Million Ideas Plants: Do Occasional Inventors Benefit from Local Highly Patenting Companies?

Carlo Menon*

October 22, 2012

Abstract

The paper investigates whether the patenting activity of the most inventive companies (stars) has any causal effect on the number of patents granted to other local inventors (comets) in the same US metropolitan area. I exploit the panel dimension of the dataset to account for various fixed effects, and I adopt an instrumental variable approach to test for causality. The results show that the effect of stars on comets is overall positive, it is stronger with a time lag, and the spillovers are not confined within narrow technological categories. The implications for local development policy are discussed.

Keywords: localized knowledge spillovers, patents, innovation.

JEL classification: R10; O31

*STI Directorate, OECD. Email address: Carlo.menon@oecd.org. This article is a deeply revised version of a previous working paper titled "Stars and Comets: an Exploration of the Patent Universe". I thank Alfonso Gambardella, John Van Reenen, and other partecipants to the Zvi Griliches Seminar on Economics of Innovation (Barcelona, July 2011) for useful comments. The views expressed in the article are those of the author and do not involve the responsibility of the OECD or the Bank of Italy.

1 Introduction

In a recent paper, Greenstone, Hornbeck and Moretti (2010, henceforth GHM) discuss the importance of understanding whether 'million dollar' plants raise the productivity of local incumbents. They find a substantial effect: five year later the opening of a new large plant, the productivity of incumbent plants located in the same US county is 12% higher. The authors argue that the findings are extremely relevant both for economic theory, since they provide evidence on the mechanisms underlying the agglomeration of economic activities, and for local development policies, which often subsidize the location of large industrial investments.

At least since Marshall (1980), we know that knowledge spillovers are one of the leading mechanisms of agglomeration economics. This paper focuses primarily on this channel, which is not directly addressed by GHM. Exploiting a rich patent database for the United States, I identify *Million Ideas Plants* by proxying the company size with the stock of owned patents. I then assess whether the aggregate number of patents developed by inventors working for star companies (star patents) has any causal effect on the number of patents granted to other inventors (comet patents) located in the same Metropolitan Statistical Area (MSA). A priori, the effect is not necessarily positive: in a general equilibrium framework, positive knowledge spillovers produced by star companies may be counterbalanced by upward pressures on nominal wages. Depending on the relative strength of the two mechanisms, the net effect could be null or negative.

Causality is inferred through an original identification strategy based on two-stage least squares. I build an instrumental variable for star patents based on the interaction among the historical presence of star companies in a given MSA, and on the contemporaneous variation in innovation activity of other plants of the same company. Using the NBER/USPTO patent database, I estimate a model where the number of comet patents produced in a given city, time period, and technological category is a function of the number of star patents developed in the same city, period, and category. I exploit the panel dimension of the dataset to account for time, city and technology fixed effects. Consistently with economic theory, results show that positive effects prevail with a broad sectoral classification and a time lag, while negative effects are stronger in the short run and within narrow technological sectors. The net effect, however, is generally positive and significant.

This paper fills two important gaps in the related economic literature. First, by providing empirical evidence on the effect of knowledge spillovers, it improves our understanding of the mechanisms of urban agglomeration. Local knowledge spillovers have gained the interest of economists at least since Marshall (1980), but the quantification of their importance has been difficult. Second, the study is related to the wide literature on the economics of innovation and patenting, by exploring under-investigated aspects of patent data, i.e., the skewness of the distribution of patents across inventors and companies, and the connections between more and less productive inventors.

The paper also offers interesting insights for policy. It is well known that innovation activity is highly concentrated in a small number of cities and regions; these spatial disparities have pushed a number of policies aimed at enhancing local innovation (Agrawal et al., 2012), often based on subsidizing the location of R&D labs of large companies. Very little is known, however, about the effectiveness of these policies, i.e., whether they produce any additional effects on the innovation performance of local firms, or, rather, are just a windfall for large companies, at the taxpayer's cost. More generally, the paper contributes to the lively debate on cluster policies: a very popular local development tool among policy makers and academics alike (e.g., European Commission, 2003; OECD, 2001; Porter, 1998), they still lack of substantial empirical evidence for the alleged benefits, according to a number of economic studies (Accetturo and de Blasio, 2012; Duranton, 2011; Martin, Mayer and Mayneris, 2011; Duranton, Mayer and Mayneris, 2010).

The next paragraph reviews the relevant literature on patents and innovation; the third one introduces and discusses the definition of stars and comets; the fourth one describes the empirical methodology and the fifth presents the results; the sixth paragraph concludes.

2 Patents, localized knowledge spillovers, and the size of innovation

Patent data have become extremely popular in the economic literature in the last two decades, as they represent an easy and accessible way to proxy for innovation, which is generally very hard to measure. Furthermore, the availability of citation linkages further spurred more interest in patent data: for the first time, researchers had a tool to "trace" knowledge spillovers, which previously had been considered one of the most difficult variable to define empirically. A popular book by Jaffe and Trajtenberg (2005), and the free availability of the USTPO dataset from the NBER website, also contributed to multiply the empirical applications based on patent data.

A significant part of this literature has focused on the geographic component of innovation, with a particular interest in the spatial decay of knowledge spillovers. A seminal contribution by Jaffe et al. (1993) shows that a cited-citing patent couple is twice as likely to be in the same US metropolitan area as a couple of technologically similar patents with no citation links.[1]

[1] These findings have been strongly criticized by Thompson and Fox-Kean (2005), who argue that the methodology underlying the construction of the control group is seriously flawed. With a more robust approach, based

Similarly, Peri (2005) examines the flows of citations among 147 European and US regions to find that "only 20% of average knowledge is learned outside the average region of origin", and Jaffe (1989) demonstrates that academic research has large effects on the number of private patents developed in the same US state. Finally, Carlino et al. (2007) use patent data for a cross-section of US metropolitan areas to investigate the relationship between urban density and innovation intensity (as measured by patents per capita) finding a positive and robust association. All these contributions (and many similar which I omit for brevity) highlight that knowledge spillovers have a geographically limited distance decay.

The nature and causes of knowledge spillovers are still debated. For instance, Breschi and Lissoni (2009), building on previous contributions by Breschi and Lissoni (2001), Zucker et al (1998), and Almeida and Kogut (1999), highlight that defining localized knowledge spillovers as an *externality* can be misleading, as most of the knowledge diffusion may take place through market interactions - namely the spatially-bounded mobility of inventors among workplaces - rather than through informal contacts. Using data on US inventors' applications to the European Patent Office, they were able to show that after controlling for inventors' labour mobility and the related professional network, the role of proximity in explaining knowledge diffusion is greatly reduced.

These issues are related to the growing interest in peer effects in science and in the spillovers originating from star scientists. Azoulay et al. (2010) exploit the exogenous variation in the number of "superstar scientists" in US universities due to the sudden death of these individuals to estimate the loss in productivity of their collaborators. They find an average 5-10% decline in their average publication rates, starting 3-4 years after the superstars' death and enduring over time, but no differential effect for co-located collaborators. Waldinger (2010) estimates the effect of the dismissal of scientists from Germany universities during Nazism. Similarly to Azoulay et al., he finds a strong effect on coauthors (13-18%), but no significant effects at department level. Therefore, both studies challenge the existence of localized positive spillovers originating from stars in academic environments.

Similarly, the advocates of the "death of distance" theory argue for a decreasing importance of the role of spatial proximity following the progress of communication technologies (*e.g.*, Friedman, 2005; Quah, 1999; Cairncross, 1997). On the other side, a few studies maintain that technological progress has actually increased the scope for proximity for innovative activities due to the greater importance of face-to-face contacts and agglomeration externalities (*e.g.*, Coyle, 1999). The few empirical assessments of the issue seem to support the "death of distance" hypothesis (Griffith

on a finer technological classification of patents, the main results of the paper disappear.

et al., 2007; Ioannides et al., 2008), indeed suggesting that localized knowledge spillovers are fading over time.

