perla and Tonetti (2014)), few studies have analysed learning in teams. Burstein and Monge-Naranjo (2009) develop a setting where a manager affects the knowledge of identical workers in her team. Akcigit et al. (2018) develop a model where patents are developed in teams, but the learning process happens in pairwise interactions from the whole population. Moreover, while highlighting that more productive leaders are able to manage bigger teams of inventors, their theoretical framework does not encompass a matching process between leaders and supporters within a team. Finally, Herkenhoff et al. (2018) estimate a search and match model with dynamic types where learning happens in a heterogeneous team, with size of only two coworkers.

Jarosch et al. (2021) develop a more flexible model and structurally estimate it using a rich German employer-employee dataset. However, in both papers, knowledge is measured through wages of workers, which may be an imperfect measure of their productivity or the quality of their output. In addition, my research maintains the focal point on learning in teams, while focusing on inventors of the same patent and the matching process responsible for creating the teams themselves.

The theoretical part of this paper relates to models of sorting that have built on the framework pioneered by Roy (1951). Among them, Rosen (1982) is the most relevant for this paper, as it introduces endogenous formation of groups and hierarchies. Moreover, the model I propose can be linked to the recent study of Luttmer (2015), where a manager and a worker are paired according to the comparative advantage in learning, so that fast learners are matched with the most productive managers.[1]

Last but not least, this study relates to the literature on knowledge diffusion, networks and citations. Ductor et al. (2014) use data on co-authorship in economics to show that knowing the co-authorship network of an economist provides extremely modest information on her future productivity. Head et al. (2019) show that controlling for personal ties among mathematicians significantly reduces the probability of citing another paper in the field.

The rest of the paper is organized as follows. Section 2 introduces the theoretical framework and the predictions of the model. Section 3 describes the data and how the main variables have been constructed. Section 4 shows the empirical strategy as well as the results in the cross-section analysis. Section 5 explores the empirical evidence on inventors' life-cycle and their sorting into leaders and supporters over time. Section 6 concludes.

2 Theoretical framework

I propose a simple model where a unit mass of heterogeneous agents, characterized by productivity $Z_i \in [0, \infty)$, choose between two roles: they can be either leaders or supporting inventors. Leaders and supporters produce a patent in teams of one leader and multiple supporters. Leaders choose how many supporters to manage, while supporters choose which leader to join from a frictionless market with perfect information. An intuitive interpretation would be that leaders choose how many spots to open, while supporters choose which spots to fill, having perfect information on the number of spots available and the characteristics of the leader managing the spots. The maximization problem of leaders is similar to the model proposed by Akcigit et al. (2018), while the maximization problem of supporters provides new predictions on the match between leaders and supporters.

2.1 Set up

A team of one leader and $n \in [0, \infty)$ supporters produces a patent of quality $q = n^{\eta}\left(Z_i^l\right)^{1-\eta}$, where $\eta \in [0, 1]$ denotes the span of control of the leader in a fashion similar to Lucas (1978). A leader with productivity Z_i^l chooses the size of his team of supporters in order to maximize the quality of the patent produced, net of a disutility cost due to monitoring supporters. This disutility is assumed to be linear in the number of supporters

[1] For other research encompassing both learning and comparative advantage, please refer to Jovanovic (1979).

monitored, with marginal disutility given by $m > 0$. Leaders solve the following maximization problem:

$$\max_{n \geq 0} U_i^l(n; Z_i^l) = n^\eta \big(Z_i^l\big)^{1-\eta} - mn \tag{1}$$

Note that a leader can choose not to monitor supporters, thus choosing $n = 0$ and working alone to produce a patent of zero quality. In that case, his utility would be equal to zero.

A supporter with productivity Z_i^s optimally chooses which leader to join, as characterized by his utility Z^l. I assume that there is a frictionless market for leaders, such that supporters can join the leader they want. Moreover, I assume that there is perfect information, so that supporters internalize the optimal team size choice of leaders. Supporters maximize their utility, which is proportional to both their productivity and the productivity of the leader, net of a disutility coming from sharing the leader with other supporting inventors. This disutility is increasing in n and decreasing in Z_i^s, as to capture the fact that, for a given team size, a very productive supporter would benefit more from interacting briefly with the leader than a low productive supporter. Even though the model I am proposing is static, this reduced-form assumption would capture the intuition provided by a dynamic model where the productivity of supporters follows a law of motion increasing in the leader productivity but decreasing in team size.[2] Supporters solve the following maximization problem:

$$\max_{Z^l \geq \overline{Z}} U_i^s(Z^l; Z_i^s) = \big(Z_i^s\big)^\alpha \big(Z^l\big)^{1-\alpha} - \frac{n^*(Z^l)}{Z_i^s} \tag{2}$$

where $\alpha \in (0, 1)$ is the supporter's preference shifter for his own productivity. Note that $Z^l \geq \overline{Z}$ is a feasibility constraint: supporters cannot choose to join a leader that does not admit supporters in his team.

Before solving their respective maximization problems, agents with productivity Z_i choose whether to become leaders or supporters by maximizing the optimal utility that they would get in each role:

$$\max_{j \in \{l,s\}} \{U_i^l(n^*; Z_i), U_i^s(Z^{l*}; Z_i)\} \tag{3}$$

If agent i chooses to become a leader, his leader productivity is $Z_i^l = Z_i$; if he chooses to become a supporter, his supporter productivity is $Z_i^s = Z_i$.

2.2 Agents' optimal choices

Since agents choose whether to become leaders or supporters knowing their optimal choice for each option, the model is to be solved backwards. Consider an agent with productivity Z_i who has chosen to be a leader. The solution to his maximization problem is to manage a team of $n^*(Z_i)$ supporters, with

$$n^*(Z_i) = \left[\frac{\eta}{m}\right]^{\frac{1}{1-\eta}} Z_i \tag{4}$$

[2] An example would be $Z_{i,t+1}^s = (1-\delta)Z_{i,t}^s + f(Z_{i,t}^l, n)$, with δ being productivity depreciation, $f_1(\cdot,\cdot) > 0$ and $f_2(\cdot,\cdot) < 0$.

By plugging this value back into his utility function, this choice brings him the optimal utility value equal to

$$U_i^l(n^*; Z_i) = Z_i \left[\frac{\eta}{m}\right]^{\frac{1}{1-\eta}} \left[\left[\frac{\eta}{m}\right]^{\eta} - m\right] \tag{5}$$

Similarly, consider an agent with productivity Z_i who has chosen to be a supporter. The solution to his maximization problem is to join a leader characterized by productivity Z^{l*} with

$$Z^{l*}(Z_i) = (1-\alpha)^{\frac{1}{\alpha}} \left[\frac{m}{\eta}\right]^{\frac{1}{\alpha(1-\eta)}} (Z_i)^{\frac{1+\alpha}{\alpha}} \tag{6}$$

The utility level of a supporter associated to his optimal choice of leader is

$$U_i^s(Z^{l*}; Z_i) = (1-\alpha)^{\frac{1}{\alpha}} \left[\frac{m}{\eta}\right]^{\frac{1}{\alpha(1-\eta)}} \left[\left[(1-\alpha)\frac{m}{\eta}\right]^{1-\alpha} - 1\right] Z_i^{\frac{1}{\alpha}} \tag{7}$$

By plugging the two optimal utilities back into the maximization problem described in (3), it follows that agents become leaders if and only if their productivity Z_i is lower than a threshold $\underline{Z}$ or bigger than a threshold $\overline{Z}$:

$$Z_i \in \left[0, \underline{Z}\right] \qquad \text{or} \qquad Z_i > \overline{Z} \tag{8}$$

where $\overline{Z} \equiv \frac{\zeta^l}{\zeta^s}\zeta^{l,s}$ and $\underline{Z} \equiv \overline{Z}^{\frac{\alpha}{1+\alpha}} \left[(1-\alpha)\zeta^l\right]^{-\frac{1}{1+\alpha}}$ with

$$\zeta^l \equiv \left[\frac{m}{\eta}\right]^{\frac{1}{1-\eta}}$$
$$\zeta^s \equiv (1-\alpha)^{\frac{1}{1-\alpha}}$$
$$\zeta^{ls} \equiv \frac{\left[\left(\frac{\eta}{m}\right)^{\eta} - m\right]^{\frac{\alpha}{1-\alpha}}}{\left[\left[(1-\alpha)\frac{m}{\eta}\right]^{1-\alpha} - 1\right]^{\frac{\alpha}{1-\alpha}}}$$

Note that $\underline{Z}$ can be found as the inverse of the FOC of supporters evaluated at $Z^l = \overline{Z}$.

The optimal choice of team size of leaders is

$$n^*(Z_i) = \begin{cases} 0 & \text{if } Z_i \in \left[0, \underline{Z}\right] \\ \left[\frac{\eta}{m}\right]^{\frac{1}{1-\eta}} Z_i & \text{if } Z_i > \overline{Z} \end{cases} \tag{9}$$

A formal proof to Equation 9 is provided in Appendix C. The intuition behind it is that individuals with productivity $Z_i \in (0, \underline{Z}]$ would prefer to be supporters and match with a leader with productivity $Z^l < \overline{Z}$. However, this is not a feasible choice, as inventors with a productivity level marginally smaller than Z^l prefer to be supporters of a very productive leader, rather than leaders of unproductive supporters. Therefore, in equilibrium these inventors with productivity $Z_i \in (0, \underline{Z}]$ are better off by producing patents by themselves, rather than joining a leader with productivity $\overline{Z}$.

Note that, for the solution to be feasible, there must be a strictly positive share of agents in both roles. This is achieved by adding some additional restrictions to the size of the parameters m and η.

Lemma 1. *$\overline{Z} > 0$ if $m \in \left[1, \frac{\eta}{1-\alpha}\right)$ and $\eta > 1 - \alpha$.*

Proof. See Appendix C. □

2.3 Model predictions

The model provides several predictions that can be assessed by the stylized facts brought by the data on patents. I divide them into three different sets of contributions, based on whether they relate to the optimal choice of leaders, the optimal choice of supporters, or the interaction between inventors in the two roles.

Leaders ability and team size. The model shows that more productive leaders can overcome monitoring disutility and span of control better than less productive leaders, and therefore manage bigger teams of supporters (see also Akcigit et al. (2018)).

Lemma 2. *More productive leaders manage bigger teams.*

Proof. See Appendix C. □

In the empirical section I test this prediction and find that leaders' productivity is different and increasing across teams of different size at the 99% level.

Supporters ability and team choice. More productive supporters prefer to join very productive leaders, even though they face higher disutility due to having to share the leader with more co-workers. The intuition behind this result is that the marginal disutility from sharing a supervisor is smaller for very productive supporters, who would benefit more from interacting with the leader than a low productive supporter. By incurring smaller disutility costs, very productive supporters are thus able to join big teams led by very productive leaders.

Lemma 3.

(i) More productive supporters join more productive leaders.

(ii) More productive supporters join bigger teams.

Proof. See Appendix C. □

In the empirical section I test this prediction and find that the productivity of supporters joining teams of at least 4 members is bigger than the productivity of supporters joining a team of 2 members at the 95% level. The predictions of Lemma 3 lead directly to the result that there is a positive correlation between supporters productivity and patent quality. In fact, since patent quality increases in the leader's productivity and there is positive assortative matching between leaders and supporters, it follows that patent quality is positively correlated with supporters productivity.

Lemma 4. *Patent quality is positively correlated with the productivity of supporters working in that patent.*

Proof. See Appendix C. □

In Section 4 I test this prediction as well as the assumption that patent quality is correlated with leaders' productivity. I find a highly statistically significant positive correlation between patent quality and leaders' productivity as well as supporters' productivity, with the latter being smaller in magnitude than the former.

Non-linear relationship. The last set of predictions is the main contribution of this research and concerns the relationship between leaders' and supporters' productivity. Specifically, this model predicts a non-linear relationship between leaders' and supporters' productivity, such that supporters' productivity is increasing in the leader's productivity at a decreasing rate in leader's productivity or team size. The intuition behind these results is that a very productive leader collaborates with more productive supporters, as shown in Lemma 3, but he is not going to find supporters as productive as he is because they are likely leaders in other teams. Moreover, since more productive leaders manage bigger teams (Lemma 2), they are further away from the cut-off with respect to other leaders, and as such cannot match with supporters as productive as they are.

Lemma 5.

(i) Supporters productivity is increasing in the leader productivity, at a decreasing rate in leaders productivity.

(ii) Supporters productivity is increasing in the leader productivity, at a decreasing rate in team size.

Proof. See Appendix C. □

While formal proofs are provided in Appendix C, the main intuition behind these last two results is given by the existence of a cut-off productivity defining which agent becomes a supporter or a leader, combined with different functional forms pairing leaders to team size and supporters to leaders. Specifically, while team size increases linearly in leaders productivity, leaders and supporters match in a non-linear way. The inverse of the supporters FOC is an increasing concave function on leaders productivity, such that supporters productivity increases in leaders productivity, but at a decreasing rate.[3]

These results are tested empirically in Section 4, where I find that the correlation between leaders' and supporters' productivity decreases as team size increases, and the difference between these correlations is statistically significant at the 99% level.

3 Data

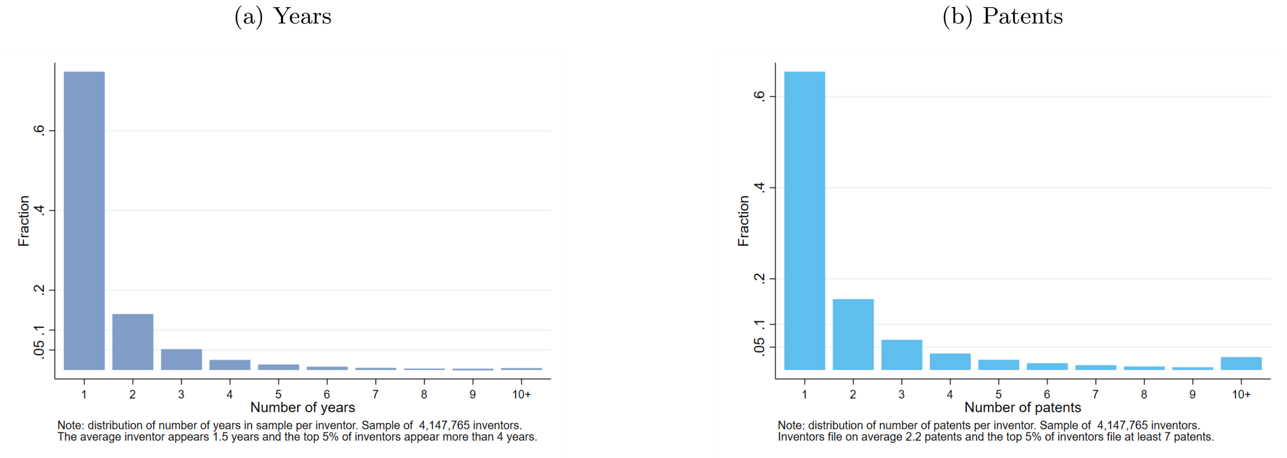
Validating the predictions of the model ideally requires data on teams, characteristics of their members, as well as characteristics of the output produced by the teams. The patent data collected by the OECD has all the required features. The OECD patent datasets, which are fully derived from the European Patent Office's Worldwide Statistical Patent Database (PATSTAT), collect information on patents filed at the EPO between 1978 and 2017. Specifically, I combine three of these datasets. First, the REGPAT database, January 2021 edition, which has information on both inventors and applicants of each patent. Second, the HAN database,

[3] Please refer to Figure 11 in Appendix C for a visual representation of the first order conditions.

