Stars and Comets: an Exploration on the Patents' Universe

Carlo Menon*
Department of Geography and Environment
London School of Economics
Email: c.menon@lse.ac.uk

This version: 20 May 2009 - **WORKING DRAFT**

Abstract

The analysis of patents' and citations' data has become a privileged source of evidence on localized knowledge spillovers and innovation. Nevertheless, one aspect has been overlooked: the patents' distribution across inventors is extremely skewed, as many inventors register one or a few patents, while a small number of inventors register many patents. To our knowledge, the previous empirical literature has not discussed the different 'scales' of local innovation patents may originate from. A first contribution of this paper is therefore to discuss the issue. A second contribution is to investigate whether patents originating from different scales of innovation are located in different cities. A third contribution - which constitutes the main scope of the paper - is to test whether the concentration of the activity of star inventors is beneficial to the local productivity of other kinds of innovation - namely the ones led by more occasional, and less prolific, inventors.

*Riccardo Crescenzi, Steve Gibbons, Henry Overman, Olmo Silva, and partecipants to the First SERC Conference in London provided extremely useful comments.

1 Introduction

The analysis of patents' and citations' data has become a privileged source of evidence on localized knowledge spillovers and innovation. Nevertheless, one aspect has been generally overlooked: the patents' distribution across inventors is extremely skewed, as many inventors register one or a few patents, while a small number of inventors register many patents. Innovations developed by inventors at the opposite extremes of the distribution are unlikely to be the outcome of an homogenous innovation "black box". Interestingly, this peculiar characteristic of the patenting activity recalls the more general 'innovation paradox' recently highlighted in the literature (e.g. Acs and Audretsch, 1990). While big companies massively invest in formal R&D activities, many new products and processes are generated by small and medium enterprises, with little or no reported investments in R&D. The latter kind of innovation process is therefore more likely to be based on learning-by-doing and incremental innovation, being thus intrinsically different from the activity of 'professional scientists'.

To our knowledge, all of the previous empirical literature on 'local innovation' based on patent data has not discussed the different scales of innovation patents may originate from. A first contribution of this paper is therefore to discuss the issue. A second contribution is to investigate whether patents originating from different scales of innovation are located in different cities. A third contribution - which constitutes the main scope of the paper - is to test whether the concentration of the activity of star inventors is beneficial to the local productivity of other kinds of innovation - namely the ones led by more occasional, and less prolific, inventors.

In order to achieve that, using the USPTO/NBER database we identify two exemplificative categories of inventors situated in the two tails of the distribution: we define as *stars* those inventors who are highly productive in a given period of time - while we define as *comets* those inventors that develop only one or a few patents. A preliminary data inspection at MSA level shows how the association with establishment births and other MSAs' structural characteristics and number of patents is significantly different for the two patents' categories. This confirms that i) the categorization is not trivial, ii) the two categories relate to different innovation processes, and iii) stars and comets are concentrated in different cities, especially after controlling for the general distribution of the patenting activity.

The location of investments of big companies is increasingly influenced, directly or indirectly, by local policy makers: the attraction of 'million dollar plants' is seen as a successful policy targeted at increasing the productivity of incumbent (small) firms through technological spillovers (Greenstone et al, 2008). Similarly, local policy makers may be keen in attracting R&D labs of big companies whithin their jurisdiction. We, however, find some evidence suggesting that the direct impact of stars on the local economy is negligible. However, the lack of direct effects might be compensated by indirect effect operating through an increase of the activity of comet inventors, which in turn may justify the

provision of public money to place-marketing policies.

Therefore, in the second part of our empirical analysis we assess whether the activity of star inventors is beneficial to the production of comet patents, and try to quantify this effect. More specifically, using the NBER/USPTO rich patents' database we 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. We exploit the panel dimension of our dataset to account for various fixed effects, and adopt an instrumental variable approach to avoid a potential endogeneity bias.

2 Patents and the Geography of Innovation

Patents' 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 an economic activity which is generally very hard to measure, i.e., innovation. Furthermore, the availability of citation linkages has added even more interest on patents data: for the first time, researchers had a tool to 'trace' knowledge spillovers, which previously had been considered as one of the most intangible concepts in economic theory. A popular book by Jaffe et al (2005), and the free availability of the USTPO dataset from the NBER website, further contributed to multiply the empirical applications based on patents' data.