The relationship between highly inventive companies and inventors in small firms have been much less explored: to the best of my knowledge, contributions on the subject are confined to the role played by academic star scientists on other researchers (*e.g.*, Azoulay et al., 2010; Oettl, 2011), while industrial patenting is not considered. The only exceptions are Fons-Rosen (2010), and a very recent working paper by Agrawal et al. (2012). Fons-Rosen (2010) uses data on the entry of foreign firms into Central and Eastern Europe during the 1990s to analyze the effect on knowledge flows on local incumbent inventors; he compares the MNEs which won the privatization bids with the control group of those which also applied to the bid but lost, finding that winners receive 20% more citations by local inventors, on average, than losers. Differently from this paper, its analysis is at national level and is limited to patent citations. Agrawal et al. (2012) explore the spatial distribution of large and small (patenting) labs across US MSAs, finding that the birth rate of new start-ups (defined using patents filed for by inventors who were previously employed by large labs) is higher in metropolitan areas which are more diverse, i.e., where large and small labs coexist.

Occasional inventors are important, since they may give birth to new entrepreneurial projects and spin-offs. Balasubramanian and Sivadasan (2011) in a recent paper link patent records to Census firm data for the US, in order to assess the impact of patents on firm performance. They focus in particular on firms that patent for the first time, and find a significant and large effect of the first patent on firm growth (but, interestingly, little change in factor productivity). This would suggest that "occasional" patents have a relevant market value. There is also large empirical evidence on the primary role played by young and small firms in innovation (Acs and Audretsch, 1990) and employment growth (Audretsch, 2002; Haltiwanger et al., Forthcoming). Furthermore, for the local development policies, patents by smaller companies are probably more relevant than patents filed for by large corporations. To the extent that the latter are the outcome of formal R&D activity of large companies, they may have weaker implications on the local economy. Since patenting firms are generally large (Balasubramanian and Sivadasan, 2011), they are often multilocated, and the productivity gains of these inventions are spread across the different plants (and localities).

3 Stars and comets

The analysis is based on the NBER/USPTO database, which lists all the patents granted in the United states from 1975 to 1999.[2] For each patent, the database contains the name and city of residence of the inventor(s), the name of the applicant(s),[3] an unique applicant identifier added by the NBER working group on patents (based on the standardization of the name of company and ancillary information), the application and grant year, and the number of citation received. Patents are classified according to the synthetic technological classification developed by Hall et al. (2001) who define five technological categories: Chemical (excluding Drugs); Computers and Communications (C&C); Drugs and Medical (D&M); Electrical and Electronics (E&E); Mechanical.[4] Following a common practice in the patent literature, the geographical location of the patent is derived from the location of the first inventor. More details on the data, including the geocoding process, are reported in Appendix A.[5]

At first glance, the abundance of data makes a micro analysis at inventor or applicant level the most appealing alternative. A deeper view, however, clarifies that this is not feasible, because the dataset is about patents, not inventors or applicants, implying that when an inventor or applicant is not patenting, their location and their activity status are unknown. Furthermore, there is not an unique inventor identifier in the dataset, and the only information available is their full name and their city of residence. Spelling errors are frequent. As a consequence, the longitudinal tracking of inventors would require a fuzzy matching of names and cities of residence, with inevitable errors which can easily be non random (*e.g.*, more frequent in cities where duplicate surnames are more common, or with a higher rate of inventors with foreign origins). The problem would be perhaps negligible if the focus was only on very productive inventors or applicants; but given that I am interested also in comets, the issue is crucial.

The analysis is therefore run at city level, focusing on the number of *patents* produced by two groups of applicants: *star* patents and *comet* patents. The classification of patents into the two groups is based on the total number of patents granted to the applicant in the whole period of analysis (1980-1999): stars patents are assigned to the most inventive companies in a given technological categories, while the remaining patents are defined as comets. The most inventive companies are defined as those ranked among the top 50 in their technological category. The large majority of those companies are located in more than one MSA, which implies that they

[2]The dataset is described in details in Hall et al., 2001.

[3]The applicant is the legal entity - either a company or an individual - which owns the right to exploit the invention. In the large majority of cases, the applicant is the employer of the inventor.

[4]The sixth technological category, called "other", is a residual classification and is excluded.

[5]The paper has recently been extended until the 2006 or 2008, depending on the version. However, inventors data are not publicly available yet; without information on the city of residence of inventors,it is not possible to correctly geolocate the patents, therefore the date cannot be used here.

can plausibly be the target of location incentives. As it will be further detailed, this is also a crucial feature of the identification strategy. In the robustness section, I repeat the empirical exercise changing the threshold (limiting the definition to the top 25 or the top 75 companies), finding little variation in the results. The threshold is based on an absolute number, rather than on a quantile (*e.g.*, top 1%), since the total number of active firms is difficult to define in the patent database.[6] Also, since the patent literature offers many examples of large companies filing for patents for reasons unrelated to new inventions (*e.g.* patent thickets), and considering that generally such non-inventive patents are not cited by other patents, I exclude from the star group all patents which do not receive any citations. In the robustness section I replicate the analysis with forward citations weighted patents obtaining comparable results.

The following step is the definition of the temporal dimension of the analysis. The data used are rather imprecise in the time dimension for the following reasons: the information is based on the year in which the patent is granted,[7] which is generally 2-3 years after the first application; and it is not possible to know how long an inventor has been working on a patent before applying for it. Timing when local knowledge spillovers may have effect is equally difficult: it could be while both the source and destination inventors are working on their respective patents, but it could also happen a few years after the star has applied for (or been granted) it. By inspecting the data I found that the median and mean value of the citation lag of patents in the same MSA is four years, and I therefore choose to adopt periods of the same length (Kerr, 2008, also adopt a period of the four years).[8] This is a reasonable choice in order to "average out" some of the measurement error in the temporal dimension. Five time periods of four years each are therefore defined, spanning from 1980 to 1999. Econometric analysis is generally limited to the last three periods (from 1989 to 1999), as MSA controls are unavailable for periods 1 and 2. I define five periods, however, as the first is used to build the instrumental variables and lagged variables.

Star patents account for 9% of the total patents granted in the period 1980-1999, while comet patents are the remaining 91%. Star inventors are around 28% of the total. The metropolitan area with the highest share (more than 80%) of star is Poughkeepsie, NY (home to a large IBM plant), while in Bakersfield, CA, almost all patents are comets.

[6]Similarly to inventors' identifiers, the unique identifier for small companies is not fully reliable due to spelling errors, homonymy, and changes of name across time; the identifier for large companies is somehow more reliable, due to their smaller number and the notoriety of their different denominations.

[7]The reason why I use the grant year, rather than the application year, is to avoid the bias given by data truncation. More precisely, using the application year would automatically exclude all the patents not granted (but applied for) before 1999, as they are not included in the dataset. This subsample could easily be non-random, *e.g.* better patents may take longer to be examined, etc.

[8]I restricted the calculation to patent couples with a maximum citation lag of ten years, as longer lags are unlikely to be related to knowledge spillovers. The citation lag is calculated as the difference between the grant year of the citing and cited patents.

3.1 Preliminary evidence on the location of stars and comets

Once controlling for the overall distribution of patenting activity, which is highly spatially concentrated, the distribution of comets and stars patents across MSAs shows a remarkable spatial concentration (figure 1 and 2). Some MSAs like Rochester, NY or Saginaw-Bay and City-Midland, MI are clearly specialized in stars, while other show a stronger presence of comets (*e.g.*, Reno, NV and Omaha, NE-IA). In order to explore in further detail the spatial distribution of star and comet clusters, Appendix B presents the result from simple regressions of the share of stars and comets on a set of MSA-specific coovariates. The results show that comet patents are negatively associated with the total number of patents and positively with the total number of firms. Conversely, star patents are positively associated with both the number of patents and a proxy for labour productivity, suggesting that star patents are more frequently located in cities with a large number of patents and a more skilled workforce. The other explanatory variables are not significant. Therefore, comets and star patents are not evenly distributed over cities, the share of stars being significantly larger in bigger metropolitan areas. This is a first insight policy makers should consider when promoting subsidies to attract stars.

3.2 Why should stars affect comets?

An increase in the number of star patents, due to an increase in the productivity or in the number of star inventors, may have both positive and negative effects on the number of comet patents in the same city.