July 2020 edition, which allows to harmonize the names of applicants over time. Third, the Quality Index database, January 2021 edition, which contains a wide set of quality indicators for each patent filed at the European Patent Office.[4] The EPO automatically assigns a unique identifier to each applicant, inventor, and patent, which allows me to merge the three datasets together. However, the identifier created for each inventor is based on unique combinations of name and address of residence. Since both of them are subject to typing mistakes, due for example to special and accentuated characters in the names or differences in abbreviations, it is not rare that the same inventor is assigned to distinct identifiers, one for each patent filed. As such, an important step in the data cleaning process is the disambiguation of inventors' names. I rely on the algorithm devised by Raffo and Lhuillery (2009) and I use homonymy and quasi-homonymy of both names and addresses of residence within a region to disambiguate inventors.[5] This cleaning process leaves me with a baseline dataset of 3,490,866 patents filed by 4,147,765 inventors and 508,087 applicants.

By disambiguating inventors' name and assigning them a unique reliable ID, I can track inventors over time and therefore exploit the panel feature of the dataset. Since inventors' production of patents during their lifetime is highly skewed, the dataset is unbalanced: the average inventor appears for 1.5 filing years, but 1% of inventors appear in the data for at least 7 years (Figure 1 panel (a)). Similarly, this unbalance can be seen from the number of patents attributed to each inventor: the average inventor files 2.2 patents, but 2.5% of inventors produce at least 10 patents (Figure 1 panel (b)).

Figure 1: Inventors presence in the data


(a) Years

(b) Patents

Note: distribution of number of years in sample per inventor. Sample of 4,147,765 inventors.
The average inventor appears 1.5 years and the top 5% of inventors appear more than 4 years.

Note: distribution of number of patents per inventor. Sample of 4,147,765 inventors.
Inventors file on average 2.2 patents and the top 5% of inventors file at least 7 patents.

While the disambiguation process is extremely helpful in tracking inventors over time, it does not allow to track inventors across space: an inventor filing two patents while residing in different addresses is considered as two different inventors, even after the disambiguation process. This is the main limitation of the data I am using and, to the best of my knowledge, a limitation of other data on patents filed at the EPO as well.

[4] For more information on this database, please refer to Squicciarini et al. (2013).

[5] Please refer to Appendix B for detailed information on the disambiguation process and data validation, as well as for further summary statistics not presented in the main text.

3.1 Teams

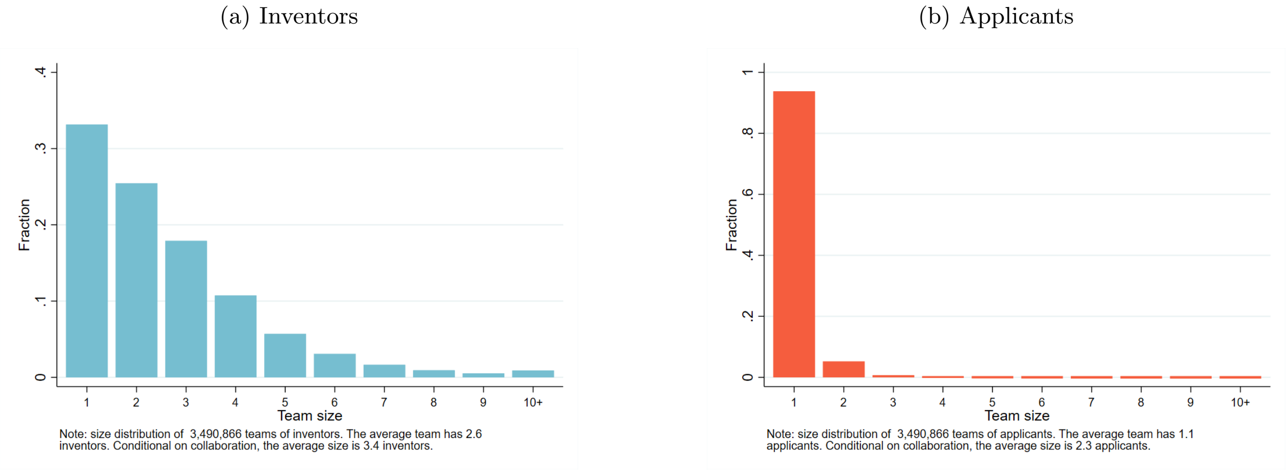
Since the EPO assigns a unique identifier to each patent, I can identify common patents between inventors and use them to define teams. This paper focuses on inventors' collaboration, but I could apply the same methodology to applicants as well. Table 1 and Figure 2 show that collaboration is not a rare event among inventors. 89% of inventors in the data has filed a patent with another inventor at least once in their lifetime. Moreover, only roughly 30% of patents are produced by an inventor working alone, while the remaining ones are produced in teams with average size of 3.4 inventors.[6] This is in stark contrast with the behaviour of applicants, as almost all patents are produced by an applicant alone.

Table 1: Collaborators among inventors and applicants

	Inventors	Applicants
Total	4,147,765	508,087
Collaborators	3,683,611	154,631
Percentage	89%	30%

Note: Sample of 3,490,866 patents filed at EPO between 1978 and 2017 for which both applicant and inventors are known. *Collaborators* are defined as agents being co-inventor or co-applicant in at least one patent.

Figure 2: Team size distribution


(a) Inventors

(b) Applicants

Note: size distribution of 3,490,866 teams of inventors. The average team has 2.6 inventors. Conditional on collaboration, the average size is 3.4 inventors.

Note: size distribution of 3,490,866 teams of applicants. The average team has 1.1 applicants. Conditional on collaboration, the average size is 2.3 applicants.

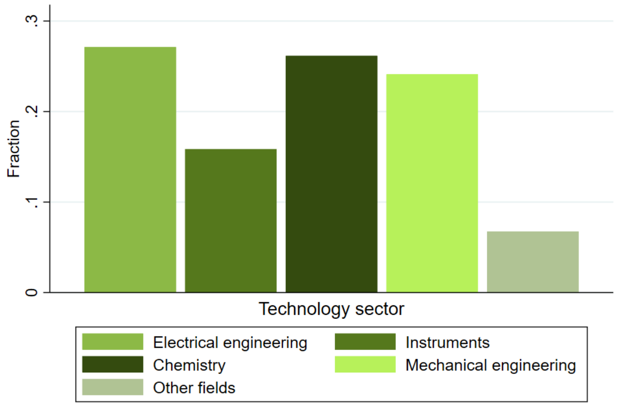
3.2 Technology and quality

Patents are pertinent to different fields and are heterogeneous in their quality. Information on these two aspects come from the Quality Index Database, which assigns each patent to one of the 35 technology fields defined by Schmoch (2008) and provides several measures of quality. The 35 technology fields rely on the International Patent Classification (IPC) codes contained in the patent documents.

[6] As an additional validation step on top of the one described in Appendix B, the summary statistics shown on panel a of Figure 2 match the information reported on Akcigit et al. (2018) for patents filed at the EPO between 1977 and 2010. In their study, the average team size is 2.6 inventors, and 3.4 inventors when conditioning on multi-inventor patents.

Figure 3 shows the distribution of patents in the sample across technology fields, aggregated into 5 technology sectors. The data appear to be fairly balanced, especially between electrical engineering, chemistry, and mechanical engineering sectors.[7]

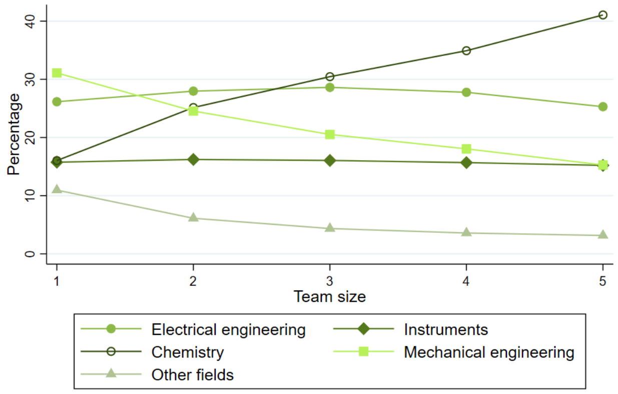
Figure 3: Patents by technology sector



Note: distribution of patents by technology sector. Sample of 3,489,209 patents. 2,793 patents with unknown technology field have been excluded.

The distribution of patents by technology sector varies across team size of inventors (Figure 4), with Chemistry being more common among patents of at least five inventors rather than solo inventors. This already hints towards the fact that bigger teams of inventors may be substantially different than smaller teams.[8]

Figure 4: Technology sector by team size



Note: technological category as percentage of patents developed by inventors' team size. 5 stands for teams with at least five inventors. Sample of 3,489,209 patents; 2,793 patents with unknown technological category are excluded.

Being able to measure quality of patents is an important component of this research. The vast majority of quality measures rely on citations. To ease the analysis throughout the paper, I group the indices into two categories. I label the first category *Citation indices*, as it comprises measures relying only on citations, and the second category *Technology indices*, as these measures also rely on the International Patent Classification

[7] For a detailed distribution of patents across all 35 technology fields, as well as a description of how technology fields are aggregated into sectors, please refer to Table 24 in Appendix B.

[8] A similar pattern can be seen for applicants' team size (Figure 10 in Appendix B).

(IPC) codes contained in the patent documents.

The most important index in the first category is forward citations: the number of citations received by a patent in a given window of time. The count also includes self-citations, following the findings of Hall et al. (2005) suggesting that they are more valuable than citations from external patents when assessing their market value. Citations are classified in different categories. Of particular importance are citations falling in the X and Y categories, as they question the inventive step of the filed patent. The Quality Index Database provides measures based on forward citations counting all categories or only the X and Y categories, and in different time windows (5 or 7 years). Since forward citations received in a 5 year window in all categories is the measure of quality most used in the literature, it is going to be my baseline measure of quality throughout the paper, unless stated otherwise. Another important measure is backward citations, which refers to the number of citations done by the patent. NPL citations instead refers to citations made by the patent to Non-Patent Literature: among the others, scientific papers, conference proceeding, and databases.

Among the technology indices, scope refers to the number of distinct 4-digit subclasses of the IPC codes the invention is allocated to. A patent has a broader scope if it falls into more technological subclasses. The generality measure refers to the range of 4-digit subclasses of IPC codes that cite the patent: a patent is more general if it is cited by other patents belonging to a wide range of codes. Originality refers to the range of 4-digit subclasses of IPC that are cited by the patent. It can be interpreted as the inverse of an Herfindahl index of concentration of citations. Lastly, radicalness refer to the range of 4-digit subclasses of IPC that are cited by the patent and that differ from the subclasses of the patent itself. The last two indices differ in the fact that radicalness only accounts for technological categories that are different from those of the patent under consideration, and does not account for the distribution of citations. Formally, for a given patent p, the last two indices are defined as follows:

$$\text{Originality}_p = 1 - \sum_{j}^{n_p} s_{pj}^2 \qquad \text{Radicalness}_p = \sum_{j}^{n_p} \frac{\text{CT}_j}{n_p}$$

where s_{pj} is the percentage of citations made by patent p to technology class j, n_p is the total number of IPC classes cited by p, and CT_j denotes the count of IPC-4 digit codes of patent j cited in patent p that are not allocated to patent p.

All indices are normalized with respect to the maximum value exhibited by other patents in the same cohort, which is defined by pairs of filing year and technology field. This normalization prevents the measure to be affected by any potential bias linked to technology or time. Table 2 provides summary statistics of the quality indices. The distribution appears to be highly skewed, as only a handful of patents are highly successful.

3.3 Specialization and productivity

Being able to track inventors over time, together with knowing the technology field of a patent and the amount of citations it receives, allows me to define specialization and productivity of inventors. I define the specialization of an inventor as the most frequent technology field of his past patents. This definition allows the

Table 2: Patents' quality indices

	N	Mean	Min	Max	SD
Citation indices:					
Forward citations (5y)	2,774,730	.033	0	1	.075
Forward citations (7y)	2,500,051	.038	0	1	.079
Forward citations (5y) – XY	2,774,730	.024	0	1	.077
Forward citations (7y) – XY	2,500,051	.030	0	1	.083
Backward citations	3,490,866	.095	0	1	.098
NPL citations	3,490,866	.021	0	1	.064
Technology indices:					
Scope	3,490,866	.209	0	1	.128
Generality	1,119,158	.407	0	1	.321
Originality	3,397,393	.699	0	1	.244
Radicalness	3,490,866	.307	0	1	.263

Note: quality indices for 3,490,866 patents filed between 1978 and 2017. Measures relying on a 5-year time window are available only up until 2012; those relying on a 7-year time window are available only up until 2010. All indices have been normalized with respect to the maximum value exhibited by other patents in the same cohort (defined as filing year and technology field).

specialization of each inventor to change over time, depending on his patenting activity. Productivity of an inventor is instead measured by the cumulative sum of normalized citations of his past patents. This definition of productivity implicitly assumes it to be weakly increasing over time. Note that, given their lagged nature, both measures are available only for inventors that appear in the data in at least two different filing years. Since information about the age of inventors is not available, I use the filing years to measure age of inventors at a given point in time as the difference between that year and the filing year of the first patent filed by the inventor.

Apart from being interesting per se, the measures of specialization, productivity, and age play a key role in distinguish inventors within the same team into leaders and supporters. Throughout the paper, I identify leader inventors of each patent as those specialized in the technology field of the patent and with the highest productivity. If no inventor in the team is specialized in the same technology field as the one of the patent, then the leader inventor is simply the most productive member of the team. If there is no specialized inventor and all inventors have zero productivity, then I define the leader inventor as the oldest one. Note that a team may have multiple leaders and inventors filing a patent alone are trivially leaders of that patent. Within a patent, all inventors that are not leaders are identified as supporting inventors, or simply supporters.

Table 3 shows the summary statistics of productivity for all inventors in each year. Since this measure is derived from forward citations, it carries its skewness. Moreover, since forward citations are available only up until 2012 and the productivity measure is lagged by one year, inventors' productivity is available only for the time window 1979-2013. The second and third row show the distribution of productivity for leader and supporting inventors, respectively. Note that if an inventor appears in multiple patents in the same filing year, he could be leader in some patents and supporter in others; this explains why the sum of the total number of leaders and supporters is bigger than the full sample size.

Table 3: Inventors' productivity

	N	Mean	Min	Max	sd
Full sample	5,211,737	.071	0	9.989	.325
Leaders	3,261,129	.085	0	9.989	.371
Supporters	2,215,049	.067	0	9.988	.077

Note: productivity distribution for 5,211,737 inventor × year observations in the time window 1979-2013.

Throughout the paper, I show that the main results hold when using the definition of leader inventors adopted by Akcigit et al. (2018): leaders are the most productive members of the team, or the oldest members of the team if all inventors have zero productivity. This definition allows for multiple leaders within a team as well. Moreover, I perform robustness checks using the cumulative count of past patents as an alternative definition of productivity. This alternative definition accounts only for the quantity of past patents, not for their quality.[9]

4 Empirical strategy and results

4.1 Quality by team size and productivity

I first investigate whether large team of inventors produce better patents, even when controlling for their average productivity. I start by estimating the following specification:

$$\text{Quality}_p = \sum_{s=1}^{5} \beta_s \, \mathbf{I}\{\text{size}_p = s\} + \beta \, \text{Z}_p + \delta_{a(p)} + \delta_{t(p)} + \delta_{y(p)} + \varepsilon_p \tag{10}$$

where p indicates the patent, $a(p)$ the applicant of the patent, $t(p)$ the technology field of the patent, and $y(p)$ the filing year of the patent. $\mathbf{I}\{\text{size}_p = s\}$ are indicator variables taking value of 1 if the patent is developed by a team of s inventors, where $s = 5$ represents teams of at least five inventors. Z_p is the average productivity of inventors in patent p, while $\delta_{a(p)}$, $\delta_{t(p)}$, and $\delta_{y(p)}$ are fixed effects for applicant, technology field, and filing year of the patent, respectively. Errors ε_p are three-way clustered at the applicant, technology field, and year level.