A significant part of this wide literature has focused on the geographic component of innovation, with a particular interest on the spatial decay of knowledge spillovers. A seminal contribution by Jaffe et al (1993) showed as a cited patent is twice more likely to be in the same US metropolitan area as the citing one, compared to a couple of technologically similar patents with no citation links. On the same line, Peri (2005) examined 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". Carlino et al (2007) used patents data for a cross-section of US metropolitan areas to investigate the relationship between urban density and innovation intensity (as measured by patents pro-capita) finding a positive and robust association, with the caveat, however, that many omitted variables might, in facts, explain the positive correlation.[1] All these contributions (and many similar which we omit for brevity) highlight that knowledge spillover have a geographically limited distance decay.

Previous contributions, however, did not take into consideration an important feature of patents' data, i.e., the skewness of the distribution of patents

[1] The authors include a robustness test based on IV estimation, but, in our opinion, all the instruments are arguably exogenous, as they may affect patenting through e.g. productivity. Also, it is not very clear how sorting of very productive inventors and companies into denser cities influences their results.

across inventors.[2] This is in part due to the fact that until very recently an unique identifier for inventors was not available in the NBER/USPTO database and therefore calculating the distribution of patents by inventors was unfeasible. Thanks to the efforts of Trajtenberg et al (2006), who 'estimated' the unique id using an ad-hoc algorithm, we know that out of 1,600,000 inventors listed in the NBER dataset in the period 1975-99, 60% of them registered just one patent, the 30% from 2 to 5, and only the 0,15% (2,402 inventors) more then 50.

The peculiar distribution of patents by inventors reveals that the innovation process which patenting is a proxy for is an extremely composite phenomenon. However, The fact that innovation produced by small and big companies is intrinsically different is well known and debated in the literature. In particular, robust evidence on two distinct aspects of small firm innovation are posing a challenging 'innovation paradox'. First, small firms have a much higher ratio of patents developed to R&D expenditures (Griliches,1990) than big companies. If we substitute patents with innovations introduced to the market and R&D with employment, the result is equivalent: the ratio is much higher for small firms (Acs and Audretsch, 1990) . The authors argue that this can be due to the higher permeability of small companies to local public R&D inputs (i.e., university research) (Acs et al, 1992). An alternative explanation could be that small companies rely on alternative innovation inputs, based on lerning-by-doing and incremental innovation, rather than formal scientific research. Second, small firm innovation is not at all a residual phenomenon, accounting for most of the innovative activity in many sectors (Acs and Audretsch, 1990). In passing, it is also worth mentioning that small firms account for most of the employment growth in the US in the last decades (Audretsch, 2002). Furthermore, Balasubramanian and Sivadasan (2008) in a recent working paper link patents' data with Census firm data for US, being able to assess the impact of patents on firms' performance. They focus in particular on firms that patent for the first time, and find a significant and large effect on firm growth (but, interestingly, little change in factor productivity). To the extent that first-time patenting firms are much more likely to belong to the group of 'occasional' inventors, we deduct that this kind of innovative activity bears strong economic importance. Partly contradicting the previous findings, the analysis of patents' value as inferred by their renewal shows that "small entities – individuals, corporations with fewer than 500 employees and non-profit organizations – have patent values that are on average less than half as large as the values obtained by large corporations" (Bessen, 2008, p. 944).

To sum up, the main conclusion we draw from the mentioned literature is the following: patenting activity - as a proxy of a more general innovation activity - of comets and stars are two distinct phenomena, and this distinction need to be carefully considered when examining patents' data.

[2] Among the closest contribution we could find, we mention: Silverberg and Verspagen (2007), who analysed in depth the skewness of the distribution of citations across patents; Zucker and Darby (2007) looked at the linkages with private companies of a small sample of star inventors;

3 Stars and Comets

Our analysis is based on the NBER/USPTO database, which lists all the patents granted in US from the 1969 to 1999. We added to this dataset the inventors' unique ID developed by Trajtenberg et al (2006). As the latter is available only since the 1975, our period of analysis is restricted accordingly. More details on the data are reported in Appendix A.