As I discussed earlier in the paper, positive effects may occur through knowledge spillovers. The idea that cities foster the diffusion of knowledge goes back to Marshall (1890) and it is the backbone of endogenous growth theory (Lucas, 1988). Duranton and Puga (2004) provide an excellent survey of the "learning mechanisms" posing the microeconomic foundations of the existence of cities. However, there is still limited evidence on the channels through which knowledge spillovers take place (Feldman and Avnimelech, 2011). In the specific context of patenting in cities, it is possible to think about at least five different channels:

a) Informal (or tacit) knowledge spillovers: star inventors and comet inventors develop informal (personal) contacts due to residential proximity or other kind of face-to-face interactions. Thanks to frequent direct contacts with the star inventor, the comet inventor obtain ideas or hints on their work.

b) Formal knowledge spillovers: star inventors transfer their expertise to comet inventors in more formal ways, *e.g.* during seminars or conferences.

c) Workplace contacts: (future) comet inventors may have the opportunity to work in a star company, without necessarily being inventors themselves (they may be employed in different duties, or they may leave the institution at an early stage of their career).

d) Workplace mobility and spin-off: active star inventors leave star companies and create their own company, or they are hired by a smaller local company. As correctly pointed out by Breschi and Lissoni (2009) and Almeida and Kogut (1999), the previous experience may be fully priced into the inventor's wage, thus in this case the spillover is not an externality.

e) Display/attraction effects: the presence of many labs of big companies may attract comets to a locality, as they may expect to enjoy the effects of points a, b, and c. For a young firm, the location in a successful city may be also a positive signal to potential founders. This is therefore an indirect form of positive knowledge spillover.

All the five mechanisms may require some time to become effective, thus they may be found in the data with a time lag. Their technological boundaries are also fuzzy: given that they are often involuntary, tacit knowledge spillovers may be technologically complementary to other, market-mediated, forms of learning, which are a direct and conscious objective of the inventor. In other words, the inventor may look (and pay, even in opportunity costs) for spillovers which are closely related to her field of specialization and which she can capitalize with less risk. On the other hand, the outcome of cross-technology spillovers may be unpredictable, therefore those are not an intentional goal the inventor is ready to pay for, since it would be a too risky investment. This line of reasoning recalls the theory of "cross-fertilization of ideas" developed by Jacobs (1969) and the related economies of diversity, later formalized and empirically validated by Glaeser et al. (1992).

Potential negative effects may be derived in a general equilibrium approach to local labour markets (Moretti, 2011), and they may mainly occur through an increase in nominal wages. Indeed, a raise in innovation activity in a local star plant corresponds to an upward shift in the demand for local scientists, which in turn raises local nominal wages in the sector, at least in the short run (in the longer run, workers may migrate in from other cities, but the inflow is limited by the local supply of housing which affects real wages). Both mechanisms affect negatively the number of local comet patents, since local scientists become more costly, without a corresponding increase in productivity (assuming zero knowledge spillovers). The actual impact of these effects depends on the skill substitutability among star and comet inventors, on the elasticity of supply of labour (also through migration). Since the latter is likely to be rigid in the short run, the negative effects are expected to be stronger in the short term, and then fade over time. Also, I expect the negative effects to be stronger within narrowly-defined technological sectors, since skill substitutability of workers is higher, and correspondingly the wage effect is larger.

In the following of the analysis, I estimate a simple reduced form model, taking into account only the net effects of all the aforementioned mechanisms. Exploring the effect of individual channels would be much more demanding - also in terms of data - and it is above the scope of the present paper.

It is worth noticing that all the mechanisms may, in theory, work also in the opposite direction (from comets to stars); the empirical methodology is therefore designed to be robust to reverse causality.

4 Analysis

This section investigates whether the production of star patents in a city affects the production of comet patents in the same city and period, and quantifies this effect. The model also allows a one period lag in the spillover effects of stars. The following dynamic panel with fixed effects is estimated:

$$\begin{aligned} Comets_t^{ik} = \beta_1 \cdot Stars_t^{ik} + \beta_2 \cdot Stars_{t-1}^{ik} + \beta_2 \cdot Comets_{t-1}^{ik} + \theta_1 \cdot \sum_{j \neq i} Z_t^{jk} + \\ + \theta_2 \cdot \sum_{j \neq i} Z_{t-1}^{jk} + \gamma \cdot Totemp_t^i + \phi^i \delta^k + \delta^k \tau_t + \varepsilon_t^{ik} \quad (1) \end{aligned}$$

where i, k, and t index MSAs, categories, and periods, respectively; stars, comets are the number of patents in the respective groups, Z is a control specific to the MSA/category pair, X is a set of MSA time-variant controls, and δ, τ, ϕ are category, time, and MSA fixed effects. The five technological categories are the following: Chemical (excluding Drugs); Computers and Communications (C&C); Drugs and Medical (D&M); Electrical and Electronics (E&E); Mechanical.[9] The analysis is limited to periods 3-4-5, as MSA controls are not available for previous periods. All the variables are expressed in logarithmic form.

In order to check the consistency of the results across different specifications, regressions are based on four different estimations of model 1. The first is an inconsistent OLS estimation including all the controls but the (endogenous) lagged comets. The second and the third are 2SLS estimates of the contemporaneous and lagged star variable, respectively, excluding all other continuous controls and including all fixed effects; details on the instrumental variable strategy are reported in the next section. Finally, the fourth estimation includes all the controls, the full set of fixed effects, and the lagged dependent variable; in this case, the model is estimated in first differences and the flow of lagged comets at time *t-1* is instrumented with the level value at time

[9] The sixth technological category, called "other", is a residual classification and is excluded. This does not affect the coefficients but increase precision of the estimates.

t-2, following the traditional Arellano and Bond (1991) technique. For easy of comparability with the second and third estimations, the fourth estimation is also based on 2SLS.[10]

The MSA/category control (Z) is total number of stars in technological categories different from *i*. It is worth noting that this variable might be endogenous: it is not possible to exclude that comet inventors produce knowledge spillovers benefiting stars in the other technological categories. However, the inclusion of this variable has a limited effect on the main coefficients, especially with the 2SLS estimator. As the latter is robust to omitted variables bias, the estimate of the coefficient for the variable of interest (the number of star patents) is consistent even excluding the (endogenous) control.[11]

The total MSA employment (*totemp*) is also included to control for time-variant agglomeration economies and size effects.

Finally, as mentioned aboveall regressions include various fixed effects, controlling for technological category and MSA time invariant factors, for time-specific shocks, and for technological category shocks.

4.1 Technological category specification

Data are aggregated following two different technological classifications, i.e., the technological category and subcategory breakdown developed by Hall et al., 2001. In addition, the model is also estimated with aggregated data. The three different level of technological aggregations may give interesting insights on the technological boundaries of knowledge spillovers.

4.2 The choice of the MSA as areal unit

Ideally, the spatial unit at which individual observations are aggregated should match the spatial decay of both knowledge spillovers and labour market clearing forces. Since both boundaries are indefinable entities, the spatial definition should entail a substantial degree of approximation; furthermore, data limitation are particularly stringent at a detailed geographical level. With respect to labour market analysis, the choice of commuting-defined areas, like the MSAs in US, is now widely considered to be a viable option.[12] The definition of the spatial decay of knowledge spillovers is more debated: while several studies have adopted spatial areas as large as US States

[10]In the fourth estimation, the set of endogenous variables are the first differences of stars, the first lag of the first differences of stars, and the first lag of the first differences of comets; the excluded instruments are the contemporaneous level and first lag of the ad-hoc exogenous instrument described in the next section, and the first lag of the first difference of comets. The 2SLS estimation comes at a cost of reduced efficiency as compared to GMM, but efficiency does not appear to be a major issue in this context. Comparable results based on GMM estimations are available from the author upon request.

[11]Attempts to instrument the variable with the sum of the instrument in the other categories provide similar results, but estimates were less precise, due to the large number of endogenous variables and instruments.

[12]See Menon (2012) for a discussion of the statistical properties of MSAs.

(*e.g.* Jaffe, 1989; Peri, 2005), available evidence suggests that the effect of knowledge spillovers may fade out within a few miles (Arzaghi and Henderson, 2008; Jofre-Monseny et al., 2010).

Since the reduced-form effect we estimate is supposed to be a mix of labour market and knowledge spillovers mechanisms, the Metropolitan Statistical Area is the most sensible spatial unit of analysis, among the limited number of available options. As a consequence, the effect of short-decay knowledge spillovers may be underestimated. It is therefore appropriate to specify that the analysis takes into account only MSA-level knowledge spillovers, which may not fully reflect other short-decay spillovers.

4.3 Instrumental variable estimation of the star variables

Estimates of equation 1 can be inconsistent due to reverse causality or omitted variable biases, especially for the main variable of interest (the number of star patents). For instance, comets may affect the productivity of stars, and a dynamic university (or public subsidies) may attract a large number of comet and star inventors to the same city. I therefore create an instrumental variable for the number of star patents in order to allow a causal interpretation of the results.