The results of this specification are shown in Table 4. On average, patents filed by more inventors or by a team with higher productivity show higher quality levels. This positive correlation holds with other measures of quality or productivity, with the notable exception of radicalness (Table 12, Table 13, and Table 14 in Appendix A). The team size coefficients shown in Table 4 use $\mathbf{I}\{\text{size}_p = 1\}$ as baseline and are statistically different from each other at the 99% level when a Wald test on any pairwise combination is performed. The correlation is higher than what it seems at first sight; for example, the increase in citations received by a patent filed by two inventors with respect to a patent filed by a solo-inventor is in the order of magnitude of 10% of the average amount of citations. Similarly, a one standard deviation increase in the average productivity of inventors in the team corresponds to an increase in citations equal to 9% of the average amount of citations.

[9] For a summary statistics of the distribution of this alternative measure, please refer to Table 23 in Appendix B.

Table 4: Citation indices by team size

	(1) Forward (5y)	(2) Forward (7y)	(3) Forward (5y) – XY	(4) Forward (7y) – XY	(5) Backward	(6) NPL
$\mathbf{I}\{\text{size}_p = 2\}$	0.352***	0.389***	0.253***	0.281***	0.145***	0.196***
	(0.031)	(0.034)	(0.024)	(0.028)	(0.025)	(0.027)
$\mathbf{I}\{\text{size}_p = 3\}$	0.704***	0.786***	0.519***	0.606***	0.340***	0.302***
	(0.063)	(0.069)	(0.044)	(0.048)	(0.042)	(0.038)
$\mathbf{I}\{\text{size}_p = 4\}$	1.051***	1.178***	0.826***	0.971***	0.553***	0.376***
	(0.080)	(0.089)	(0.059)	(0.068)	(0.058)	(0.043)
$\mathbf{I}\{\text{size}_p = 5\}$	1.735***	1.921***	1.438***	1.641***	0.998***	0.556***
	(0.140)	(0.159)	(0.109)	(0.129)	(0.091)	(0.060)
Z_p	0.696***	0.763**	0.555***	0.623***	0.426***	0.213***
	(0.082)	(0.082)	(0.071)	(0.086)	(0.098)	(0.072)
Applicants FE	X	X	X	X	X	X
Tech. field FE	X	X	X	X	X	X
Year FE	X	X	X	X	X	X
Applicants clusters	149,121	137,108	149,121	137,108	155,032	155,032
Tech. fields clusters	35	35	35	35	35	35
Year clusters	35	33	35	33	36	36
N	2,503,053	2,248,873	2,503,053	2,248,873	2,632,947	2,632,947
Adj. R^2	0.101	0.115	0.068	0.075	0.324	0.141

Summary statistics:						
Dep. var. average	3.400	3.945	2.524	3.069	9.912	2.240
(std. dev.)	(7.657)	(8.088)	(7.836)	(8.439)	(9.866)	(6.708)
Z_p average	0.144	0.141	0.144	0.141	0.146	0.146
(std. dev.)	(0.451)	(0.442)	(0.451)	(0.442)	(0.453)	(0.453)

Note: The dependent variable is a quality index. Quality indices have been rescaled to take value in $[0, 100]$. Coefficients are with respect to $\mathbf{I}\{\text{size}_p = 2\}$. Z_p is the average productivity of inventors in the team. Standard errors in parentheses are robust to three-way clustering at the applicant, technology field, and year levels. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

I further examine the correlation between patent quality and the productivity of the inventors producing it, through the lenses of the theoretical framework described in Section 2. In the model, patent quality is defined as a function which is increasing in team size and leader's productivity. Moreover, one of the predictions of the model is that supporters' productivity is positively correlated with the quality of the patent they are producing (Lemma 4). Therefore, I estimate a slightly different specification, where I disentangle the average productivity of inventors of a patent into leader's productivity and average supporters' productivity:

$$\text{Quality}_p = \sum_{s=2}^{5} \beta_s \, \mathbf{I}\{\text{size}_p = s\} + \beta_1 \, \mathrm{Z}_p^l + \beta_2 \, \mathrm{Z}_p^s + \delta_{a(p)} + \delta_{t(p)} + \delta_{y(p)} + \varepsilon_p \tag{11}$$

where Z_p^l is the productivity of a leader of patent p, while Z_p^s is the average productivity of his supporters working in patent p. Please note that in order to distinguish between leaders and supporters, all patents produced by solo inventors must be excluded ($s = 1$).

The results of this specification seem to confirm the positive relation between inventors' productivity and patent quality, even when inventors are divided into leaders and supporters (Table 5). Since the team size used as reference in this specification is different compared to the previous specification, the magnitude of the results of this specification are not directly comparable with those reported in Table 4. Nevertheless, the magnitude of the coefficient is still quite sizeable: for example, the increase in citations received by a patent filed by three inventors with respect to a patent filed by two inventors is in the order of magnitude of 9.5% of the average amount of citations received by patents with at least two inventors.
Similar results can be obtained by identifying leaders as the most productive inventors within a team regardless of their specialization (Table 15 in Appendix A). The positive correlation between leaders or supporters' productivity and patent quality weakens if quality is measured by the technology indices, while the relationship between team size and quality is still positive and statistically significant for most measures, regardless of the definition of leaders used (Table 16 and Table 17 in Appendix A).

4.2 Productivity by team size

The model in Section 2 predicts that both leaders and supporters sort into bigger teams according to their productivity (Lemma 2 and Lemma 3). I investigate whether bigger teams are formed on average by more productive inventors by estimating the following specification:

$$\mathrm{Z}_{ip} = \sum_{s=1}^{5} \beta_s \, \mathbf{I}\{\text{size}_p = s\} + \delta_{a(p)} + \delta_{t(p)} + \delta_{\text{age}(i)} + \delta_{c(i)} + \varepsilon_{ip} \tag{12}$$

where Z_{ip} is productivity of inventor i working on patent p and $\mathbf{I}\{\text{size}_p = s\}$ is once again an indicator variable for the team size. The specification also comprises a battery of fixed effects: $\delta_{a(p)}$ and $\delta_{t(p)}$ are fixed effects for the applicant and the technology field of the patent, while $\delta_{\text{age}(i)}$ and $\delta_{c(i)}$ are fixed effects for the age and the cohort of the inventor, which is defined as the first year in which he appears in the data. Standard errors are clustered at the inventor, applicant, and technology field level.

Table 5: Citation indices by team size

	(1) Forward (5y)	(2) Forward (7y)	(3) Forward (5y) – XY	(4) Forward (7y) – XY	(5) Backward	(6) NPL
$\mathbf{I}\{\text{size}_p = 3\}$	0.340***	0.383***	0.243***	0.305***	0.183***	0.087***
	(0.043)	(0.048)	(0.030)	(0.036)	(0.026)	(0.021)
$\mathbf{I}\{\text{size}_p = 4\}$	0.650***	0.740***	0.539***	0.655***	0.380***	0.155***
	(0.061)	(0.066)	(0.042)	(0.050)	(0.049)	(0.037)
$\mathbf{I}\{\text{size}_p = 5\}$	1.308***	1.446***	1.084***	1.246***	0.767***	0.326***
	(0.109)	(0.127)	(0.083)	(0.100)	(0.076)	(0.064)
Z_p^l	0.437***	0.470***	0.366***	0.408***	0.199***	0.094***
	(0.064)	(0.065)	(0.057)	(0.070)	(0.043)	(0.032)
Z_p^s	0.394***	0.426***	0.299***	0.325***	0.245*	0.204***
	(0.090)	(0.079)	(0.086)	(0.083)	(0.123)	(0.070)
Applicants FE	X	X	X	X	X	X
Tech. field FE	X	X	X	X	X	X
Year FE	X	X	X	X	X	X
Applicants clusters	50,508	45,554	50,508	45,554	52,936	52,936
Tech. fields clusters	35	35	35	35	35	35
Year clusters	34	32	34	32	35	35
N	1,098,657	964,816	1,098,657	964,816	1,168,140	1,168,140
Adj. R^2	0.108	0.122	0.079	0.089	0.299	0.151
Summary statistics:						
Dep. var. average	3.589	4.128	2.924	3.506	8.797	2.269
(std. dev.)	(7.945)	(8.383)	(8.241)	(8.805)	(9.349)	(6.544)
Z_p^l average	0.473	0.467	0.473	0.467	0.475	0.475
(std. dev.)	(1.015)	(0.998)	(1.015)	(0.998)	(1.020)	(1.020)
Z_p^s average	0.103	0.100	0.103	0.100	0.104	0.104
(std. dev.)	(0.349)	(0.342)	(0.349)	(0.342)	(0.350)	(0.350)

Note: The dependent variable is a quality index. Quality indices have been rescaled to take value in $[0, 100]$. Coefficients are with respect to $\mathbf{I}\{\text{size}_p = 2\}$. Z_p^l is the productivity of the leader of patent p, while Z_p^s is the average productivity of supporting inventors working in patent p. Standard errors in parentheses are robust to three-way clustering at the applicant, technology field, and year levels. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 6: Productivity by team size

	(1) Productivity	(2) Productivity	(3) Productivity	(4) Productivity
$\mathbf{I}\{\text{size}_p = 2\}$	0.004	0.015***		
	(0.005)	(0.005)		
$\mathbf{I}\{\text{size}_p = 3\}$	0.006	0.031***	0.014***	0.001
	(0.006)	(0.008)	(0.003)	(0.001)
$\mathbf{I}\{\text{size}_p = 4\}$	0.010	0.049***	0.031***	0.005**
	(0.007)	(0.010)	(0.006)	(0.002)
$\mathbf{I}\{\text{size}_p = 5\}$	0.012	0.078***	0.058***	0.011**
	(0.008)	(0.013)	(0.010)	(0.004)
Applicant FE	X	X	X	X
Tech. field FE	X	X	X	X
Age FE	X	X	X	X
Cohort FE	X	X	X	X
Inventors clusters	3,347,318	2,188,822	1,857,308	1,587,020
Applicants clusters	283,256	242,233	189,490	93,790
Tech. fields clusters	35	35	35	35
Sample	Full	Leaders	Leaders; $s > 1$	Supporters
N	7,320,271	4,128,317	3,272,891	3,124,065
Adj. R^2	0.229	0.274	0.277	0.204

Summary statistics:				
Dep. var. average	0.143	0.169	0.175	0.109
(std. dev.)	(0.556)	(0.630)	(0.653)	(0.438)

Note: The dependent variable is the productivity of inventors working on a patent. Coefficients with respect to $\mathbf{I}\{\text{size}_p = 1\}$ for the first two columns, with respect to $\mathbf{I}\{\text{size}_p = 2\}$ for the last two columns. Column (2) considers only leaders. Column (3) restricts the sample to leaders in teams of at least two inventors. Column (4) considers only supporters. Standard errors in parentheses are robust to three-way clustering at the applicant, technology field, and inventor levels. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Average productivity of inventors does not seem to change significantly by team size (column (1) of Table 6). However, distinguishing between leaders and supporters unveils a different picture. The second and third column show the results of estimating this specification on the sample of leaders only. The results suggest that the average productivity of leaders increases in team size, whether I consider all leaders (column (2)) or only leaders in teams with at least two inventors (column (3)). Specifically, leaders of teams with size 4 are on average more productive than leaders of teams with size 2, and this difference corresponds to 18% of the average productivity of leaders in teams of at least two inventors. This relation is less visible when I consider the subsample of supporters (column (4) of Table 6), even though it is still statistically significant at the 95% level for teams of at least size 4. These results confirm the predictions brought by the model described in Section 2. A similar pattern holds when defining leaders as in Akcigit et al. (2018) or when using the cumulative count of past patents as measure of productivity (Table 18 and Table 19 in Appendix A). While the results for the full sample and the leaders subsamples are robust throughout all alternative measurements, the positive correlation between supporters' productivity and sample size becomes not statistically significant when productivity is measured by past patents. Overall, these robustness checks point to the fact that the quality of past patents produced, not only their quantity, may be an important component for the assessment of inventors productivity.

4.3 Assortative matching

The main focus of this research is on the possible assortative matching mechanism between leader and supporting inventors. The simple theoretical framework described in Section 2 suggests that more productive supporters join teams managed by more productive leaders (Lemma 3), but the correlation between their productivities decreases in magnitude as team size increases (Lemma 5).
I test if this prediction is displayed in the data by estimating the following specification:

$$\mathrm{Z}_p^s = \beta\ \mathrm{Z}_p^l + \delta_{a(p)} + \delta_{t(p)} + \delta_{y(p)} + \delta_{s(p)} + \varepsilon_p \tag{13}$$

where Z_p^s is the average productivity of supporters in patent p and Z_p^l is the productivity of leader in patent p. $\delta_{a(p)}$, $\delta_{t(p)}$, $\delta_{y(p)}$, and $\delta_{s(p)}$ are fixed effects for applicant, technology field, year, and inventors' team size of patent p. The error term ε_p is clustered at the filing year, applicant, and technology field level.
To check whether the relationship between leader's and supporters' productivity varies by team size, I estimate a specification enhanced with an interaction term between Z_p^l and team size. The model suggests that these interaction terms would be negative and increasing in absolute value over team size: supporters productivity is increasing in leader productivity, but at a decreasing rate as team size becomes bigger (Lemma 5).

The results are shown in Table 7, where column (1) displays the estimation of Equation 13 while column (2) shows the results for the specification augmented with the interaction terms. Both columns show evidence of assortative matching between leaders and supporters within patents of the same size and technology field, filed by the same applicant in the same year. On average, a one standard deviation increase in the productivity of the leader corresponds to an increase in the average productivity of his supporters of the order of magnitude of 160% of the average productivity of all supporters in the sample. These results are consistent with the predictions declared in Lemma 3. However, the specification shown on column (2) unveils a more nuanced relationship between leaders and supporters productivity that depends on size. Specifically, a one standard deviation in leader productivity in teams of two inventors corresponds to an increase in average supporters' productivity equivalent to 177% of the average supporters productivity. The magnitude of the change lowers to 145% for leaders in teams of size 5. The results of the interaction terms suggest that having a highly productive leader is more relevant for smaller teams than for bigger ones. This is closely linked to the predictions of Lemma 5: since the inverse of supporters optimal choice of leaders is concave in leaders productivity, supporters productivity increases at a lower rate in leader's productivity for teams of bigger size. This result will be tested directly in the next subsection. The results shown in Table 7 are quite robust to alternative definitions of leaders and/or productivity: leader's and supporters' productivities are still positively correlated, but the interaction coefficients lose statistic significance when productivity is measured by patents (Table 20 in Appendix A). This once

Table 7: Supporters' productivity and leader's productivity

	(1) Sup. Prod.	(2) Sup. Prod.
Z_p^l	0.163*** (0.011)	0.180*** (0.014)
$\mathrm{Z}_p^l \times \mathbf{I}\{\text{size}_p = 3\}$		-0.012** (0.006)
$\mathrm{Z}_p^l \times \mathbf{I}\{\text{size}_p = 4\}$		-0.017*** (0.005)
$\mathrm{Z}_p^l \times \mathbf{I}\{\text{size}_p = 5\}$		-0.033*** (0.009)
Applicant FE	X	X
Tech. field FE	X	X
Year FE	X	X
Size FE	X	X
Applicants clusters	52,936	52,936
Tech. fields clusters	35	35
Filing clusters	35	35
N	1,168,140	1,168,140
Adj. R^2	0.295	0.296

Interpretation of the effect:		
Coeff. as % of avg. dep. var.	157%	173%
1 sd as % of avg. dep. var.	160%	177%

Note: The dependent variable is the average productivity of supporters in the patent. Z_p^l is the productivity of the leader of the patent. Coefficients for the interactions in column (2) are estimated with respect to $\mathbf{I}\{\text{size}_p = 2\}$. Standard errors in parentheses, robust to three-way clustering at the applicant, technology field, and filing year levels. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

again suggests that the quality of patents is an important component of inventors productivity.