We identify five time periods of four years each, and five "observational windows" of eight years for each of them - they are listed in table 1. We use the four-years periods, instead of single year intervals, to take into account the imprecision of our temporal dimension, due the following characteristics of the data. First, we use the year in the patent is granted, which is generally 2-3 later the year of application. Furthermore, we do not know how long the inventor had been working on a patent before applying for it. Equally difficult is to time when local knowledge spillovers may have effect - it could be while the source and destination inventors are both working on their respective patents or after the star has applied (or has been granted) for it. By inspecting the data we found that the median and mean value of the citation lag of patents in the same MSA is four years, and we therefore choose to adopt periods of the same length.[3] This seems a reasonable choice in order to average out some of the measurement error.

However, the definition of comet and star patents in a given period is based on the characteristics of their inventors in the relative observational windows, rather then the period. Observational windows are longer than periods (and overlapping) to allow for some flexibility in the definition of the two categories: a star may be an highly productive inventor even a few years before getting the patents granted, or, conversely, they can still produce knowledge spillovers a few years after the last patent has been granted. Similarly, the longer observational window is important for comets as well, as it avoids including in the category inventors who do not satisfy the requirements in the years immediately before, or after, the given period.

Table 1: Period classification

Period	Years	Obs Windows
1	1978-1981	1976-1983
2	1982-1985	1980-1987
3	1986-1989	1984-1991
4	1990-1993	1988-1995
5	1994-1997	1992-1999

In each period, a patent is defined as the outcome of a "star inventor" if its first author has developed five other patents or more (as first author) in the

[3] We restricted the calculation to patents 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.

relative observational window.[4] Such patent is therefore defined as a star patent. Similarly, I define "comet inventors" patents' (first) authors which developed less than three patents in the observational window, and less then six till that point in time; the patents they develop are defined as comets. As a further restriction, comets must not have as assignee a company which is assignee of 50 patents or more in the whole dataset. Thus, Individual inventors listed as stars cannot become comets in a following period, while a comet can potentially become a star; this, however, happens for only 1% of comet inventors listed in the dataset.

It is important to note that the patent, rather than the inventor, is the final unit of observation. This choice depends on the fact that inventors are observed only when they patent, which makes extremely hard to set up a meaningful analysis at the inventor level.

The analysis is generally limited to the last three periods, as MSA controls are unavailable for period 1 and 2. We defined the latter, however, as they are used to build the instrumental variables and in regressions which includes time lags of the patents variables.

Star patents account for the 26% of the total patents granted in the period 1986-1997, while the corresponding share of comet patents is equal to 11%. On the inventors' side, among all the unique inventors listed in the five periods (534,120), around 5% of them are listed as stars at list once, while for comets the same share is equal to 15%. Looking at single periods, star inventors are the 7-9% of the total, while comets are the 14-16%. The 'star' status appears to be quite persistent across time: around the 40% of stars in given period where stars also in the previous period. The share goes down to 15% with a two periods lag.

Interesting facts emerge also from the analysis of citations' data. Table 2 reports the shares of citations originating from stars, comets, and other patents (rows) directed to stars and patents (columns). Compared to patents that are neither comets nor stars (third row), comets (first row) are more likely to cite comets, and less likely to cite stars. The opposite is true for stars: they are more likely to cite stars, and less likely to cite comets. The pattern is similar also when looking at citations within technological categories (not showed). We interpret it as a further evidence that the stars/comets categorization, although stylized and partially arbitrary, do identify different groups of patents, supposedly related to different innovation processes. On the other hand, we notice that comets do cite stars, although at a smaller rate than other patents. The hypothesis that comets may benefit from knowledge spillovers from stars is thus not rejected by citations data.

Citations may also be useful to inspect the average 'value' of different categories of patents. Although quite debated and 'noisy', the association of number of received citations with the market value of the patents has been convincingly argued (Hall et al, 2001). We use citations' data to explore whether patents and comets significantly differ from other patents in this dimension, by regress-

[4] The threshold has been chosen as it approximately limit the upper quartile of the inventors' distribution in term of patents pro-capita.