The intuition for the instrument builds on the fact that assignees of stars are generally multilocated. Most star companies are located in several MSAs in different US states. Table 1 lists the top 25 assignees in the period under examination (1980-1999), reporting the number of different MSAs and states where at least 100 patents are developed, and the highest share of patents developed in an individual MSA: only two companies are located in only one MSA (Ford Motor and Procter & Gamble), while all the remaining assignees are located in several different cities and states. Smaller assignees of star patents show a similar pattern. Therefore, an exogenous variation in the productivity of star inventors in a given MSA and period may arise from the interaction of two factors: i) an historical presence of inventors working for a given company in that MSA, and ii) a US-wide increase in the productivity of this company in the given period. To the extent that the first factor is path-dependent and exhibits some inertia over time, it is exogenous to contemporaneous MSA-specific factors once MSA fixed effects are introduced in the specification. At the same time, I expect the productivity of star inventors working for the same companies - but in different cities - to be correlated, due to sharing a similar company strategies and resources, competition pressure, market demand, etc. I then suppose that a US-wide productivity shift in a given company translates into MSA-specific productivity shocks in proportion to the number of inventors working for that company in the given MSA.

The IV strategy is close in spirit to the approach of Bartik (1991) and Blanchard and Katz (1992), among others, who instrument regional economic growth interacting the lagged sectoral structure of a region with the contemporaneous national sectoral trend. In the next section the construction of the instruments is explained in detail.

4.3.1 Instrumental variable construction

The instrumental variable is calculated as follows:

a) For the first period, I calculate the share of star inventors active in a given MSA and with a given assignee. In the case of star inventors with multiple MSAs or assignees in the same period, the modal one is chosen.

b) For each period, each assignee, and each MSA, I calculate the average number of patents produced by star inventors in that period in the whole US, excluding the given MSA.

c) For each MSA, period, and assignee, I multiply the number of inventors in the first period calculated at point a) by the average number of patents produced by star inventors sharing the same assignee in period t calculated in b). Subsequently, I sum the outcome by MSA, period, and technological category (if an inventor has patented in different categories in the same period, the modal one is chosen). The result is the second instrumental variable for total number of star patents in period t, by MSA and category.

Formally, it can be summarized by the following equation:

$$IV_{ikt} = \Sigma_a(StarsInv_{ika1} \cdot AvPat_{iat})/Pat_{ik,t=1} \tag{2}$$

where i indexes MSAs, t periods, k technological categories, and a the assignees. In the few cases in which the value of point b is missing (because there are not other stars with the same assignee in other MSAs), it is replaced with the contemporaneous US-wide average productivity of stars in the same technological category. The value of the IV is then divided by the total number of star patents in period 1 in the MSA-category to neutralize the scale factor.

The validity of the IV relies on an exclusion restriction for point a), i.e., once MSA fixed effects are controlled for, the number of star inventors working for a given assignee in the first period has no independent effect on the number of comet patents developed in period n in the same MSA/category; and on an assumption of exogeneity for b), i.e., given that stars and comets have different assignees, I assume that the average productivity of an assignee in the whole US (calculated excluding the given MSA) has no independent effect on the productivity of comets of that MSA.

It could be, however, that the address of residence of a few inventors does not truly reflect the location of their workplace while working on the patent; this can be due to errors in the data or geocoding process, or to a subsequent change in the inventor's address. This would threaten the exogeneity of the IV, as the productivity of those inventors would not be exogenous to local unobservables. When building the IV, I therefore drop from the sample all the company-MSA pairs with less than 100 patents; i.e., I exclude those patents located in MSA where the given assignee is not patenting with regularity.

5 Results

The regression results are reported in table 2-4.[13] As noted in the previous section, the estimations are based on four different specifications - OLS with controls, 2SLS with contemporaneous stars only, 2SLS with lagged stars only, and first-difference dynamic panel with the full set of variables - at three aggregation levels: MSAs (table 2), MSAs and five technological categories (table 3), and the MSAs and 27 technological subcategories (table 4). All columns with IV estimation also report the Angrist and Pischke (2009, pp. 217-18) first-stage F statistics for tests of weak identification when there is more than one endogenous regressor. When just one variable is considered to be endogenous, the test is equivalent to the traditional first stage F-statistic.[14]In most cases, results from first-stage regressions confirm that the instrument is strong, especially at category and subcategory level.[15]

At MSA level (table 2), the contemporaneous effect of stars is always significant, and the coefficient ranges from 0.8 to 0.15. With a one period lag, the effect of star patents is still significant and the 2SLS coefficient ranges between 0.10 and 0.17. These results suggest that technological boundaries are not particularly important for knowledge spillovers, and that they require some time before becoming effective (when both are included, the lagged coefficient is always larger than the contemporaneous one). The other control variables are not significant.

When the sample is decomposed into technological categories (table 3), the contemporaneous coefficient is now not significant, while the 2SLS elasticity for the lagged star variable is equal to 23% in both the 2SLS regressions in which it is included (col. 3-4). The control variables are

[13]Standard errors are clustered at at the cross-sectional unit of observation, i.e., the MSA, the MSA-category, or the MSA-subcategory. Alternative estimates based on clustering at the state-year pairwise combination give almost identical standard errors. Since the distribution of total patents across MSAs shows a large variance, all regressions are (analytically) weighted by the total number of patents over the period of analysis (see Angrist and Pischke, 2008, for a detailed discussion on the suitability of weighted regressions when the sample is composed by grouped individual observations). We also dropped all the MSA-Category pairs with less then 10 patents over the whole period of analysis. Unweighted regression results and full sample results are qualitatively similar but less precise. They are available from the author upon request.

[14]The Angrist-Pischke (AP) first-stage F statistic is calculated for each individual endogenous regressors by "partialling-out" linear projections of the other endogenous regressors. The AP test will fail to reject if a particular endogenous regressor is unidentified. Values of the AP first-stage F can be compared to the Stock-Yogo (2002, 2005) critical values for the Cragg-Donald F statistic with K1=1.

[15]The only case in which the AP statistic is critically low is in col. 4 of table 2; the low value refers to the 1st stage regression of the lagged comet IV. The results of this column should therefore be interpreted with caution, although they are fully consistent with more robust specifications.

significant only in the OLS regression (col. 1).

When the sample is further decomposed into 27 technological subcategories (table 4), results are qualitatively similar, although the size of the main coefficients is reduced. While the lagged star coefficients are still positive and overall significant, its 2SLS value is now equal to 0.12. The contemporaneous coefficient is small and (weakly) significant only in col. 4. The lagged number of comet patents in the same MSA and subcategory has a significant and large effect, suggesting that spillovers within the comet group are more technologically bounded than spillovers from stars to comets.

Summing up, the results suggest that i) the effect of stars on comets is overall positive, ii) it is stronger with a time lag, and iii) positive stars-to-comets spillovers are not confined within narrow technological categories. Consistently with the theoretical predictions, the results also suggest that factor cost effects (upward pressure on wages), which push the coefficient downward, are stronger within narrow technological categories and in the short run, while knowledge spillovers take a few years before being effective, and may be technologically complementary.

Why OLS estimates are downward biased? There are three plausible explanations for that: negative selection, measurement error, and local average treatment effect (LATE). Negative selection may arise because, in general, those star inventors that are more "exposed" to comet inventors might produce less knowledge spillovers than the average star inventor. In other words, star inventors localized in "comet cities" may be "worse" than star inventors localized in "star cities". As this lower quality is unobserved, it introduces a (downward) bias in the OLS estimates. Another plausible explanation for the downward bias could be a measurement error in the star variable: the intensity of activity of star inventors in a locality is approximated by the number of patents they produce, but the measure is clearly noisy, as patents are heterogeneous in quality. To the extent that the measurement error of the instrumental variable is independent from the one in the endogenous variable, IV estimates may eliminate the "attenuation bias" of the OLS coefficient. The independence of the two errors is plausible as the variables are measured using patents in different localities (in the specific city and in the whole US excluding that city, respectively). Finally, to the extent that the elasticity of the endogenous regressor to changes in the instrumental variables is not constant across groups, 2SLS estimates may correspond to a local treatment effect, rather than to an average treatment effect (ATE) (Imbens and Angrist, 1994). In this specific context, it is likely that the elasticity of the endogenous variable to the instrument is higher for incumbent plants, since one of the component of the instrument is the historical presence of star inventors in the MSA. Incumbent plant inventors may have a stronger effect, since they are more connected with local comets; this may explain an higher local treatment effect.