A potential issue comes directly from the definition of leader itself. Since leaders are defined as the most productive inventors within a team, specialized in the same technology field as the patent they are working on, the leader productivity might increase mechanically in supporters productivity: if supporters' productivity increases, it is more likely that one of them becomes the leader. This mechanical bias might be especially strong when leaders are identified based on productivity only, regardless of their sectoral specialization, as in Akcigit et al. (2018). To address this issue, I perform a series of robustness checks, whose results are reported in Table 8. The first one is displayed in column (1) and employs a *leave-out* specification, where I compute Z_p^s as the average productivity of supporters in patent p excluding the most productive supporter in the team. The positive relation between leaders and supporters decreases in magnitude, but it is still statistically significant and sizeable, as one standard deviation in leaders' productivity is associated with an increase in average productivity of supporters corresponding to 131% of the average in the sample. A second set of robustness checks involves using a

Table 8: Robustness checks

	(1) Sup. Prod. leave-out	(2) Sup. Prod.	(3) Sup. Prod.	(4) Sup. Prod. leave-out
Z_p^l	0.085*** (0.008)	0.022*** (0.001)	0.025*** (0.001)	0.004*** (0.001)
$\mathrm{Z}_p^l \times \mathbf{I}\{\text{size}_p = 3\}$			-0.002** (0.001)	
$\mathrm{Z}_p^l \times \mathbf{I}\{\text{size}_p = 4\}$			-0.003*** (0.001)	
$\mathrm{Z}_p^l \times \mathbf{I}\{\text{size}_p = 5\}$			-0.007*** (0.001)	
Applicant FE	X	X	X	X
Tech. field FE	X	X	X	X
Year FE	X	X	X	X
Size FE	X	X	X	X
Applicants clusters	17,931	73,606	73,606	27,967
Tech. fields clusters	35	35	35	35
Filing clusters	35	35	35	35
Data	Real	Placebo	Placebo	Placebo
N	394,587	1,373,553	1,373,553	402,608
Adj. R^2	0.256	0.034	0.034	0.028

Interpretation of the effect:				
Coeff. as % of avg. dep. var.	96%	69%	78%	27%
1 sd as % of avg. dep. var.	131%	70%	79%	33%

Note: Column (1) shows results of a leave-out specification, where the dependent variable is the average productivity of supporters computed excluding the most productive supporters within the team. Column (2)-(4) shows results using placebo data: all inventors are reassigned randomly to teams and divided into leaders and supporters within each patent. The dependent variable for column (2) and (3) is the average productivity of supporters. Column (4) uses a leave-out specification: the dependent variable is the average productivity of supporters, excluding the most productive ones within the team. Z_p^l is the productivity of the leader of the patent. Coefficients for the interactions in column (3) are estimated with respect to $\mathbf{I}\{\text{size}_p = 2\}$. Standard errors in parentheses are robust to three-way clustering at the applicant, technology field, and filing year levels. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

placebo dataset to estimate the specifications described so far. Specifically, I reassign inventors, with their real productivities, randomly to teams and subsequently distinguish inventors between leaders and supporters within each team. Column (2) and column (3) of Table 8 show the results for Equation 13 and the one augmented with all the interactions, respectively. There is still evidence of assortative matching between leaders' and supporters' productivity, which is entirely due to the mechanical bias embedded in the definition of leaders itself. However, by comparing these results with those shown in Table 7, I can quantify that at most only 45% of the positive relation between leaders' and supporters' productivity can be accounted for by this mechanical bias. This leaves out a still sizeable lower bound positive relationship equivalent to 90% of the average of the dependent variable: on average, a one standard deviation in the leaders' productivity is associated to roughly doubling the average supporter productivity. Moreover, using the placebo data, the difference in the correlation between leader's and supporters' productivity across teams of distinct size shrinks substantially. The difference between the interaction coefficients of Table 7 and Table 8 points to the fact that the match between leaders and supporters cannot be the result of randomness, but it is dictated by a specific assortative matching.

Lastly, column (4) reports the results for the *leave-out* average supporter productivity using placebo data. Since the magnitude of the coefficient decreases sharply, this specification confirms the suspicion that most of the mechanical bias comes from the most productive supporters within the team. Comparing the results shown in column (4) of Table 8 with those reported in column (1) of the same table once again ensures that only less than a third of the positive relationship between leaders' and supporters' productivity is accounted for by the mechanical contribution of the definition. Therefore, even in the most conservative case, it still holds true that a one standard deviation increase in leaders' productivity is associated, on average, to roughly doubling the average productivity of supporters.

4.4 Inventors' age

Since age could be a proxy for experience and productivity, I inspect whether the average age of inventors in a team increases with its size. To do so, I estimate the following specification:

$$\text{Age}_{ip} = \sum_{s=1}^{5} \beta_s \, \mathbf{I}\{\text{size}_p = s\} + \delta_{a(p)} + \delta_{t(p)} + \delta_{y(p)} + \varepsilon_{ip} \tag{14}$$

where Age_{ip} is the age of inventor i working on patent p, $\mathbf{I}\{\text{size}_p = s\}$ is once again an indicator function for team size, and $\delta_{a(p)}$, $\delta_{t(p)}$, and $\delta_{y(p)}$ are fixed effects for applicant, technology field, and filing year associated to a patent p. As before, Age_{ip} is measured as the difference between the filing year of the patent, $y(p)$, and the first year in which inventor i appears in the data.

The estimates (first column of Table 9) suggests that larger teams are composed by younger inventors, after controlling for applicant, technology field, and year. Inventors working in teams of 4 are, on average, younger than solo inventors by half of a year, which corresponds to 17% of the average age of all inventors. This result seems to be driven mainly by supporters. In fact, considering leaders and supporters separately leads to a different pattern (second and third column of Table 9, respectively). Both leaders' and supporters' average age

Table 9: Age by team size

	(1) Inv. Age	(2) Inv. Age	(3) Inv. Age	(4) Inv. Age
$\mathbf{I}\{\text{size}_p = 2\}$	-0.344***	0.107***		
	(0.038)	(0.022)		
$\mathbf{I}\{\text{size}_p = 3\}$	-0.475***	0.263***	0.153***	0.012
	(0.049)	(0.034)	(0.016)	(0.012)
$\mathbf{I}\{\text{size}_p = 4\}$	-0.552***	0.405***	0.289***	0.035**
	(0.054)	(0.043)	(0.028)	(0.017)
$\mathbf{I}\{\text{size}_p = 5\}$	-0.681***	0.609***	0.493***	0.047**
	(0.059)	(0.060)	(0.051)	(0.021)
Inventor FE	-	-	-	-
Applicant FE	X	X	X	X
Tech. field FE	X	X	X	X
Year FE	X	X	X	X
Inventors clusters	4,001,427	2,571,431	2,197,102	1,942,554
Applicants clusters	333,582	283,862	224,497	113,299
Tech. fields clusters	35	35	35	35
Sample	Full	Leaders	Leaders; $s > 1$	Supporters
N	8,976,320	4,986,964	3,994,544	3,908,482
Adj. R^2	0.153	0.271	0.271	0.104

Summary statistics:				
Dep. var. average	3.200	3.490	3.455	2.780
(std. dev.)	(3.958)	(4.344)	(4.318)	(3.293)

Note: The dependent variable is inventors' age. Coefficients with respect to $\mathbf{I}\{\text{size}_p = 1\}$ for the first two columns, with respect to $\mathbf{I}\{\text{size}_p = 2\}$ for the last two columns. Column (2) considers only leaders. Column (3) considers only leaders working in a team of at least two inventors. Column (4) considers only supporters. Standard errors in parentheses, robust to three-way clustering at the applicant, technology field, and inventor levels. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

is increasing in team size, but the relationship is smaller and not very significative for supporters. This result is robust to defining leaders according to Akcigit et al. (2018) (Table 21 in Appendix A).

5 Evidence on dynamic sorting

Although the theoretical framework described in this paper is static, it can be seen as the starting point for a more complex dynamic setting. Therefore, this section exploits the panel data feature of the dataset.

Of the 3.5 million inventors with known productivity, 33% never play the role of leader in a patent. Considering the sub-sample of inventors appearing for more than one year, the figure decreases to 22%. This seems to suggest that, even though productivity may change over time, there are some inventors that never make the cut to become leaders.

Table 10 provides information for the remaining inventors; those that are leaders at least once in their lifetime. It shows the distribution of two variables characterizing the minimum requirements for becoming leaders. The

Table 10: Minimum age and minimum experience for leadership

	N	Mean	Min	Max	SD
Minimum age:					
Full sample	2,328,119	1.290	1	32	1.261
Age > 1	626,835	3.940	2	35	2.913
Productivity > 0	413,068	4.368	2	35	3.250
Minimum experience:					
Full sample	2,328,119	0.195	0	205	0.976
Age > 1	626,835	1.793	1	205	1.947
Productivity > 0	413,068	2.451	1	205	2.588

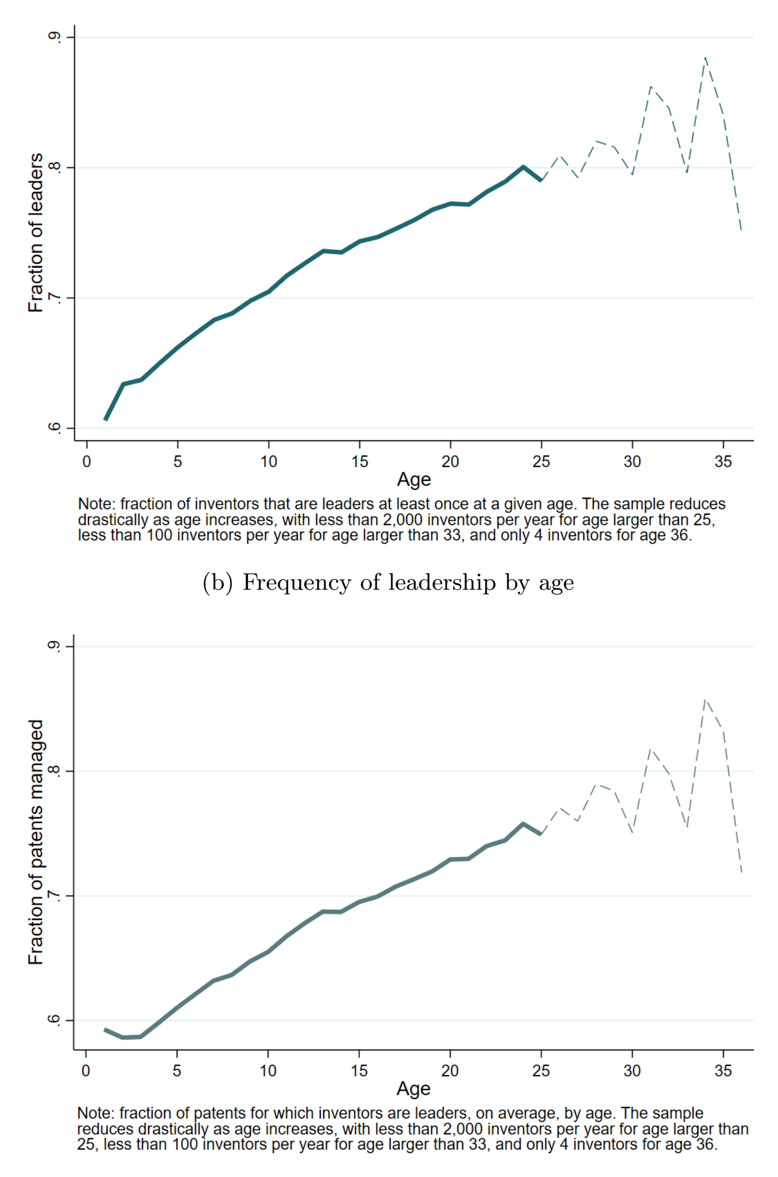
Note: *Minimum age* is the age at which an inventor becomes leader for the first time; *Minimum experience* is the number of patents filed by an inventor before becoming leader. The full sample consists of all inventors appearing in the data between 1978 and 2013 that have been leaders at least once. The *Age >1* sub-sample excludes patents filed by inventors in their first year in the sample. The *Productivity > 0* sub-sample restricts the sample to inventors with positive productivity.

first is the minimum age for leadership, that is the age at which inventors become leaders of a patent for the first time. A first look indicates that, if they are ever leaders, inventors tend to play this role since the very beginning. However, this is mainly due to the fact that the majority of inventors appears in the data for only one year, as shown in panel (a) of Figure 1. In order to account for this, the second row of Table 10 excludes the patents produced in the first year an inventor appears in the data. Notice that the sample size reduces, as this constraint is essentially ruling out all inventors appearing in the data for only one year. The third row further restricts the analysis to patents where the leader had a positive productivity, in order to exclude those cases where an inventor is leader either because all his teammates have zero productivity as well or because he is the only one specialized in the same technological sector. The average minimum age for leadership increases to roughly 4 years in both sub-samples, meaning that, on average, inventors become leaders for the first time 4 years after they have filed their first patent. The second variable shown in Table 10 is the minimum experience for leadership, that is the number of patents filed by an inventor before becoming leader for the first time. Once again, the difference between the full sample and the two sub-samples highlights the bias introduced by the definition of leader itself for the first time an inventor appears in the data. By discarding the patents produced in the first year an inventor appears in the data, or the patents produced when the leader had zero productivity, the average minimum experience for leadership increases to roughly 2 patents. This means that, on average, an inventor has to play the role of supporter in filing 2 patents before becoming leader of a team for the first time. The distribution of the minimum requirements for leadership points towards the existence of a learning process such that inventors are supporters in a team before becoming leaders in the next team they join. In other words, there are some inventors that are not born as leaders; they have to be supporters first and do the leg work of becoming leaders of their own team.

At the other end of the spectrum, there are those inventors that are always leaders of their team. 54% of the inventors appearing between 1978 and 2013 are leaders in all the teams they have ever joined. The figure decreases to 34% for the sub-sample of inventors appearing for more than one year. These are the superstar inventors: those that are born as leaders.

Figure 5: Leadership status

(a) Leadership by age



Note: fraction of inventors that are leaders at least once at a given age. The sample reduces drastically as age increases, with less than 2,000 inventors per year for age larger than 25, less than 100 inventors per year for age larger than 33, and only 4 inventors for age 36.

(b) Frequency of leadership by age

Note: fraction of patents for which inventors are leaders, on average, by age. The sample reduces drastically as age increases, with less than 2,000 inventors per year for age larger than 25, less than 100 inventors per year for age larger than 33, and only 4 inventors for age 36.

Another aspect worth investigating is whether leadership status changes by age. Panel (a) of Figure 5 shows that the fraction of inventors that are leaders at least once at a given age increases with age. In fact, 60% of inventors of age 1 are leaders in at least one patent in that year, compared to roughly 80% of inventors aged 25. Similarly, almost 90% of 34 years old inventors are leaders in at least one of the patents they have produced in that year. However, it is important to bear in mind that the sample is reduced drastically as age increases, with only 0.06% of inventors surviving at least 25 years (dashed line in Figure 5). Since it is possible for inventors to produce multiple patents by year, I also consider the fraction of patents for which an inventor is leader at a given age. Panel (b) shows that this measure is increasing with age as well: on average, an inventor is leader of roughly 60% of the patents he produces in his first year and of 75% of the patents he produces when he is

25 years old. While the first measure points to the fact that over time it is more likely that an inventor plays a leadership role at least once in that year, the second measure tells that, over time, inventors tend to be leaders for more of their patents in a given year. As such, one could interpret the first measure as an extensive margin of leadership status and the second measure as an intensive margin of leadership status in each year.
The two measures behave in a very similar way, with the notable exception of the first years of age of inventors. Once again, this is due to the fact that some inventors appear in the dataset for the first time either alone or working with other *new* inventors, thus being all classified as leaders of their patent.

Lastly, I investigate whether inventors move to teams of different size over the years by estimating a specification similar to Equation 14, except for the addition of an inventor fixed effect δ_i:

$$\text{Age}_{ip} = \sum_{s=1}^{5} \beta_s \, \mathbf{I}\{\text{size}_p = s\} + \delta_{a(p)} + \delta_{t(p)} + \delta_i + \varepsilon_{ip} \tag{15}$$

where the error term ε_{ip} is three-way clustered at the applicant, technology field, and inventor level.