Table 2: citations' shares, comets and stars

Citing/Cited	Comets	Stars
comets	16.2	16.8
stars	7.5	34.7
other patents	9.8	24.2

Table 3: Regression of citations received

Dep. var.	Citations received
Star patent	0.129***
	(0.00331)
Comet patent	0.0123***
	(0.00422)
Citations made	0.0152***
	(0.000184)
Constant	1.822***
	(0.00727)
Observations	902,105

Robust standard errors in parentheses

ing the number of received citations on 'comet' and 'star' dummies, over the whole sample of patents in period 3-4-5. We also include time and technological category dummies, their interactions, and a variable reporting the number of citations made to control for the heterogeneous propensity to cite among different groups of patents (in a finer way than it could be absorbed by category and time dummies). Given that the dependent variable is a overdispersed count, the model is a negative binomial. Coefficients, reported in table 3, show that star patents are indeed significantly more cited than other patents. The effect is positive also for comet patents, but the size of the coefficient is much smaller.

3.1 Preliminary evidence on location of stars and comets

In this section, we present some descriptive statistics which i) show how stars and comets are located in different places, and ii) substantiate the validity of stars and comets as good proxies for the output of different innovation processes.

If we look at the distribution of comet, star, and other patents over total employment across MSAs,[5] we find a strong correlation (Table 4), which implies

[5] Counties are grouped into MSAs according to the 1993 definition, based on 1990 Census data. Counties not included into MSAs are also individually included in the sample. The analysis, therefore, covers the whole US territory.

that innovative activity is overall spatially concentrated. When plotting the shares of comets and stars on the total patents, however, we see that there is a fair degree of dispersion in both the distributions, remarkably driven by a long right tail (Figure 1, 2).

Table 4: Patents by MSAs over total employment, rank correlation

	comets	stars	other patents
comets	1	0.42	0.59
stars	0.42	1	0.61
other patents	0.59	0.61	1

We can go further by looking at patterns of partial correlation with MSAs structural characteristics, setting up a simple panel regression for periods 3-4-5 based on the following equations:

$$Share(Comets)_{it} = \beta_1 X_{it} + \delta_t + \epsilon_{it} \tag{1}$$

$$Share(Stars)_{it} = \beta_2 X_{it} + \delta_t + \epsilon_{it} \tag{2}$$

where i indexes MSAs and t times, X is a matrix of MSA-specific coovariates, $\beta_1$and β_2 are vectors of coefficients, and δ is a time fixed effect. The variables included in X are the (log of) the total patents in the MSA which are neither stars or comets, the (log of) total MSA employment, the share of employment in manufacturing, the Herfindahl diversity index of 3-digit SIC sectors, and the number of plants with less than 500 employees. The equation are estimated with an OLS regression on the pooled samples, with standard errors clustered at MSA level. The results - reported in table 5 - clearly show how the two vectors of coefficient are different (as confirmed by the Hausman test). In particular, comet patents are positively (partially) associated with the number of small firms, while the total number of other patents and the Herfindahl index have a negative coefficient (the latter thus meaning that a more diversified city is associated with more comets). Conversely, star patents are positively associated with both the number of other patents and the Herfindahl index, suggesting that star patents are more commonly located in specialized cities.

Our interpretation of these results is the following: comet patents are associated to more general innovation activities, and therefore are more likely to be located in innovative hotpots with a diversified economy and many small firms; in such cities the pool of patents is not necessarily large, as innovations may be introduced to the market in other forms. On the other hand, the activity of stars is more strongly associated with formal R&D and patenting, thus is more likely to be located where the pool of patents is large, and the structure of the local economy is dominated by big companies.

Table 5: Regression of comets/stars shares at MSA level

VARIABLES	(1) comets (share)	(2) stars (share)
Oth. pat.	-0.0412***	0.0237***
	(0.00361)	(0.00821)
Tot. emp.	0.00291	-0.00365
	(0.00595)	(0.0119)
Herfindahl	-0.284**	0.677**
	(0.127)	(0.335)
Manuf. share	0.0573	0.0694
	(0.0448)	(0.0902)
N. plant <500 emp.	0.0351***	-0.00776
	(0.00625)	(0.0135)
	(0.0140)	(0.0320)
Observations	1289	1289
R^2	0.230	0.039

Period dummies included in all the specifications
Robust standard errors in parentheses

We also look at the association with establishment births, by regressing the latter variable on the number of star and comet patents developed in the same MSA, plus some other controls (total employment, Herdindahl index, and average establishment employment - all lagged by one period to avoid simultaneity bias), for period 4 and 5 (period 3 is dropped due to data restrictions). Again, the results (table 6) show a differentiated pattern for stars and comets: while comets have a significant effect, comparable to the effect of other patents, star patents have a negative coefficient, which become insignificant once the MSA-specific controls are included.