5.1 Robustness

The first robustness test challenges the choice of limiting the definition of star company to the top 50 companies within a given technological category. I replicate the analysis with two different ranking thresholds, equal to 25 and 75, respectively. The results of the first-differences IV estimation with the lagged dependent variables are reported in table 5, all the other are available from the author upon request. The values of the main coefficients are close to those presented above, and my general conclusions are unaffected by the change in the threshold.[16]

Another point of concern is the choice of considering only the first author of the patent in the geolocalization process.[17] This is based on the assumption that the first author is the leading scientist, but it would introduce a bias if authors are listed in alphabetical order. Therefore, in table 6 I check whether authors whose surname begins with one of the first letters of the alphabet are more likely to be reported as first author, compared with second or third authors, finding that differences in probability are very low and fade out after the first five-six letters. This evidence therefore suggests that the first author should be the project leader.

The choice of patent count as a measure of productivity of star inventors may also be questioned, since patents are very heterogeneous in quality and value. As a consequence, star patent counts can be a very noisy proxy. Although I exclude from the star group all patents which do not receive any citations, this might not be enough. It we interpret the patent value heterogeneity as a classic measurement error leading to attenuation bias, to the extent that the measurement error in the instrumental variable is independent from those in the endogenous variable, the 2SLS estimation strategy would be sufficient to get rid of the bias. Unfortunately, given that both the variables refer to the same company, the assumption of the independence of the measurement error may not hold. There is, however, another solution available, i.e., weighting star patents by the number of forward citations, since the latter has been shown to be a reasonably good proxy for patent value (Hall *et al.*, 2005). I can then replicate the analysis using the quality-corrected measure of star patents. The results of the first-differences specification are reported in table 6, all the other tables are available upon request. Again, coefficient values are close to those of the main specifications.

[16] The star coefficients tend to slightly increase with a lower threshold, while the lagged comet comet tend to decrease. This is due to the fact that part of the effect attributed to comet patents with a higher threshold is mechanically shifted to the star group once a lower threshold is used. The change in the coefficient value is however small, and seldom statistically significant.

[17] In the patent literature, using only the firs author is probably the most common option, although some researchers also use fractional count or multiple allocation.

6 Conclusions

This paper assesses whether the number of patents developed by inventors working for the most inventive companies (star patents) has any causal effect on the number of patents granted to other inventors (comet patents) located in the same MSA. The two categories of patents - stars and comets - are identified according to the total number of patents owned by their assignees.

Economic theory predicts that an increase in innovation activity of star companies affect the production of comet patents positively through knowledge spillovers, and negatively through increased local wages. The empirical findings are coherent with the theoretical framework: results show that positive effects prevail; they are stronger with a time lag and are not necessarily bounded within sectors, providing support for the relevance of economies of diversity. On the other side, net effects are smaller in the short run and within narrow technological sectors. Results survive to a number of potentially demanding robustness tests. These findings are in line with a substantial stream of research proving the economic relevance of localized knowledge spillovers.

The findings also bring in relevant implications for local development policies. As discussed by GHM, policy makers are increasingly keen in subsidizing the local investments of large companies, with the idea that these may generate agglomeration spillovers and benefit local firms. Do my findings provide ground to these policies? Given the positive effect of stars on the productivity of comets, the attraction of stars to a city may have a positive effect on the local economic environment: in the medium run, stars positively affect comets, which in turn might foster the birth of new plants, the innovation output of small businesses, and the generation of new employment. Thus, even though R&D labs of big corporations may have only a limited direct effect on the local economy, as most the of the employment and value added is located elsewhere, they might still be beneficial. However, the attraction of stars may impact sectors and time periods which are not those directly affected by the policy intervention, making difficult for policy makers to target specific sectors and to seeing benefits in the short term.

Furthermore, to the extent that stars and comets tend to locate in different places in absence of policy intervention (as suggested by their uneven distribution across MSAs), attracting stars where comets are might not be a successful policy, as stars in "comets' places" may be less productive. In other words, the same location for comets and stars would end up to be sub-optimal for (at least) one of the two categories. Therefore, interfering in the location choice of stars (or comets) in order to increase their spatial proximity to comets may lead to a much weaker effect than expected. Considering also that this kind of "attraction policy" can be quite costly, the findings are probably more on line with the skeptical arguments of economists questioning the alleged benefits of cluster policies (Duranton, 2011), rather than with the thesis of their proponents.

References

Agrawal A., I. Cockburn, A. Galasso, A. Oettl, 2012. Firm Size Diversity and Regional Innovation. Working paper.

Almeida, P., and B. Kogut, 1999. Localization of knowledge and the mobility of engineers in regional networks. *Management Science* 45.

Angrist J. D. and J-S. Pischke, 2008. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press.

Azoulay, P., J. S. Graff Zivin, and J. Wang, 2010. Superstar Extinction. *Quaterly Journal of Economics* 125:2, 549-589.

Balasubramanian N, J Sivadasan, 2011. What Happens when Firms Patent? New Evidence from US Economic Census Data, *The Review of Economics and Statistics*, vol. 93;1, pp 126-146 .

Bartik T.J., 1991. *Who Benefits from state and Local Economic Development Policies?* W.E. Upjohn Institute.

Bessen J, 2008. The value of U.S. patents by owner and patent characteristics. *Research Policy* 37:5.

Blanchard O., L. F. Katz, 1992. Regional Evolutions, *Brookings Papers on Economic Activity.* 23:1. pp 1-76.

Breschi, S. and Lissoni, F.,2001. Knowledge spillovers and local innovation systems: a critical survey. *Industrial and Corporate Change.* 10: 975–1005.

Breschi S., and F. Lissoni, 2009. Mobility of skilled workers and co-invention networks: an anatomy of localized knowledge flows.*Journal of Economic Geography* 9 pp. 439–468.

Cairncross F., 1997. *The death of distance: how the communications revolution will change our lives.* Harvard Business School Press. Boston.

Carlino G.A., S. Chatterjee, R.M. Hunt, 2007. Urban density and the rate of invention. *Journal of Urban Economics*, 61: 389-419.

Coyle, D., 1999. *The Weightless World: Strategies for Managing the Digital Economy.* MIT Press.

Duranton, G., 2011. California Dreamin': The feeble case for cluster policies. *Review of Economic Analysis*, 3(1), 3-45

Duranton, G., P. Martin, T. Mayer and F. Mayneris, 2010. *The Economics of Clusters: Lessons from the French Experience*. OUP Catalogue, Oxford University Press.

Duranton G., D. Puga, 2005. From sectoral to functional urban specialisation. *Journal of Urban Economics*, 57:2, 343-370.

Ellison, G.D., E.L. Glaeser, and W.R. Kerr, 2010. What Causes Industry Agglomeration? Evidence from Coagglomeration Patterns. *The American Economic Review*, 100:3, pp 1195-1213

Feldman MP, G. Avnimelech, 2001. Knowledge spillovers and the geography of innovation revisited: a 20 years' perspective on the field on geography of innovation. In D. B. Audretsch, O. Falck, S. Heblich,A. Lederer (eds), Handbook of Research on Innovation and Entrepreneurship, Edward Elgar.

Fons-Rosen, C., 2010. Knowledge Flows Through FDI: the Case of Privatisations in Central and Eastern Europe. Working paper.

Friedman, T., 2005. The World is Flat. New York: Farrar, Strauss and Giroux.

Glaeser, E.L., Kallal, H.D., Scheinkman, J.A., and Shleifer, A., 1992. Growth in cities. *Journal of Political Economy* 100, 1126-1152..

Greenstone M., R. Hornbeck, E. Moretti, 2010. Identifying Agglomeration Spillovers: Evidence from Winners and Losers of Large Plant Openings. *Journal of Political Economy*, 118(3).

Griffith R., L. Sokbae, J. Van Reenen, 2007. Is Distance Dying at Last? Falling Home Bias in Fixed Effects Models of Patent Citations. NBER Working Paper 13338.

Griliches Z, 1990. Patent Statistics as Economic Indicators: A Survey. *Journal of Economic Literature*, 28: 4.