These results are shown in Table 11 and consider four different samples; the full sample, a sub-sample of only leaders, a sub-sample of leaders of teams with at least one supporter, and a sub-sample of supporters. The results suggest that, irrespectively of the sample considered, inventors tend to move to bigger teams as they age. The estimates are more sizeable for leaders: on average, a leader is slightly more than one year older when he works in a team of size 5 with respect to when he works alone. This corresponds to 22% of his average age. For inventors, irrespectively of their role, the same comparison would correspond to 11% of their average age. As far as supporters are concerned, a supporting inventor working in a team of at least five inventors is 0.4 years older than when he works in a team of size two, on average. The magnitude of this difference is 12% of the average age of supporters. This result is robust to defining leaders following Akcigit et al. (2018) (columns (4) to (6) of Table 21 in Appendix A).

The definition of age used in this paper captures experience. The empirical evidence shown in this section underlines the relevance of time and experience on the accumulation of productivity. Older, more experienced inventors are more likely to be leaders and join bigger teams, whether they play the role of leader or supporter.

Table 11: Age by team size within inventor

	(1) Inv. Age	(2) Inv. Age	(3) Inv. Age	(4) Inv. Age
$\mathbf{I}\{\text{size}_p = 2\}$	0.127***	0.334***		
	(0.015)	(0.016)		
$\mathbf{I}\{\text{size}_p = 3\}$	0.247***	0.597***	0.252***	0.149***
	(0.024)	(0.027)	(0.013)	(0.013)
$\mathbf{I}\{\text{size}_p = 4\}$	0.328***	0.804***	0.451***	0.254***
	(0.030)	(0.036)	(0.025)	(0.019)
$\mathbf{I}\{\text{size}_p = 5\}$	0.446***	1.069***	0.712***	0.428***
	(0.043)	(0.053)	(0.042)	(0.032)
Inventor FE	X	X	X	X
Applicant FE	X	X	X	X
Tech. field FE	X	X	X	X
Year FE	-	-	-	-
Inventors clusters	1,404,201	786,681	616,418	659,763
Applicants clusters	197,752	161,543	104,246	64,288
Tech. fields clusters	35	35	35	35
Sample	Full	Leaders	Leaders; $s > 1$	Supporters
N	6,340,193	3,180,869	2,397,513	2,611,452
Adj. R^2	0.637	0.681	0.713	0.656

Summary statistics:				
Dep. var. average	4.082	4.731	4.828	3.495
(std. dev.)	(4.380)	(4.911)	(4.949)	(3.663)

Note: The dependent variable is inventors' age. Coefficients for the interactions are with respect to $\mathbf{I}\{\text{size}_p = 1\}$ for the first two columns, with respect to $\mathbf{I}\{\text{size}_p = 2\}$ for the last two columns. Column (2) considers only leaders. Column (3) considers only leaders working with at least another team-mate. Column (4) considers only supporters. Standard errors in parentheses, robust to three-way clustering at the applicant, technology field, and inventor levels. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

6 Conclusions

This paper studies the collaboration process of inventors using a rich dataset on the universe of patents filed at the European Patent Office between 1978 and 2017. I find that bigger teams of inventors create patents of higher quality and more productive inventors work together in bigger teams, after controlling for filing year, applicant, and technological sector of the patent. Dividing inventors into leaders and supporters within each team unveils further stylized facts. Most importantly, the productivity of leaders exhibits a strong positive correlation with the productivity of supporting inventors that work with them, but this correlation decreases in magnitude as team size increases. Even after controlling for possible mechanical bias driven by the definition of leaders itself, I find that, in the most conservative case, a one standard deviation increase in leaders' productivity is associated with roughly doubling the average productivity of supporters working with them.

These results confirm the predictions of a simple one-factor version of the Roy model (Roy (1951)) where agents endogenously sort into being leaders or supporters according to their productivity. Despite being static, the assumptions used in this model still capture the intuition a dynamic model would provide. Findings on inventors' movement across teams over their life-cycle suggest that the dynamic component is relevant. I plan on encompassing this model in a more complete dynamic framework where inventors learn from their current teammates and internalize the network of their future teammates in the maximization problem.

References

Akcigit, U., Caicedo, S., Miguelez, E., Stantcheva, S., & Sterzi, V. (2018). *Dancing with the stars: Innovation through interactions* (tech. rep.). National Bureau of Economic Research.

Burstein, A. T., & Monge-Naranjo, A. (2009). Foreign know-how, firm control, and the income of developing countries. *The Quarterly Journal of Economics*, *124*(1), 149–195.

Doherr, T. (2016). Inventor mobility index: A method to disambiguate inventor careers. *ZEW-Centre for European Economic Research Discussion Paper*, (17-018).

Ductor, L., Fafchamps, M., Goyal, S., & Van der Leij, M. J. (2014). Social networks and research output. *Review of Economics and Statistics*, *96*(5), 936–948.

Hall, B. H., Jaffe, A., & Trajtenberg, M. (2005). Market value and patent citations. *RAND Journal of Economics*, 16–38.

Head, K., Li, Y. A., & Minondo, A. (2019). Geography, ties, and knowledge flows: Evidence from citations in mathematics. *Review of Economics and Statistics*, *101*(4), 713–727.

Herkenhoff, K., Lise, J., Menzio, G., & Phillips, G. M. (2018). *Production and learning in teams* (tech. rep.). National Bureau of Economic Research.

Jarosch, G., Oberfield, E., & Rossi-Hansberg, E. (2021). Learning from coworkers. *Econometrica*, *89*(2), 647–676.

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.

Jovanovic, B. (1979). Job matching and the theory of turnover. *Journal of Political Economy*, *87*(5, Part 1), 972–990.

Lee, Y.-N., Walsh, J. P., & Wang, J. (2015). Creativity in scientific teams: Unpacking novelty and impact. *Research policy*, *44*(3), 684–697.

Lucas, R. E. (1978). On the size distribution of business firms. *The Bell Journal of Economics*, 508–523.

Lucas, R. E. (2009). Ideas and growth. *Economica*, *76*(301), 1–19.

Luttmer, E. G. (2015). *An assignment model of knowledge diffusion and income inequality* (tech. rep.). Federal Reserve Bank, Research Department.

OECD. (July 2020). *HAN database.*

OECD. (Januray 2021). *Quality Index database.*

OECD. (January 2021). *REGPAT database.*

Perla, J., & Tonetti, C. (2014). Equilibrium imitation and growth. *Journal of Political Economy*, *122*(1), 52–76.

Raffo, J., & Lhuillery, S. (2009). How to play the “names game”: Patent retrieval comparing different heuristics. *Research policy*, *38*(10), 1617–1627.

Rosen, S. (1982). Authority, control, and the distribution of earnings. *The Bell Journal of Economics*, 311–323.

Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford economic papers*, *3*(2), 135–146.

Schmoch, U. (2008). Concept of a technology classification for country comparisons. *Final report to the World Intellectual Property Organisation, WIPO*.

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

Squicciarini, M., Dernis, H., & Criscuolo, C. (2013). Measuring patent quality: Indicators of technological and economic value.

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

Appendix A Robustness checks

Table 12: Technology indices by team size

	(1) Scope	(2) Generality	(3) Originality	(4) Radicalness
$\mathbf{I}\{\text{size}_p = 2\}$	0.265***	0.598***	0.455***	0.158
	(0.051)	(0.123)	(0.108)	(0.096)
$\mathbf{I}\{\text{size}_p = 3\}$	0.459***	1.167***	0.859***	0.250*
	(0.097)	(0.146)	(0.155)	(0.123)
$\mathbf{I}\{\text{size}_p = 4\}$	0.754***	1.952***	1.411***	0.280*
	(0.126)	(0.232)	(0.209)	(0.156)
$\mathbf{I}\{\text{size}_p = 5\}$	1.385***	3.164***	2.134***	-0.078
	(0.255)	(0.375)	(0.305)	(0.312)
Z_p	0.284***	0.202	0.362	-0.615***
	(0.099)	(0.234)	(0.235)	(0.192)
Applicants FE	X	X	X	X
Tech. field FE	X	X	X	X
Year FE	X	X	X	X
Applicants clusters	155,032	64,931	152,120	155,032
Tech. fields clusters	35	35	35	35
Year clusters	36	35	36	36
N	2,632,947	1,002,718	2,553,860	2,632,947
Adj. R^2	0.164	0.172	0.193	0.130

Summary statistics:				
Dep. var. average	21.053	41.008	69.510	29.732
(std. dev.)	(12.925)	(31.988)	(24.572)	(26.081)
Z_p average	0.146	0.186	0.146	0.146
(std. dev.)	(0.453)	(0.526)	(0.453)	(0.453)

Note: The dependent variable is a quality index. Quality indices have been rescaled to take value in $[0, 100]$. Coefficients with respect to $\mathbf{I}\{\text{size}_p = 1\}$. Z_p is the average productivity of inventors in the team. Standard errors in parentheses are robust to three-way clustering at the applicant, technology field, and year levels. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 13: Citation indices by team size – productivity as cumulative number of patents

	(1)	(2)	(3)	(4)	(5)	(6)
	Forward	Forward	Forward	Forward	Backward	NPL
	(5y)	(7y)	(5y) – XY	(7y) – XY		
$\mathbf{I}\{\text{size}_p = 2\}$	0.345***	0.382***	0.247***	0.274***	0.129***	0.185***
	(0.031)	(0.035)	(0.025)	(0.029)	(0.023)	(0.025)
$\mathbf{I}\{\text{size}_p = 3\}$	0.696***	0.777***	0.512***	0.598***	0.303***	0.289***
	(0.064)	(0.069)	(0.045)	(0.049)	(0.041)	(0.034)
$\mathbf{I}\{\text{size}_p = 4\}$	1.044***	1.172***	0.820***	0.964***	0.490***	0.360***
	(0.081)	(0.089)	(0.059)	(0.068)	(0.055)	(0.039)
$\mathbf{I}\{\text{size}_p = 5\}$	1.726***	1.913***	1.431***	1.633***	0.899***	0.527***
	(0.141)	(0.161)	(0.109)	(0.130)	(0.089)	(0.057)
Z_p	0.011***	0.014**	0.007***	0.009**	0.005	0.013***
	(0.003)	(0.003)	(0.003)	(0.004)	(0.006)	(0.004)
Applicants FE	X	X	X	X	X	X
Tech. field FE	X	X	X	X	X	X
Year FE	X	X	X	X	X	X
Applicants clusters	149,121	137,108	149,121	137,108	179,712	179,712
Tech. fields clusters	35	35	35	35	35	35
Year clusters	35	33	35	33	40	40
N	2,503,053	2,248,873	2,503,053	2,248,873	3,162,683	3,162,683
Adj. R^2	0.099	0.114	0.067	0.074	0.334	0.144

Summary statistics:						
Dep. var. average	3.400	3.945	2.524	3.069	9.341	2.109
(std. dev.)	(7.657)	(8.088)	(7.836)	(8.439)	(9.681)	(6.420)
Z_p average	2.941	2.748	2.941	2.748	3.478	3.478
(std. dev.)	(7.345)	(6.769)	(7.345)	(6.769)	(9.069)	(9.069)

Note: The dependent variable is a quality index. Quality indices have been rescaled to take value in $[0, 100]$. Coefficients with respect to $\mathbf{I}\{\text{size}_p = 1\}$. Z_p is the average productivity of inventors in the team, with productivity measured as cumulative count of past patents. Standard errors in parentheses are robust to three-way clustering at the applicant, technology field, and year levels. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 14: Technology indices by team size – productivity as cumulative number of patents

	(1) Scope	(2) Generality	(3) Originality	(4) Radicalness
$\mathbf{I}\{\text{size}_p = 2\}$	0.238***	0.580***	0.421***	0.157
	(0.046)	(0.121)	(0.104)	(0.096)
$\mathbf{I}\{\text{size}_p = 3\}$	0.418***	1.142***	0.775***	0.222**
	(0.082)	(0.146)	(0.149)	(0.107)
$\mathbf{I}\{\text{size}_p = 4\}$	0.666***	1.926***	1.251***	0.267*
	(0.113)	(0.233)	(0.198)	(0.148)
$\mathbf{I}\{\text{size}_p = 5\}$	1.252***	3.133***	1.974***	0.001
	(0.233)	(0.376)	(0.281)	(0.291)
Z_p	0.006	-0.031*	-0.006	-0.051***
	(0.007)	(0.016)	(0.010)	(0.010)
Applicants FE	X	X	X	X
Tech. field FE	X	X	X	X
Year FE	X	X	X	X
Applicants clusters	179,712	64,931	176,535	179,712
Tech. fields clusters	35	35	35	35
Year clusters	40	35	40	40
N	3,162,683	1,002,718	3,074,440	3,162,683
Adj. R^2	0.160	0.172	0.193	0.133

Summary statistics:				
Dep. var. average	20.965	41.008	70.054	30.528
(std. dev.)	(12.781)	(31.988)	(24.254)	(26.247)
Z_p average	3.478	3.117	3.471	3.479
(std. dev.)	(9.069)	(7.451)	(8.947)	(9.069)

Note: The dependent variable is a quality index. Quality indices have been rescaled to take value in $[0, 100]$. Coefficients with respect to $\mathbf{I}\{\text{size}_p = 1\}$. Z_p is the average productivity of inventors in the team, with productivity measured as cumulative count of past patents. Standard errors in parentheses are robust to three-way clustering at the applicant, technology field, and year levels. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 15: Citation indices by team size – alternative leader definition

	(1) Forward (5y)	(2) Forward (7y)	(3) Forward (5y) – XY	(4) Forward (7y) – XY	(5) Backward	(6) NPL
$\mathbf{I}\{\text{size}_p = 3\}$	0.337***	0.381***	0.240***	0.303***	0.179***	0.084***
	(0.043)	(0.048)	(0.029)	(0.034)	(0.026)	(0.022)
$\mathbf{I}\{\text{size}_p = 4\}$	0.640***	0.729***	0.529***	0.645***	0.374***	0.150***
	(0.061)	(0.066)	(0.041)	(0.049)	(0.049)	(0.037)
$\mathbf{I}\{\text{size}_p = 5\}$	1.283***	1.421***	1.064***	1.225***	0.751***	0.317***
	(0.108)	(0.126)	(0.082)	(0.099)	(0.076)	(0.064)
Z_p^l	0.436***	0.465***	0.362***	0.402***	0.220***	0.098***
	(0.059)	(0.058)	(0.054)	(0.066)	(0.037)	(0.028)
Z_p^s	0.351**	0.390***	0.269**	0.301**	0.156	0.195**
	(0.136)	(0.126)	(0.122)	(0.122)	(0.180)	(0.090)
Applicants FE	X	X	X	X	X	X
Tech. field FE	X	X	X	X	X	X
Year FE	X	X	X	X	X	X
Applicants clusters	50,612	45,634	50,612	45,634	53,039	53,039
Tech. fields clusters	35	35	35	35	35	35
Year clusters	34	32	34	32	35	35
N	1,101,037	966,682	1,101,037	966,682	1,170,804	1,170,804
Adj. R^2	0.108	0.122	0.079	0.089	0.299	0.151

Summary statistics:						
Dep. var. average	3.581	4.120	2.918	3.500	8.791	2.267
(std. dev.)	(7.934)	(8.372)	(8.233)	(8.798)	(9.342)	(6.539)
Z_p^l average	0.517	0.511	0.517	0.511	0.519	0.519
(std. dev.)	(1.060)	(1.043)	(1.060)	(1.043)	(1.065)	(1.065)
Z_p^s average	0.084	0.081	0.084	0.081	0.085	0.085
(std. dev.)	(0.299)	(0.291)	(0.299)	(0.291)	(0.300)	(0.300)

Note: The dependent variable is a quality index. Quality indices have been rescaled to take value in $[0, 100]$. Coefficients with respect to $\mathbf{I}\{\text{size}_p = 2\}$. Z_p^l is the productivity of the leader of patent p, while Z_p^s is the average productivity of supporting inventors working in patent p. Leaders defined as most productive inventors within the team, regardless of specialization, according to Akcigit et al. (2018). Standard errors in parentheses are robust to three-way clustering at the applicant, technology field, and year levels. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 16: Technology indices by team size

	(1) Scope	(2) Generality	(3) Originality	(4) Radicalness
$\mathbf{I}\{\text{size}_p = 3\}$	0.189**	0.534***	0.318***	0.036
	(0.076)	(0.066)	(0.087)	(0.099)
$\mathbf{I}\{\text{size}_p = 4\}$	0.446***	1.308***	0.874***	0.096
	(0.109)	(0.202)	(0.133)	(0.144)
$\mathbf{I}\{\text{size}_p = 5\}$	1.067***	2.484***	1.558***	-0.202
	(0.244)	(0.350)	(0.241)	(0.310)
Z_p^l	0.004	-0.256	0.040	-0.345***
	(0.057)	(0.183)	(0.134)	(0.099)
Z_p^s	0.517***	1.217***	0.553**	0.042
	(0.116)	(0.211)	(0.225)	(0.147)
Applicants FE	X	X	X	X
Tech. field FE	X	X	X	X
Year FE	X	X	X	X
Applicants clusters	52,936	27,397	52,068	52,936
Tech. fields clusters	35	35	35	35
Year clusters	35	34	35	35
N	1,168,140	482,045	1,135,052	1,168,140
Adj. R^2	0.172	0.182	0.197	0.141

Summary statistics:				
Dep. var. average	21.563	42.715	71.838	30.377
(std. dev.)	(13.090)	(31.718)	(23.489)	(26.138)
Z_p^l average	0.475	0.579	0.474	0.475
(std. dev.)	(1.020)	(1.152)	(1.016)	(1.020)
Z_p^s average	0.104	0.130	0.104	0.104
(std. dev.)	(0.350)	(0.404)	(0.349)	(0.350)

Note: The dependent variable is a quality index. Quality indices have been rescaled to take value in $[0, 100]$. Coefficients with respect to $\mathbf{I}\{\text{size}_p = 2\}$. Z_p^l is the productivity of the leader of patent p, while Z_p^s is the average productivity of supporting inventors working in patent p. Standard errors in parentheses are robust to three-way clustering at the applicant, technology field, and year levels. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 17: Technology indices by team size – alternative leader definition

	(1) Scope	(2) Generality	(3) Originality	(4) Radicalness
$\mathbf{I}\{\text{size}_p = 3\}$	0.195**	0.545***	0.339***	0.059
	(0.078)	(0.066)	(0.088)	(0.099)
$\mathbf{I}\{\text{size}_p = 4\}$	0.450***	1.314***	0.889***	0.118
	(0.109)	(0.197)	(0.132)	(0.145)
$\mathbf{I}\{\text{size}_p = 5\}$	1.063***	2.452***	1.548***	-0.197
	(0.245)	(0.348)	(0.241)	(0.313)
Z_p^l	0.118**	0.130	0.323***	-0.034
	(0.045)	(0.150)	(0.104)	(0.124)
Z_p^s	0.201	0.092	-0.394	-1.017***
	(0.184)	(0.322)	(0.324)	(0.264)
Applicants FE	X	X	X	X
Tech. field FE	X	X	X	X
Year FE	X	X	X	X
Applicants clusters	53,039	27,396	52,169	53,039
Tech. fields clusters	35	35	35	35
Year clusters	35	34	35	35
N	1,170,804	482,280	1,137,653	1,170,804
Adj. R^2	0.172	0.182	0.197	0.140

Summary statistics:				
Dep. var. average	21.543	42.681	71.803	30.352
(std. dev.)	(13.080)	(31.723)	(23.508)	(26.133)
Z_p^l average	0.519	0.633	0.519	0.519
(std. dev.)	(1.065)	(1.201)	(1.061)	(1.065)
Z_p^s average	0.085	0.109	0.085	0.085
(std. dev.)	(0.300)	(0.352)	(0.297)	(0.300)

Note: The dependent variable is a quality index. Quality indices have been rescaled to take value in $[0, 100]$. Coefficients with respect to $\mathbf{I}\{\text{size}_p = 2\}$. Z_p^l is the productivity of the leader of patent p, while Z_p^s is the average productivity of supporting inventors working in patent p. Leaders defined as most productive inventors within the team, regardless of specialization, according to Akcigit et al. (2018). Standard errors in parentheses are robust to three-way clustering at the applicant, technology field, and year levels. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 18: Inventors' average productivity by team size – productivity as cumulative number of patents

	(1) Prod. Pat.	(2) Prod. Pat.	(3) Prod. Pat.	(4) Prod. Pat.
$\mathbf{I}\{\text{size}_p = 2\}$	0.018	0.182**		
	(0.078)	(0.083)		
$\mathbf{I}\{\text{size}_p = 3\}$	-0.010	0.369***	0.156***	-0.014
	(0.098)	(0.128)	(0.050)	(0.029)
$\mathbf{I}\{\text{size}_p = 4\}$	-0.021	0.577***	0.342***	0.003
	(0.113)	(0.164)	(0.090)	(0.053)
$\mathbf{I}\{\text{size}_p = 5\}$	0.026	1.063***	0.801***	0.145
	(0.129)	(0.227)	(0.191)	(0.097)
Applicant FE	X	X	X	X
Tech. field FE	X	X	X	X
Age FE	X	X	X	X
Cohort FE	X	X	X	X
Inventors clusters	3,347,318	2,175,539	1,844,545	1,590,919
Applicants clusters	283,256	242,087	189,353	94,451
Tech. fields clusters	35	35	35	35
Sample	Full	Leaders	Leaders; $s > 1$	Supporters
N	7,320,271	4,116,089	3,260,672	3,136,382
Adj. R^2	0.332	0.390	0.391	0.310

Summary statistics:				
Dep. var. average	3.025	3.427	3.548	2.490
(std. dev.)	(9.623)	(10.731)	(11.218)	(7.946)

Note: The dependent variable is the productivity of inventors working on a patent, where productivity is measured as cumulative count of past patents. Coefficients with respect to $\mathbf{I}\{\text{size}_p = 1\}$ for the first two columns; with respect to $\mathbf{I}\{\text{size}_p = 2\}$ for the last two columns. Column (2) considers only leaders. Column (3) considers only leaders in teams of at least 2 inventors. Column (4) restricts the sample to supporters only. Standard errors in parentheses are robust to three-way clustering at the applicant, technology field, and inventor levels. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 19: Inventors' average productivity by team size – alternative leader definition

	(1) Prod. Cit.	(2) Prod. Cit.	(3) Prod. Cit.	(4) Prod. Pat.	(5) Prod. Pat.	(6) Prod. Pat.
$\mathbf{I}\{\text{size}_p = 2\}$	0.020***			0.210**		
	(0.005)			(0.084)		
$\mathbf{I}\{\text{size}_p = 3\}$	0.039***	0.017***	0.003***	0.435***	0.189***	0.005
	(0.008)	(0.003)	(0.001)	(0.127)	(0.050)	(0.028)
$\mathbf{I}\{\text{size}_p = 4\}$	0.062***	0.039***	0.008***	0.706***	0.434***	0.028
	(0.010)	(0.006)	(0.002)	(0.158)	(0.087)	(0.056)
$\mathbf{I}\{\text{size}_p = 5\}$	0.103***	0.077***	0.015***	1.379***	1.074***	0.190**
	(0.014)	(0.011)	(0.004)	(0.262)	(0.237)	(0.092)
Applicant FE	X	X	X	X	X	X
Tech. field FE	X	X	X	X	X	X
Age FE	X	X	X	X	X	X
Cohort FE	X	X	X	X	X	X
Inventors clusters	2,163,473	1,829,054	1,585,066	2,158,167	1,826,124	1,583,007
Applicants clusters	241,780	188,983	93,993	242,745	190,094	94,010
Tech. fields clusters	35	35	35	35	35	35
Sample	Leaders	Leaders; $s > 1$	Supporters	Leaders	Leaders; $s > 1$	Supporters
N	4,110,187	3,254,707	3,142,006	4,135,476	3,280,142	3,117,675
Adj. R^2	0.289	0.295	0.190	0.406	0.411	0.308

Summary statistics:						
Dep. var. average	0.183	0.193	0.091	3.661	3.841	2.169
(std. dev.)	(0.659)	(0.688)	(0.376)	(11.240)	(11.823)	(6.850)

Note: The dependent variable is inventors' productivity, measured as cumulative sum of past citations in columns (1)-(3), and as cumulative count of past patents in columns (4)-(6). Coefficients with respect to $\mathbf{I}\{\text{size}_p = 1\}$ for the columns (1) and (4); with respect to $\mathbf{I}\{\text{size}_p = 2\}$ for all the remaining columns. Column (1) and (4) consider only leaders. Column (2) and (5) consider only leaders in teams of at least two inventors. Column (3) and (6) consider only supporters. Leaders are defined as the most productive inventors within the team, regardless of specialization, according to Akcigit et al. (2018). Standard errors in parentheses are robust to three-way clustering at the applicant, technology field, and inventor levels. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 20: Supporters' productivity and leader's productivity – alternative definitions

	(1) Sup. Prod. Patents	(2) Sup. Prod. Patents	(3) Sup. Prod. Citations	(4) Sup. Prod. Citations	(5) Sup. Prod. Patents	(6) Sup. Prod. Patents
Z_p^l	0.169***	0.184***	0.172***	0.191***	0.186***	0.206***
	(0.018)	(0.020)	(0.010)	(0.013)	(0.015)	(0.017)
$\mathrm{Z}_p^l \times \mathbf{I}\{\mathrm{size}_p = 3\}$		-0.005		-0.014**		-0.009
		(0.010)		(0.006)		(0.009)
$\mathrm{Z}_p^l \times \mathbf{I}\{\mathrm{size}_p = 4\}$		-0.016		-0.019***		-0.021
		(0.017)		(0.005)		(0.017)
$\mathrm{Z}_p^l \times \mathbf{I}\{\mathrm{size}_p = 5\}$		-0.032*		-0.036***		-0.039**
		(0.019)		(0.009)		(0.017)
Applicant FE	X	X	X	X	X	X
Tech. field FE	X	X	X	X	X	X
Year FE	X	X	X	X	X	X
Applicants clusters	53,521	53,521	53,039	53,039	53,307	53,307
Tech. fields clusters	35	35	35	35	35	35
Filing clusters	35	35	35	35	35	35
Alternative leader definition	-	-	X	X	X	X
N	1,173,745	1,173,745	1,170,804	1,170,804	1,168,119	1,168,119
Adj. R^2	0.353	0.355	0.417	0.419	0.493	0.496

Interpretation of the effect:						
Coeff. as % of avg. dep. var.	7%	8%	202%	225%	9%	10%
1 sd as % of avg. dep. var.	125%	136%	216%	239%	170%	188%

Note: The dependent variable is average supporters productivity within team, where productivity is measured by the cumulative count of patents in columns (1), (2), (5), and (6), and as the cumulative sum of citations in columns (3) and (4). Coefficients for the interactions are estimated with respect to $\mathbf{I}\{\mathrm{size}_p = 2\}$. Z_p^l is leaders' productivity. In the columns (3)-(6), leaders are identified as in Akcigit et al. (2018): the inventors with the highest productivity within the team, without taking into account their specialization. Standard errors in parentheses, robust to three-way clustering at the applicant, technology field, and filing year levels. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 21: Inventors age – alternative leader definition

	(1) Inv. Age	(2) Inv. Age	(3) Inv. Age	(4) Inv. Age	(5) Inv. Age	(6) Inv. Age
$\mathbf{I}\{\text{size}_p = 2\}$	0.156***			0.365***		
	(0.023)			(0.017)		
$\mathbf{I}\{\text{size}_p = 3\}$	0.348***	0.190***	0.030**	0.649***	0.273***	0.158***
	(0.036)	(0.018)	(0.013)	(0.028)	(0.014)	(0.012)
$\mathbf{I}\{\text{size}_p = 4\}$	0.525***	0.362***	0.062***	0.873***	0.486***	0.271***
	(0.047)	(0.033)	(0.017)	(0.038)	(0.026)	(0.018)
$\mathbf{I}\{\text{size}_p = 5\}$	0.779***	0.614***	0.098***	1.160***	0.769***	0.463***
	(0.072)	(0.062)	(0.021)	(0.058)	(0.046)	(0.030)
Inventor FE	-	-	-	X	X	X
Applicant FE	X	X	X	X	X	X
Tech. field FE	X	X	X	X	X	X
Year FE	X	X	X	-	-	-
Inventors clusters	2,547,602	2,170,532	1,940,411	770,779	600,667	670,611
Applicants clusters	283,425	224,012	113,490	161,541	104,322	65,538
Tech. fields clusters	35	35	35	35	35	35
Sample	Leaders	Leaders; $s > 1$	Supporters	Leaders	Leaders; $s > 1$	Supporters
N	4,968,850	3,976,382	3,926,407	3,171,116	2,390,658	2,642,756
Adj. R^2	0.278	0.280	0.106	0.690	0.723	0.645

Summary statistics:						
Dep. var. average	3.557	3.539	2.969	4.839	4.968	3.502
(std. dev.)	(4.429)	(4.424)	(3.128)	(4.996)	(5.056)	(3.669)

Note: The dependent variable is the age of inventors. Coefficients are with respect to $\mathbf{I}\{\text{size}_p = 1\}$ for columns (1) and (4), with respect to $\mathbf{I}\{\text{size}_p = 2\}$ for the remaining columns. Columns (1) and (4) only consider leaders. Columns (2) and (5) only consider leaders working in teams of at least two inventors. Columns (3) and (6) consider only supporters. Leaders are defined as most productive inventors within the team, regardless of specialization, according to Akcigit et al. (2018). Standard errors in parentheses are robust to three-way clustering at the applicant, technology field, and inventor levels. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Appendix B Data

This section describes in detail the data cleaning process and the validation of the resulting dataset with alternative data sources. If also provides further summary statistics.

B.1 Data cleaning and validation

The raw data taken from the OECD assign an inventor ID to any unique name-address pair. However, the same name and address are usually written in different ways (commas, spaces, abbreviations, special and accentuated characters, etc.) so that a disambiguation strategy is necessary.

I use the algorithm recommended by Raffo and Lhuillery (2009), which consists of three steps. The first one is the cleaning step: all names are converted into lower case and cleaned from commas, points, titles (*Prof.*, *Dr.*, etc.) and references to corporations (eg, *Ing. at Toyota*). The second is the matching step, where name strings are matched based on tokens, keeping as a successful match all pairs with a Jaccard score equal to 0.9. The third and last step is the filtering step: the matches in the previous step are filtered based on the comparison of addresses at the bigram level and within the same region (NUTS3 for EU and OECD's Territorial Level 3 for the other countries). For addresses, matches are considered successful if the pair has a Jaccard score greater or equal than 0.7. While the threshold for the name match is close to the maximum value of 1, the threshold for the address match is smaller, as addresses are more subject to different abbreviations (simply think about *St.* and *Street*). Moreover, while there is not a clear consensus of what the ideal threshold level should be, the threshold for the filtering step is smaller than the threshold used for the matching step as recommended by Doherr (2016).

Table 22 below reports an illustrative example which also highlights possible flaws in the process: the inventor with raw ID 5 is a false negative, which is not captured because of a typo in the name.