Hall, B.H., A. B. Jaffe, M. Trajtenberg, 2001. The NBER Patent Citations Data File: Lessons, Insights and Methodological Tools. NBER Working Paper 8498.

Hall, B.H., A. B. Jaffe, M. Trajtenberg, 2005. Market Value and Patent Citations. *RAND Journal of Economics*, vol. 36, issue 1, pages 16-38

Haltiwanger, J. C., R. S. Jarmin, J. Miranda, Forthcoming. Who Creates Jobs? Small vs. Large vs. Young. *The Review of Economics and Statistics.*

Ioannides Y., H.G. Overman, E. Rossi-Hansberg, and K. Schmidheiny, 2008. The effect of information and communication technologies on urban structure. *Economic Policy*, 23, pp 201-242.

Kerr, W., 2008. Ethnic Scientific Communities and International Technology Diffusion. *The Review of Economics and Statistics*, 90, (3), 518-537

Jacob, J., 1969. The Economy of Cities. New York: Vintage.

Jaffe A.B., 1989. Real Effects of Academic Research. *The American Economic Review*, 79:5, pp. 957-970.

Jaffe A.B., M. Trajtenberg, 2005. *Patents, Citations, and Innovations: A Window on the Knowledge Economy.* The MIT Press.

Jaffe A.B.. M. Trajtenberg, and R. Henderson, 1993. Geographic localization of knowledge spillovers as evidenced by patent citations. *Quarterly Journal of Economics* 10, 577-598.

Jofre-Monsenya J, R. Marín-Lópeza and E. Viladecans-Marsal, 2011. The mechanisms of agglomeration: Evidence from the effect of inter-industry relations on the location of new firms. *Journal of Urban Economics*, 70:2-3, pp. 61-74

Martin, P., T. Mayer and F. Mayneris, 2011. Public support to clusters: A firm level study of French Local Productive Systems. Regional Science and Urban Economics, Elsevier, vol. 41(2), pages 108-123, March.

Moretti E., 2011, Local Labor Markets. in O. Ashenfelter, D. Card (eds.), *Handbook of Labor Economics*, vol. 5.

Oettly A., 2001. Productivity and Helpfulness: Implications of a New Taxonomy for Star Scientists. Working paper.

Peri G, 2005. Determinants of Knowledge Flows and Their Effect on Innovation.*The Review of Economics and Statistics*, vol. 87, issue 2, pages 308-322.

Silverberg G, B Verspagen, 2007. The size distribution of innovations revisited: An application of extreme value statistics to citation and value measures of patent significance. *Journal of Econometrics*, 139:2.

Quah, D., 1999. The Weightless Economy in Economic Development. CEP Discussion Paper No. 417.

Trajtenberg M., G. Shiff, R. Melamed, 2006. The "Names Game": Harnessing Inventors' Patent Data for Economic Research. NBER Working Paper No. 12479.

Waldinger F., 2010, Quality Matters: The Expulsion of Professors and the Consequences for PhD Student Outcomes in Nazi Germany. *Journal of Political Economy*, vol. 118, no. 4

Zucker, L G. and MR Darby, 2007. Star Scientists, Innovation and Regional and National Immigration NBER Working Paper Series no. 13547.

Zucker, L. G., Darby, M. R., Armstrong, J.,1998. Geographically localized knowledge:spillovers or markets? *Economic Inquiry.* 36: 65–86.

Table 1: The location of big patenting companies

Company	Nr of patents	Nr of MSAs	Nr of States	Share of the 1st MSA
GEN ELECTRIC	12892	20	14	0,49
INT BUSINESS MACHINES	12281	16	16	0,18
EASTMAN KODAK	8828	4	4	0,86
MOTOROLA	8383	7	5	0,35
AT & T	7010	10	10	0,62
E I DU PONT DE NEMOURS	5991	5	5	0,89
XEROX	5918	3	2	0,78
GEN MOTORS	5330	10	4	0,51
DOW CHEM	5197	5	5	0,49
MINNESOTA MINING & MFG	5064	3	3	0,93
MOBIL OIL	4830	4	4	0,56
TEXAS INSTR	4617	5	2	0,74
WESTINGHOUSE ELECTRIC	3663	7	7	0,46
RCA	3548	4	4	0,44
HUGHES AIRCRAFT	3377	3	2	0,91
FORD MOTOR	3135	1	1	1,00
ALLIED SIGNAL	2969	8	8	0,45
HEWLETT PACKARD	2963	7	6	0,42
UNITED TECH	2907	4	2	0,63
UNISYS	2664	11	10	0,16
EXXON RES & ENG	2599	3	3	0,71
ROCKWELL INT	2459	5	5	0,59
AMERICAN CYANAMID	2119	3	3	0,60
MONSANTO	2087	4	4	0,70
CATERPILLAR	1986	2	1	0,81

Note: the first column reports the number of patents owned by the company, the second (third) column report the number of different MSAs (US States) in which at least 100 patents have been authored by local inventors, and the fourth column reports the share of patents authored in the MSA with the largest number of authored patents.

Source: author's elaboration on NBER Patent database.

Table 2: The effect of stars on comets, level of aggregation: MSA

	(1)	(2)	(3)	(4)
Method	OLS		2SLS - F E.	2SLS - F D.
Stars (t)	0.111**	0.150**		0.079*
	(0.050)	(0.071)		(0.042)
Stars (t-1)	0.095***		0.168**	0.158**
	(0.022)		(0.079)	(0.074)
Comets (t-1)				0.223
				(0.679)
Total MSA empl.	0.243*			-0.077
	(0.132)			(0.422)
FIRST STAGE REGRESSION				
AP Stars (t)		115.2		41.83
AP Stars (t-1)			56.30	15.79
AP Comets (t-1)				1.127
Period F.E.	YES	YES	YES	YES
Observations	840	840	840	560

Note: robust standard errors, clustered at MSA level, in parentheses. The dependent variable is the number of comets.The endogenous variable are Stars (t), Stars (t-1), and Comets (t-1). The excluded instruments are IV (t) (col. 2); IV(t-1) (cols. 3); d.IV(t) and d.IV(t-1), IV(t) and Comets (t-2) (col. 4). *** $p<0.01$, ** $p<0.05$, * $p<0.1$

Table 3: The effect of stars on comets, level of aggregation: MSA-Category

	(1)	(2)	(3)	(4)
Method	OLS		2SLS - F E.	2SLS - F D.
Stars (t)	-0.001	0.080		-0.025
	(0.021)	(0.069)		(0.079)
Stars (t-1)	0.051***		0.233**	0.233***
	(0.018)		(0.097)	(0.077)
Comets (t-1)				0.471
				(0.339)
Stars oth. cats. (t)	0.089***			0.051
	(0.031)			(0.049)
Stars oth. cats. (t-1)	0.059***			-0.023
	(0.018)			(0.034)
Total MSA empl.	0.406***			-0.223
	(0.146)			(0.343)
FIRST STAGE REGRESSION				
AP Stars (t)		84.87		12.25
AP Stars (t-1)			40.16	18.97
AP Comets (t-1)				5.442
MSA*cat f.e.	YES	YES	YES	YES
Cat.*Period f.e.	YES	YES	YES	YES
Observations	7,549	7,549	7,549	4,966

Note: robust standard errors, clustered at MSA level, in parentheses. The dependent variable is the number of comets.The endogenous variable are Stars (t), Stars (t-1), and Comets (t-1). The excluded instruments are IV (t) (col. 2); IV(t-1) (cols. 3); d.IV(t) and d.IV(t-1), IV(t) and Comets (t-2) (col. 4). *** p<0.01, ** p<0.05, * p<0.1

Table 4: The effect of stars on comets, level of aggregation: MSA-Subcategory

	(1)	(2)	(3)	(4)
Method	OLS		2SLS - F E.	
Stars (t)	-0.001	0.033		0.052*
	(0.014)	(0.028)		(0.027)
Stars (t-1)	0.039***		0.120***	0.123***
	(0.011)		(0.034)	(0.034)
Comets (t-1)				0.426***
				(0.122)
Stars oth. subcats. (t)	0.058***			0.026
	(0.018)			(0.022)
Stars oth. subcats. (t-1)	0.054***			0.007
	(0.016)			(0.026)
Total MSA empl.	0.426***			-0.209
	(0.098)			(0.174)
FIRST STAGE REGRESSION				
AP Stars (t)		395.0		145.7
AP Stars (t-1)			186.6	163.2
AP Comets (t-1)				54.03
MSA*subcat f.e.	YES	YES	YES	YES
Subcat.*Period f.e	YES	YES	YES	YES
Observations	14,235	14,235	14,235	9,387