Table 22: Name disambiguation - example

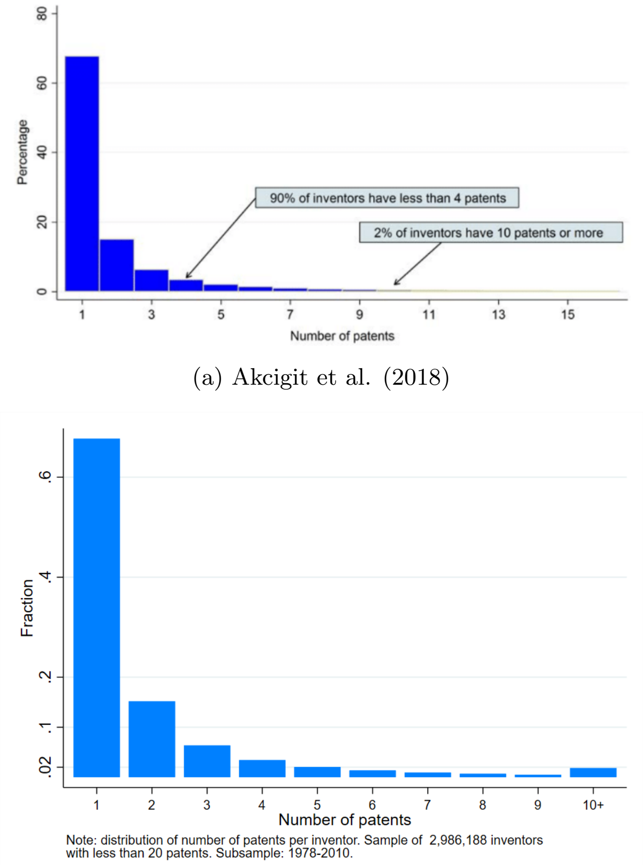
Raw ID	Name	Address	Disambiguated ID
1	Smith, John	123 Main Street, London	1
2	SMITH, John	London, 123 Main Street	1
3	Dr. John Smith	123 Main Street, London	1
4	Smith, John	123 Main St, WC1N 2AT, London	1
5	Smih, John	123 Main Street, London	5
6	Smith, John	456 Old St, London	6

The raw data provided by the OECD spans from 1977 to 2020 and contains 6,777,154 unique inventor IDs. The disambiguation process described above has reduced the number to 4,350,066 inventors for the same time period. Since this disambiguation process is crucial when assessing the number of patents an inventor has produced in a lifetime, I validate my data by comparing them with the statistics reported in Akcigit et al. (2018), as they use similar data sources.[10] Figure 6 shows the distribution of number of patents by inventor as reported in Akcigit et al. (2018) on panel (a) and as obtained with my data after disambiguation on panel (b). Since

[10] Note that Akcigit et al. (2018) use the CRIOS-PatStat dataset, a version of the PATSTAT data with inventors names already disambiguated.

Akcigit et al. (2018) use data up to 2010 and only inventors with less than 20 patents in the figure, I use a comparable subsample. The two distributions seem very similar, with more than 60% of inventors having only 1 patent, 90% having less than 4 patents, and 2% having at least 10 patents.

Figure 6: Patents per inventor – data validation



90% of inventors have less than 4 patents

2% of inventors have 10 patents or more

(a) Akcigit et al. (2018)

Note: distribution of number of patents per inventor. Sample of 2,986,188 inventors with less than 20 patents. Subsample: 1978-2010.

(b) Disambiguated data

In the empirical analysis I use a baseline dataset given by all patents filed between 1978 and 2017 for which both applicants and inventors are known. I further exclude 6 inventors with more than 500 patents produced in their lifetime and 954 inventors with extreme levels of productivity (ranging between 10 and 54). The final dataset used in the paper comprises 3,490,866 patents filed by 4,147,765 inventors and 508,087 applicants.

B.2 Other summary statistics

Figure 7: Patents by filing year

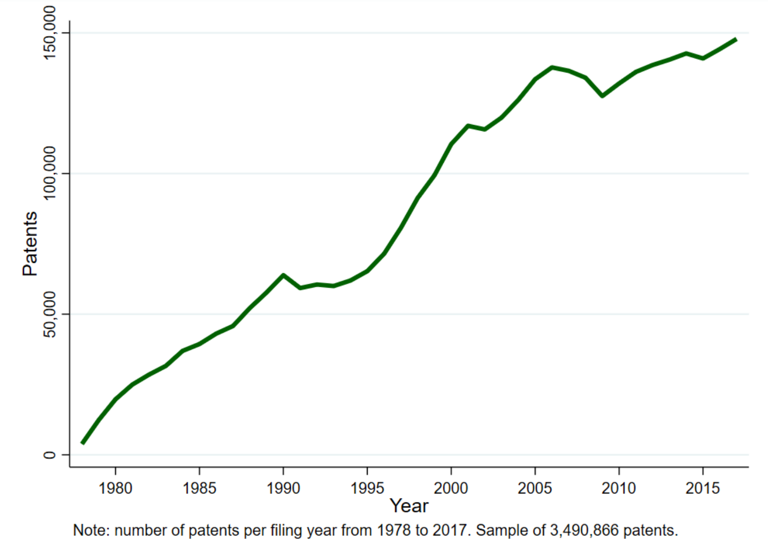


Figure 8: Team size by filing year

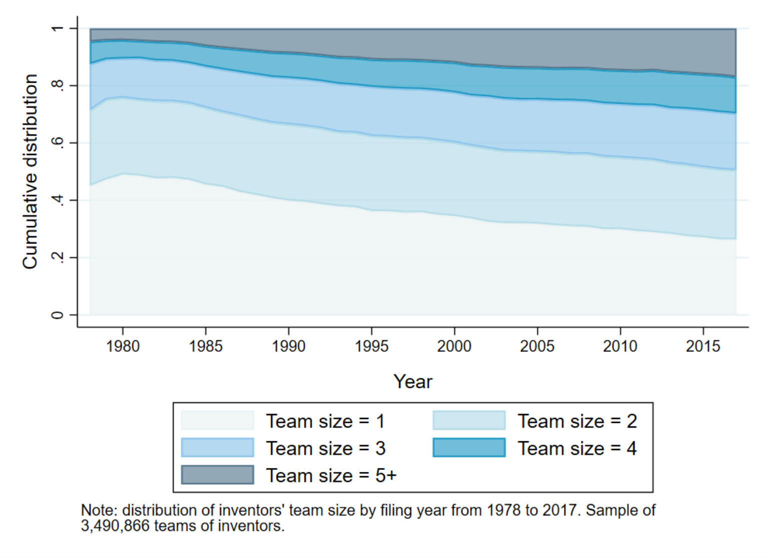
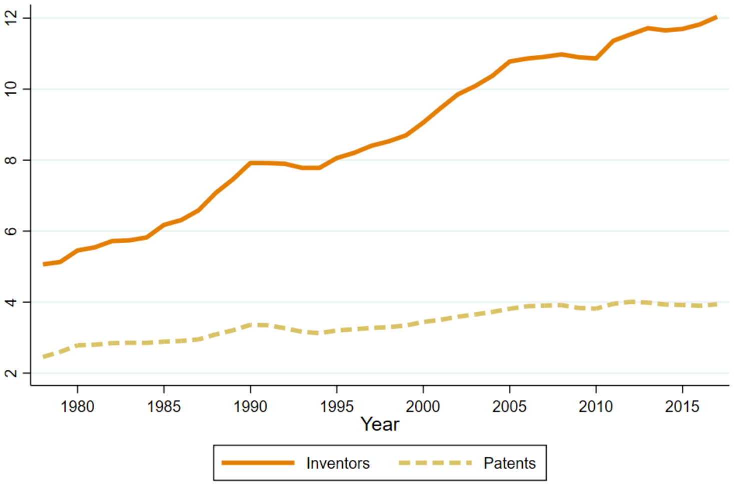


Table 23: Inventors' productivity – alternative definition

	N	Mean	Min	Max	sd
Full sample	5,211,737	1.545	0	365	4.923
Leaders	3,244,775	1.812	0	365	5.658
Supporters	2,226,384	1.557	0	354	4.959

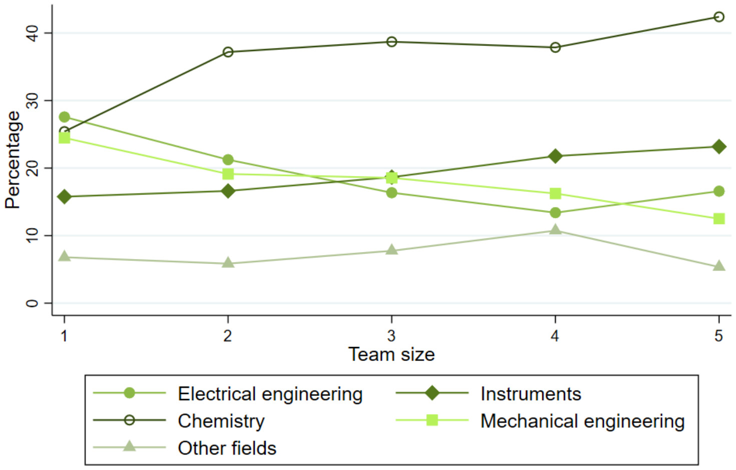
Note: productivity distribution for 5,211,737 inventor × year observations in the time window 1979-2013. Productivity defined as cumulative count of past patents.

Figure 9: Applicants' dimension by filing year



Note: average dimension of applicants per year from 1978 to 2017. Sample of 1,053,497 applicant-year observations. Dimension is either number of inventors or number of patents.

Figure 10: Technology sector by applicants' team size



Note: technological category as percentage of patents developed by applicants' team size. 5 stands for teams with at least five applicants. Sample of 3,489,209 patents; 1,657 patents with unknown technological category are excluded.

Table 24: Number of patents by technology fields

Technology field	# of patents	% in data
Electrical engineering		
1. Electrical machinery, apparatus, energy	216,047	6.19
2. Audio-visual technology	128,475	3.68
3. Telecommunications	100,802	2.89
4. Digital communication	170,328	4.88
5. Basic communication processes	35,606	1.02
6. Computer technology	187,579	5.37
7. IT methods for management	27,501	0.79
8. Semiconductors	79,939	2.29
Instruments		
9. Optics	103,564	2.97
10. Measurement	162,010	4.64
11. Analysis of biological materials	26,326	0.75
12. Control	51,625	1.48
13. Medical technology	209,509	6.00
Chemistry		
14. Organic fine chemistry	139,655	4.00
15. Biotechnology	112,015	3.21
16. Pharmaceuticals	168,093	4.81
17. Macromolecular chemistry, polymers	102,477	2.94
18. Food chemistry	31,981	0.92
19. Basic materials chemistry	97,433	2.79
20. Materials, metallurgy	70,029	2.01
21. Surface technology, coating	55,941	1.60
22. Micro-structural and nano-technology	2,575	0.07
23. Chemical engineering	89,553	2.56
24. Environmental technology	42,882	1.23
Mechanical engineering		
25. Handling	109,715	3.14
26. Machine tools	91,163	2.61
27. Engines, pumps, turbines	113,044	3.24
28. Textile and paper machines	86,950	2.49
29. Other special machines	112,486	3.23
30. Thermal processes and apparatus	55,621	1.59
31. Mechanical elements	113,138	3.24
32. Transport	159,521	4.57
Other fields		
33. Furniture, games	64,327	1.84
34. Other consumer goods	65,281	1.87
35. Civil engineering	106,018	3.04

Note: it excludes 1,657 patents with unknown technology field.

Appendix C Theoretical framework

This section provides more details on the solution of the model and the proofs to the lemmas of Section 2.

C.1 Model solution

Leaders solve

$$\max_{n\geq 0} U_i^l(n; Z_i^l) = n^\eta \big(Z_i^l\big)^{1-\eta} - mn$$

$$\begin{aligned}
\text{FOC[n]:} \qquad & \eta n^{\eta-1}\big(Z_i^l\big)^{1-\eta} - m = 0 \\
& n^*(Z_i^l) = \left[\frac{\eta}{m}\right]^{\frac{1}{1-\eta}} Z_i^l
\end{aligned}$$

Supporters solve

$$\max_{Z^l>\overline{Z}} U_i^s(Z^l; Z_i^s) = \big(Z_i^s\big)^\alpha \big(Z^l\big)^{1-\alpha} - \frac{n^*(Z^l)}{Z_i^s}$$

$$\begin{aligned}
\text{FOC}[Z^l]\text{:} \qquad & (1-\alpha)\big(Z_i^s\big)^\alpha \big(Z^l\big)^{-\alpha} - \left[\frac{\eta}{m}\right]^{\frac{1}{1-\eta}} \frac{1}{Z_i^s} = 0 \\
& \big(Z^l\big)^{-\alpha} = \left[\frac{\eta}{m}\right]^{\frac{1}{1-\eta}} \frac{1}{1-\alpha} \big(Z_i^s\big)^{-(1+\alpha)} \\
& Z^l = (1-\alpha)^{\frac{1}{\alpha}} \left[\frac{m}{\eta}\right]^{\frac{1}{\alpha(1-\eta)}} \big(Z_i^s\big)^{\frac{1+\alpha}{\alpha}}
\end{aligned}$$

Agents choose whether to become a leader or a supporter by maximizing the optimal utility that they would get in each role:

$$\begin{aligned}
& \max\{U_i^l(n^*; Z_i), U_i^s(Z^{l*}; Z_i)\} \qquad \text{where} \\
& U_i^l(n^*; Z_i) = (n^*)^\eta \big(Z_i\big)^{1-\eta} - mn^* = \left[\frac{\eta}{m}\right]^{\frac{1}{1-\eta}} \left[\left(\frac{\eta}{m}\right)^\eta - m\right] Z_i \\
& U_i^s(Z^{l*}; Z_i) = \big(Z_i\big)^\alpha \big(Z^{l*}\big)^{1-\alpha} - \frac{n^*(Z^{l*})}{Z_i} = (1-\alpha)^{\frac{1}{\alpha}} \left[\frac{m}{\eta}\right]^{\frac{1}{\alpha(1-\eta)}} \left[(1-\alpha)^{1-\alpha}\left[\frac{m}{\eta}\right]^{1-\alpha} - 1\right] Z_i^{\frac{1}{\alpha}}
\end{aligned}$$

An agent i chooses to become leader if and only if his skill Z_i is such that $U_i^l(n^*; Z_i) > U_i^s(Z^{l*}; Z_i)$, that is:

$$\begin{aligned}
& \left[\frac{\eta}{m}\right]^{\frac{1}{1-\eta}} \left[\left(\frac{\eta}{m}\right)^\eta - m\right] Z_i > (1-\alpha)^{\frac{1}{\alpha}} \left[\frac{m}{\eta}\right]^{\frac{1}{\alpha(1-\eta)}} \left[(1-\alpha)^{1-\alpha}\left[\frac{m}{\eta}\right]^{1-\alpha} - 1\right] Z_i^{\frac{1}{\alpha}} \\
& Z_i^{\frac{\alpha-1}{\alpha}} > \left[\frac{m}{\eta}\right]^{\frac{1-\alpha}{\alpha(1-\eta)}} (1-\alpha)^{\frac{1}{\alpha}} \left[(1-\alpha)^{1-\alpha}\left[\frac{m}{\eta}\right]^{1-\alpha} - 1\right] \left[\left(\frac{\eta}{m}\right)^\eta - m\right]^{-1}
\end{aligned}$$

Agents become leaders if and only if their productivity Z_i is lower than a threshold $\underline{Z}$ or bigger than a threshold $\overline{Z}$:

$$Z_i \in \big[0, \underline{Z}\big] \qquad \text{or} \qquad Z_i > \overline{Z}$$

where $\overline{Z} \equiv \frac{\zeta^l}{\zeta^s}\zeta^{l,s}$ and $\underline{Z} \equiv \overline{Z}^{\frac{\alpha}{1+\alpha}}\left[(1-\alpha)\zeta^l\right]^{-\frac{1}{1+\alpha}}$ with

$$\zeta^l \equiv \left[\frac{m}{\eta}\right]^{\frac{1}{1-\eta}}$$
$$\zeta^s \equiv (1-\alpha)^{\frac{1}{1-\alpha}}$$
$$\zeta^{l,s} \equiv \frac{\left[\left(\frac{\eta}{m}\right)^{\eta} - m\right]^{\frac{\alpha}{1-\alpha}}}{\left[\left[(1-\alpha)\frac{m}{\eta}\right]^{1-\alpha} - 1\right]^{\frac{\alpha}{1-\alpha}}}$$

Note that $\underline{Z}$ is the inverse of the supporters FOC evaluated in $\overline{Z}$: it is the productivity of supporters that match with the least productive leader with a non-zero number of supporters.