Note: robust standard errors, clustered at MSA level, in parentheses. The dependent variable is the number of comets.The endogenous variable are Stars (t), Stars (t-1), and Comets (t-1). The excluded instruments are IV (t) (col. 2); IV(t-1) (cols. 3); d.IV(t) and d.IV(t-1), IV(t) and Comets (t-2) (col. 4). *** p<0.01, ** p<0.05, * p<0.1

Table 6: The effect of stars on comets, citation-weighted stars

	(1)	(2)	(3)
Level of aggregation	MSA	MSACAT	MSASUBCAT
Stars (t)	0.100*	-0.018	0.050***
	(0.055)	(0.073)	(0.018)
Stars (t-1)	0.105*	0.268**	0.109***
	(0.054)	(0.118)	(0.034)
Comets (t-1)	0.486	0.555	0.432***
	(0.496)	(0.398)	(0.126)
Stars oth. cats. (t)		0.012	0.014
		(0.028)	(0.015)
Stars oth. cats. (t-1)		-0.014	0.013
		(0.026)	(0.014)
Total MSA empl.	-0.231	-0.329	-0.224
	(0.336)	(0.430)	(0.189)
Individual f.e.	YES	YES	YES
Cat.*Period f.e.	NO	YES	NO
Subcat.*Period f.e	NO	NO	YES
Observations	560	4,966	9,391

Note: robust standard errors, clustered at MSA level, in parentheses. The dependent variable is the number of comets.The endogenous variable are Stars (t), Stars (t-1), and Comets (t-1). The excluded instruments are d.IV(t) and d.IV(t-1), IV(t) and Comets (t-2). *** $p<0.01$, ** $p<0.05$, * $p<0.1$

Table 5: The effect of stars on comets, different thresholds

	(1)	(2)	(3)	(4)	(5)	(6)
Est. method					2SLS - F D.	
Threshold		25			75	
	MSA	MSACAT	MSASUBCAT	MSA	MSACAT	MSASUBCAT
Stars (t)	0.080**	0.018	0.047**	0.123**	0.027	0.068**
	(0.037)	(0.076)	(0.022)	(0.051)	(0.108)	(0.034)
Stars (t-1)	0.075	0.175**	0.091***	0.130*	0.261***	0.150***
	(0.083)	(0.084)	(0.028)	(0.077)	(0.087)	(0.041)
Comets (t-1)	0.277	0.637*	0.518***	0.407	0.392	0.336***
	(0.633)	(0.354)	(0.127)	(0.550)	(0.310)	(0.116)
Stars oth. cats. (t)		0.029	0.026		0.030	0.012
		(0.040)	(0.018)		(0.056)	(0.025)
Stars oth. cats. (t-1)		-0.013	0.007		-0.028	0.009
		(0.026)	(0.018)		(0.032)	(0.028)
Total MSA empl.	-0.050	-0.346	-0.300*	-0.221	-0.177	-0.131
	(0.377)	(0.362)	(0.179)	(0.343)	(0.300)	(0.166)
Individual f.e.	YES	YES	YES	YES	YES	YES
Cat.*Period f.e.	NO	YES	NO	NO	YES	NO
Subcat.*Period f.e	NO	NO	YES	NO	NO	YES
Observations	560	4,966	9,391	560	4,966	9,391

Note: robust standard errors, clustered at MSA-subcategory level, in parentheses. The dependent variable is the number of comets. The endogenous variable are Stars (t), Stars (t-1), and Comets (t-1). The excluded instruments are d.IV(t) and d.IV(t-1), IV(t) and Comets (t-2). *** p<0.01, ** p<0.05, * p<0.1

Table 7: Inventors' surname initial and patent authors' sequence

Initial	first author		second author		third author	
	Freq.	Percent	Freq.	Percent	Freq.	Percent
A	42,942	3.58	14,683	2.69	5,697	2.45
B	115,093	9.6	43,904	8.03	16,242	6.99
C	86,866	7.25	36,552	6.69	13,911	5.99
D	57,310	4.78	24,773	4.53	9,614	4.14
E	23,823	1.99	10,272	1.88	3,941	1.7
F	45,165	3.77	20,096	3.68	7,891	3.4
G	63,038	5.26	28,161	5.15	11,123	4.79
H	85,751	7.16	39,656	7.26	16,097	6.93
I	5,838	0.49	2,606	0.48	1,087	0.47
J	28,038	2.34	12,922	2.36	5,387	2.32
K	63,828	5.33	30,438	5.57	12,917	5.56
L	63,088	5.26	30,152	5.52	13,138	5.65
M	98,633	8.23	47,858	8.76	20,944	9.01
N	24,425	2.04	11,712	2.14	5,365	2.31
O	16,422	1.37	7,974	1.46	3,541	1.52
P	55,056	4.59	27,231	4.98	12,197	5.25
Q	1,854	0.15	970	0.18	386	0.17
R	55,828	4.66	26,368	4.82	12,045	5.18
S	124,636	10.4	60,864	11.14	27,666	11.9
T	37,138	3.1	18,570	3.4	8,690	3.74
U	3,582	0.3	1,769	0.32	928	0.4
V	17,480	1.46	8,525	1.56	4,342	1.87
W	63,419	5.29	30,428	5.57	14,356	6.18
X	304	0.03	247	0.05	120	0.05
Y	9,540	0.8	5,055	0.92	2,481	1.07
Z	9,282	0.77	4,735	0.87	2,297	0.99
Total	**1,198,379**	**100**	**546,521**	**100**	**232,403**	**100**

Note: the table reports the absolute and relative frequency by wich patent authors whose surname begin with the letter listed in col. 1 are reported as first, second, or third patent authors.

Source: author's elaboration on NBER Patent database.

A Data

Patent data come from the United states Patent and Trademark Office (USTPO) database as processed by the National Bureau of Economic Research (NBER), described in Hall et al., 2001. To the original dataset I add the inventors' unique identifier developed by Trajtenberg et al (2006) and the standardized assignee name available in Prof. Bronwyn H. Hall's website.[18] I am aware that the latter is not fully reliable as i) the complex ownership structure of companies may imply that differently named assignees correspond, in fact, to the same company, and ii) the same company name can be spelled in different ways (and the standardization routines cannot completely solve the problem).

I eliminate patents granted to inventors residing outside US and geolocated all the cities of residence of inventors through the ArcGis geolocator tool (based on the 2000 gazzetter of US places from US Census) and the Yahoo! Maps Ib Services. In the case where several authors are listed for the same patents and they live in different cities, the city of residence of the first author is chosen; this is a standard procedure in patent literature, and Carlino et al. (2007) show that the approximation is substantially innocuous. The geocoding operation was successful for 1,161,650 patents, which correspond to 97% of the database. I then assigned cities to counties using the ArcGis spatial join tool, and subsequently counties into MSAs (1993 definition). To my knowledge, this is the first time that patent data are geocoded (almost) entirely, without ignoring small counties.

B The spatial distribution of stars and comets

In order to explore the location pattern of stars and comets across US MSAs, I set up a simple regression for periods 3-4-5 based on the following equations:

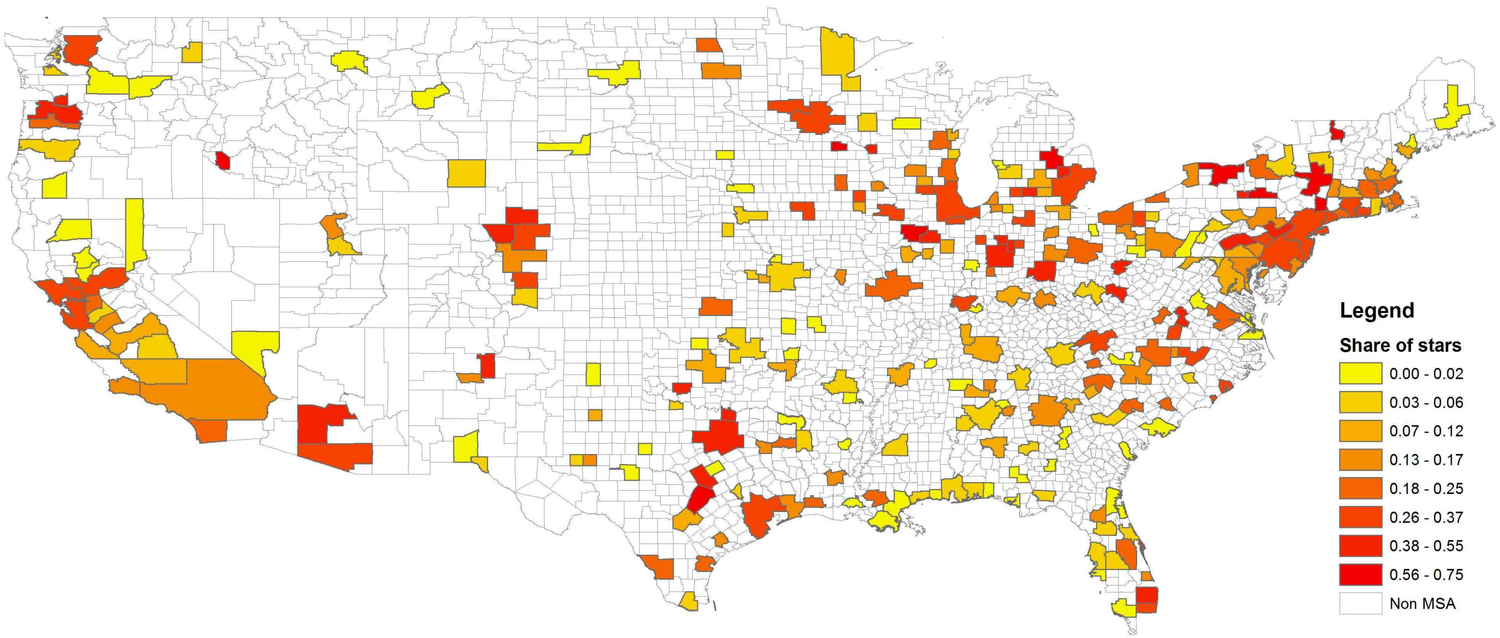
$$Share(Comets)_{it} = \beta_1 X_{it} + \delta_t + \epsilon_{it} \tag{3}$$

$$Share(Stars)_{it} = \beta_2 X_{it} + \delta_t + \epsilon_{it} \tag{4}$$

where i indexes MSAs and t periods, X_{it} is a matrix of MSA-specific coovariates, β_1 and β_2 are vectors of coefficients, and δ_t is a time fixed effect. These regressions are purely descriptive: they produce some partial correlations which are useful to assess whether stars and comets show two distinctive location patterns, depending on a few observable characteristics of cities. The variables included in X are a list of proxies of the economic structure of the MSA: total

[18] http://elsa.berkeley.edu/~bhhall/

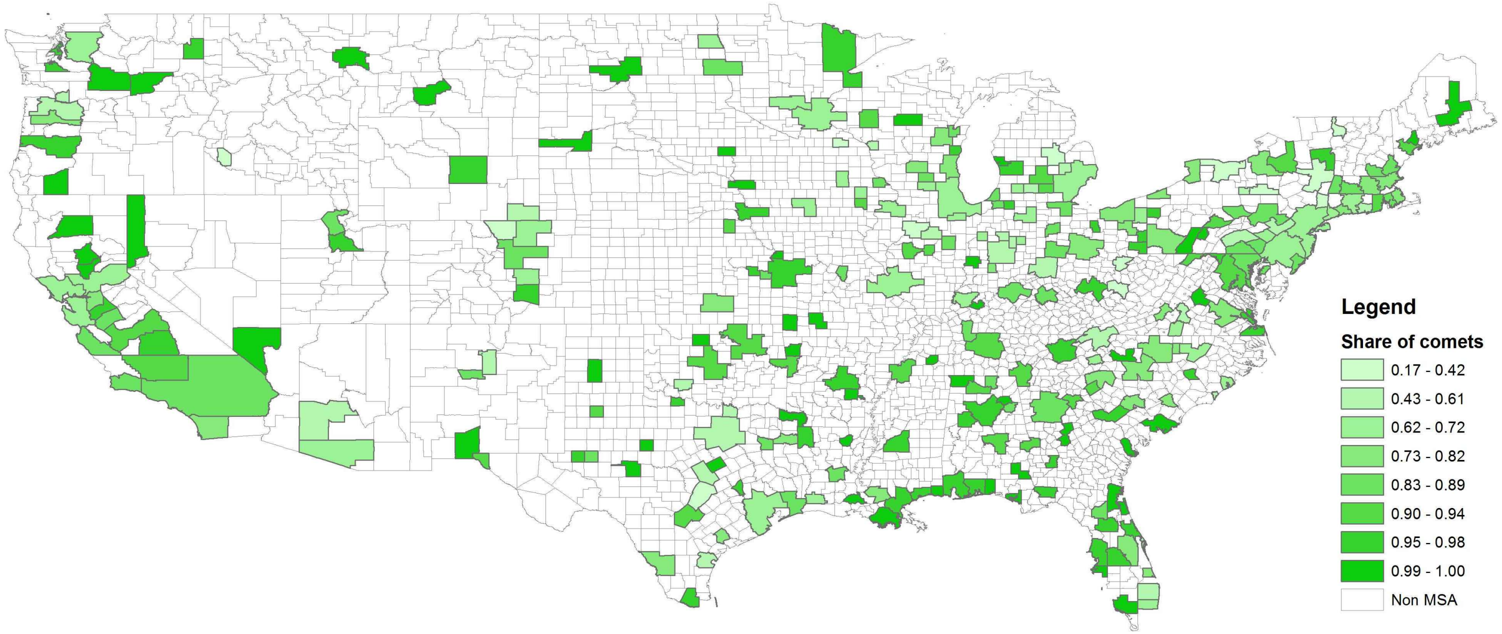
Figure 1: Share of star patents by MSA, period 5 (1994-97)



Note: the map shows the share of star patents in the total patents by MSAs in the period 1994-1997.

Source: author's elaboration on NBER Patent database and National Atlas digital maps.

Figure 2: Share of comet patents by MSA, period 5 (1994-97)



Note: the map shows the share of comet patents in the total patents by MSAs in the period 1994-1997.

Source: author's elaboration on NBER Patent database and National Atlas digital maps.

employment (*totemp*), a proxy of labor productivity (total payroll over employment, *labprod*), the share of employment in manufacturing (*manuf. share*), Herfindahl diversity index (*Herfindahl*, calculated as the sum of the squares of the share over the total of employment of 2-digit SIC sectors), and the number of plants with less then 500 employees (*n. plants <500 emp.*). I also include the number of total patents in the MSA in order to control for the size of the patenting sector in the city. The sample is restricted to the last three periods and comprehends all the MSAs. The structural variables refer to the first year of the time period, while the patent variables are summed over the whole period. All the variable are expressed in logarithmic form.

The results - reported in table 2 - show that comet patents are negatively associated with the total number of patents and positively with the total number of firms. Conversely, star patents are positively associated with both the number of patents and the proxy for labour productivity, suggesting that star patents are more frequently located in cities with a large number of patents and a more skilled workforce. The other explanatory variables are not significant.

Table 8: Regression of comets/stars shares at MSA level

	(1)	(2)	(3)	(4)
VARIABLES	Comets (share)	Stars (share)	Comets (share)	Stars (share)
Tot. emp.	-0.110***	0.444***	0.083**	-0.032
	(0.034)	(0.140)	(0.039)	(0.156)
Prod	-0.079*	0.335**	0.046	0.026
	(0.042)	(0.143)	(0.040)	(0.146)
Herfindahl	-0.042	-0.253	-0.079	-0.171
	(0.062)	(0.253)	(0.058)	(0.259)
N. plant <500 emp.	0.068*	-0.188	0.088**	-0.244
	(0.037)	(0.154)	(0.037)	(0.149)
Manuf. share	-0.020	0.153	-0.011	0.133
	(0.034)	(0.096)	(0.033)	(0.096)
Tot. MSA pat.			-0.186***	0.467***
			(0.028)	(0.074)
Period dummies	YES	YES	YES	YES
Constant	0.100	-4.661***	-1.333***	-1.545
	(0.469)	(1.774)	(0.480)	(1.767)
Observations	691	691	691	691
Number of MSAs	256	256	256	256

Note: robust standard errors, clustered at MSA level, in parentheses. *** p<0.01, ** p<0.05, * p<0.1