The optimal choice of team size of leaders is

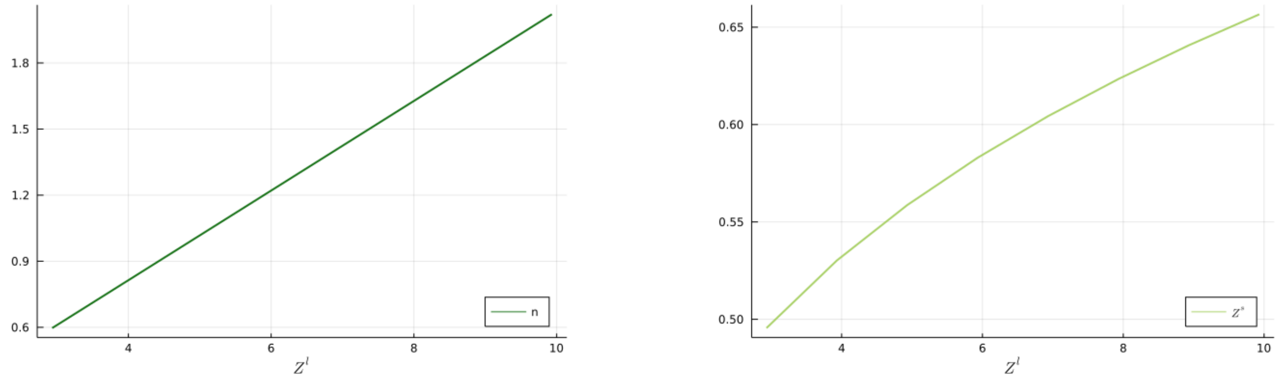
$$n^*(Z_i) = \begin{cases} 0 & \text{if } Z_i \in \left[0, \underline{Z}\right] \\ \left[\frac{\eta}{m}\right]^{\frac{1}{1-\eta}} Z_i & \text{if } Z_i > \overline{Z} \end{cases}$$

Figure 11 provides a visual representation of the first order conditions for leaders managing at least one supporter in the case where $\alpha = 0.3$, $\eta = 0.8$, and $m = 1.1$. The left panel shows the optimal team size for every leader with productivity $Z_i > \overline{Z} = 2.93$. The right panel instead shows the productivity of supporters associated with each leader with productivity $Z_i > \overline{Z} = 2.93$.

Figure 11: First Order Conditions

(a) Team size

(b) Supporters productivity



C.2 Proofs

- Equation 9: $n^*(Z_i) = \begin{cases} 0 & \text{if } Z_i \in \left[0, \underline{Z}\right] \\ \left[\frac{\eta}{m}\right]^{\frac{1}{1-\eta}} Z_i & \text{if } Z_i > \overline{Z} \end{cases}$

  Proof. I prove the first case, as the second case comes directly from the FOC of leaders.

  Consider an inventor with productivity $Z_i \in (0, \underline{Z}]$. According to the FOC, if he becomes a supporter, he chooses

$$Z^{l*} = (1-\alpha)^{\frac{1}{\alpha}}\left[\frac{\eta}{m}\right]^{\frac{1}{\alpha(1-\eta)}} Z_i^{\frac{1+\alpha}{\alpha}} < \overline{Z}$$

Since his optimal Z^{l*} is unfeasible, his best match would be the corner solution $\overline{Z}$.

If he becomes a leader, the FOC implies that he chooses

$$n^* = \left[\frac{\eta}{m}\right]^{\frac{1}{1-\eta}} Z_i > 0$$

However, since each supporter matches with a leader with productivity higher than his own, $Z^l(Z_i) > Z_i$, this leader would not find any supporter willing to match with him. Therefore, his best option would be to choose $n^* = 0$.

Individuals with productivity $Z_i \in (0, \underline{Z}]$ choose whether to become leaders or supporters based on the following problem:

$$\begin{aligned}
&\max\{U_i^l(0; Z_i), U_i^s(\overline{Z}; Z_i)\} \qquad \text{where} \\
&U_i^l(0; Z_i) = 0 \\
&U_i^s(\overline{Z}; Z_i) = Z_i^\alpha \overline{Z}^{1-\alpha} - \frac{1}{Z_i}\left[\frac{\eta}{m}\right]^{\frac{1}{1-\eta}} \overline{Z}
\end{aligned}$$

It follows that

$$U_i^l(0; Z_i) > U_i^s(\overline{Z}; Z_i) \Leftrightarrow Z_i < \left[\frac{\eta}{m}\right]^{\frac{1}{(1-\eta)(1+\alpha)}} \overline{Z} \tag{16}$$

Since $\underline{Z} < \left[\frac{\eta}{m}\right]^{\frac{1}{(1-\eta)(1+\alpha)}} \overline{Z}$, it follows that all inventors with productivity $Z_i \in (0, \underline{Z}]$ choose to become leaders with $n^* = 0$ supporters. □

- Lemma 1: $\overline{Z} > 0$ if $m \in [1, \frac{\eta}{1-\alpha})$ and $\eta > 1 - \alpha$.

Proof. $\overline{Z} \equiv \frac{\zeta^l}{\zeta^s}\zeta^{l,s}$. For any $\alpha \in (0,1)$, $\eta \in [0,1]$, and $m > 0$, $\zeta^l > 0$ and $\zeta^s > 0$.

Assume $m \geq 1$, then for any $\alpha \in (0,1)$ and $\eta \in [0,1]$, the numerator of $\zeta^{l,s}$ is negative:

$$\left[\left[\frac{\eta}{m}\right]^\eta - m\right]^{\frac{\alpha}{1-\alpha}} < 0$$

If $m < \frac{\eta}{1-\alpha}$, then for any $\alpha \in (0,1)$ and $\eta \in [0,1]$, the denominator of $\zeta^{l,s}$ is negative:

$$\left[\left[(1-\alpha)\frac{m}{\eta}\right]^{1-\alpha} - 1\right]^{\frac{\alpha}{1-\alpha}} < 0$$

Therefore, it follows that $\zeta^{l,s} > 0$ if $m \in [1, \frac{\eta}{1-\alpha})$, where $\frac{\eta}{1-\alpha} > 1$ if and only if $\eta > 1 - \alpha$. □

- Lemma 2: More productive leaders manage bigger teams.

Proof. This result comes directly from the FOC of leaders. Specifically,

$$\frac{dn^*(Z_i)}{dZ_i} = \left[\frac{\eta}{m}\right]^{\frac{1}{1-\eta}} > 0 \tag{17}$$

since $\eta > 1 - \alpha$ is strictly positive and $m \in \left[1, \frac{\eta}{1-\alpha}\right)$ is finite. □

- Lemma 3: More productive supporters (i) join more productive leaders; (ii) join bigger teams.

Proof.

(i) This result comes directly from the FOC of supporters. Specifically,

$$\frac{dZ^{l*}(Z_i)}{dZ_i} = \frac{1+\alpha}{\alpha}(1-\alpha)^{\frac{1}{\alpha}} \left[\frac{m}{\eta}\right]^{\frac{1}{\alpha(1-\eta)}} Z_i^{\frac{1}{\alpha}} > 0$$

since $\alpha \in (0,1)$.

(ii) By plugging the optimal choice of supporters into the optimal choice of leaders, I obtain the optimal number of supporters joined by a supporter with productivity Z_i:

$$n^*(Z_i) = \left[\frac{\eta}{m}\right]^{\frac{1}{1-\eta}} \left[(1-\alpha)^{\frac{1}{\alpha}} \left[\frac{m}{\eta}\right]^{\frac{1}{\alpha(1-\eta)}} (Z_i)^{\frac{1+\alpha}{\alpha}}\right] =$$
$$= (1-\alpha)^{\frac{1}{\alpha}} \left[\frac{m}{\eta}\right]^{\frac{1-\alpha}{\alpha(1-\eta)}} (Z_i)^{\frac{1+\alpha}{\alpha}}$$

Therefore, taking the first order derivative it follows that

$$\frac{\mathrm{d}n^*(Z_i)}{\mathrm{d}Z_i} = \frac{1+\alpha}{\alpha}(1-\alpha)^{\frac{1}{\alpha}} \left[\frac{m}{\eta}\right]^{\frac{1-\alpha}{\alpha(1-\eta)}} Z_i^{\frac{1}{\alpha}} > 0$$

□

- Lemma 4: Patent quality is positively correlated with the productivity of supporters working in that patent.

Proof. By pluggin leaders FOC into the definition of patent quality, it is possible to obtain patent quality as a function of leaders productivity only:

$$q(Z_i^l) = n^*(Z_i^l)^\eta (Z_i^l)^{1-\eta} = \left[\frac{\eta}{m}\right]^{\frac{1}{1-\eta}} Z_i^l$$

The inverse of supporters FOC is

$$Z^s(Z_i^l) = K(Z_i^l)^{\frac{\alpha}{1+\alpha}} \qquad \text{with} \quad K \equiv (1-\alpha)^{-\frac{1}{1+\alpha}} \left[\frac{m}{\eta}\right]^{-\frac{1}{(1-\eta)(1+\alpha)}}$$

The correlation between quality and supporters productivity is

$$
\begin{aligned}
Corr(q(Z_i^l), Z^s(Z_i^l)) &= \frac{\mathbf{E}\big[q(Z_i^l)Z^s(Z_i^l)\big] - \mathbf{E}\big[q(Z_i^l)\big]\mathbf{E}\big[Z^s(Z_i^l)\big]}{\sqrt{\mathbf{E}\big[(q(Z_i^l))^2\big] - \mathbf{E}\big[q(Z_i^l)\big]^2}\sqrt{\mathbf{E}\big[(Z^s(Z_i^l))^2\big] - \mathbf{E}\big[Z^s(Z_i^l)\big]^2}} = \\
&= \frac{\mathbf{E}\big[(Z_i^l)^{\frac{1+2\alpha}{1+\alpha}}\big] - \mathbf{E}\big[Z_i^l\big]\mathbf{E}\big[(Z_i^l)^{\frac{\alpha}{1+\alpha}}\big]}{\sqrt{\mathbf{E}\big[(Z_i^l)^2\big] - \mathbf{E}\big[Z_i^l\big]^2}\sqrt{\mathbf{E}\big[(Z_i^l)\big]^{\frac{2\alpha}{1+\alpha}} - \mathbf{E}\big[(Z_i^l)^{\frac{\alpha}{1+\alpha}}\big]^2}} = \\
&= \frac{\mathbf{E}\big[(Z_i^l)(Z_i^l)^{\frac{\alpha}{1+\alpha}}\big] - \mathbf{E}\big[Z_i^l\big]\mathbf{E}\big[(Z_i^l)^{\frac{\alpha}{1+\alpha}}\big]}{\sqrt{Var(Z_i^l)}\sqrt{Var\big((Z_i^l)^{\frac{\alpha}{1+\alpha}}\big)}}
\end{aligned}
$$

The denominator of this fraction is positive. As far as the numerator is concerned, since $f(x) = x \cdot x^{\frac{\alpha}{1+\alpha}}$ is a convex function for any $\alpha \in (0, 1)$, it is possible to apply the Jensen's inequality:

$$\mathbf{E}\big[f(x)\big] > f\big(\mathbf{E}\big[x\big]\big)$$

and therefore obtain that the numerator is positive:

$$\mathbf{E}\big[(Z_i^l)(Z_i^l)^{\frac{\alpha}{1+\alpha}}\big] > \mathbf{E}\big[Z_i^l\big]\mathbf{E}\big[(Z_i^l)^{\frac{\alpha}{1+\alpha}}\big]$$

and $Corr(q(Z_i^l), Z^s(Z_i^l)) > 0$. □

- Lemma 5: Supporters productivity is increasing in the leader productivity, (i) at a decreasing rate in leaders productivity; (ii) at a decreasing rate in team size.

Proof.

(i) This comes directly from the inverse of the supporters FOC being increasing and concave. The inverse of supporters FOC is:

$$Z^s(Z_i^l) = K(Z_i^l)^{\frac{\alpha}{1+\alpha}} \qquad \text{with} \quad K \equiv (1-\alpha)^{-\frac{1}{1+\alpha}} \left[\frac{m}{\eta}\right]^{-\frac{1}{(1-\eta)(1+\alpha)}} > 0$$

The first and second order derivative of this function are

$$
\begin{aligned}
\frac{\mathrm{d}Z^s}{\mathrm{d}Z^l} &= K\frac{\alpha}{1+\alpha}(Z^l)^{-\frac{1}{1+\alpha}} > 0 \\
\frac{\mathrm{d}Z^s}{\mathrm{d}^2 Z^l} &= -K\frac{1}{1+\alpha}(Z^l)^{-\frac{2+\alpha}{1+\alpha}} < 0
\end{aligned}
$$

(ii) Recall that the inverse of leaders' FOC is

$$Z^l = n\left[\frac{m}{\eta}\right]^{\frac{1}{1-\eta}}$$

and the inverse of supporters' FOC is

$$Z^s(Z_i^l) = K(Z_i^l)^{\frac{\alpha}{1+\alpha}} \qquad \text{with} \quad K \equiv (1-\alpha)^{-\frac{1}{1+\alpha}} \left[\frac{m}{\eta}\right]^{-\frac{1}{(1-\eta)(1+\alpha)}} > 0$$

Consider two teams of size n_1 and n_2 such that $n_1 < n_2 = n_1 + \delta$ with $\delta > 0$. The productivity of a leader managing a team of size n_j is Z_j^l, and it holds that

$$\begin{aligned} Z_1^l &= n_1 \left[\frac{m}{\eta}\right]^{\frac{1}{1-\eta}} \\ Z_2^l &= n_2 \left[\frac{m}{\eta}\right]^{\frac{1}{1-\eta}} = \delta \left[\frac{m}{\eta}\right]^{\frac{1}{1-\eta}} + Z_1^l \end{aligned}$$

Similarly, the productivity of a supporter joining a team of size n_j is Z_j^s, and it holds that

$$\begin{aligned} Z_1^s &= K(Z_1^l)^{\frac{\alpha}{1+\alpha}} = K \left[\frac{m}{\eta}\right]^{\frac{\alpha}{(1+\alpha)(1-\eta)}} n_1^{\frac{\alpha}{1+\alpha}} \\ Z_2^s &= K(Z_2^l)^{\frac{\alpha}{1+\alpha}} = K \left[\frac{m}{\eta}\right]^{\frac{\alpha}{(1+\alpha)(1-\eta)}} \big(\delta + n_1\big)^{\frac{\alpha}{1+\alpha}} \end{aligned}$$

Supporters productivity is increasing in team size at a decreasing rate because

$$\begin{aligned} \frac{\mathrm{d}Z_2^s}{\mathrm{d}\delta} &= K\frac{\alpha}{1+\alpha} \left[\frac{m}{\eta}\right]^{\frac{\alpha}{(1+\alpha)(1-\eta)}} \big(\delta + n_1\big)^{-\frac{1}{1+\alpha}} > 0 \\ \frac{\mathrm{d}Z_2^s}{\mathrm{d}^2\delta} &= -K\frac{\alpha}{(1+\alpha)^2} \left[\frac{m}{\eta}\right]^{\frac{\alpha}{(1+\alpha)(1-\eta)}} \big(\delta + n_1\big)^{-\frac{2+\alpha}{1+\alpha}} < 0 \end{aligned}$$

□