We do not claim causality at this stage - many variables are potentially omitted and we can't exclude a reverse causality bias - but, nevertheless, the associations we have analysed support two main findings: first, once controlling for the general distribution of patenting activities, comet and star patents are developed in different places; second, star patents seem to have a much weaker connection with the local economy than comet patents. To the extent that the former are developed in R&D labs of big companies, while the former are the by-product of the innovative activity of small firms, the finding is not surprising.

3.2 Why should stars positively affect comets?

Even though we assume comet and star patents are the outcome of substantially different innovation processes, this does not necessarely exclude that their

Table 6: Regression of establishment births at MSA level

VARIABLES	(1) Tot. births	(2) Tot. births
Total comets	0.304***	0.121***
	(0.0620)	(0.0320)
Total stars	-0.119***	-0.0193
	(0.0374)	(0.0200)
Total oth. patents	0.487***	0.151***
	(0.0716)	(0.0404)
Herfindahl Index t-1		2.556
		(2.024)
Tot. emp. t-1		0.601***
		(0.0263)
Av. emp. t-1		-1.457***
		(0.104)
Observations	418	418
R^2	0.707	0.907

Period dummies included in all the specifications
Robust standard errors in parentheses

activity may have some point of contacts. Stars may be positively affecting the production of local patent comets through four main mechanisms:

a) Informal knowledge spillovers: star inventors and comet inventors develop informal contacts due to proximity, which in turn facilitate the activity of the latter (e.g., they may obtain hints on their work).

b) Formal knowledge spillovers: star inventors may transfer their expertise to comet inventors in more formal ways, e.g. in occasion of seminars, conferences, and the like.

c) Workplace contacts: (future) comet inventors may have the opportunity to work in an institution where stars are employed, without necessarely becoming stars themselves (they may be employed in different mansions, or they may leave the institution at an early stage of their career).

d) Display/attraction effect: the presence of many labs of big companies may attract comets to a locality, as innovative entrepreneurs may assume that they can enjoy the effects of point a, b, c.

Interestingly, only the first point Unfortunately, the data do not allow us to disentangle the different effects.

4 Analysis

In the present section we investigate whether the activity of star inventors is beneficial to comet inventors, and try to quantify this effect. We therefore estimate the following model:

$$Comets_{ikt} = \beta \cdot Stars_{ikt} + \gamma X_{it} + \delta_k + \tau_t + \phi_i + \delta\tau_{kt} + \varepsilon_{ikt} \qquad (3)$$

where i, k, and t index MSAs, categories, and periods, respectively; Stars and Comets are the number of patents in the respective category, X is a set of MSA time variant controls, and δ, τ, ϕ are category, time, and MSA fixed effects. The six technological categories are the following: Chemical (excluding Drugs); Computers and Communications (C&C); Drugs and Medical (D&M); Electrical and Electronics (E&E); Mechanical; and Others.

The unit of observation is the MSA-category pair; the choice is motivated by the assumption that patenting in each of the six different technological categories is hardly linked to the activity in other categories. This seems to be confirmed by citation data, which show that 80% of citation linkages are bounded within the same category. The analysis is limited to periods 3-4-5, as MSA controls are not available for previous periods, and sample is restricted to the MSA-category pairs in which at least 25 patents have been granted in the given period.

We opt for a log-linear specification because the dependent variable is an extended count variable (with a long right tail and skewed to the left), which approximates the normal distribution after the log transformation. The side effect of the log transformation is the loss of the zeros, but they are less then 5% of observations. In the following section, we perform some robustness tests based on a Poisson model with the natural count variable and find compatible results.

The MSA-specific time-variant variables included in the matrix X are the following:

i) log of total employment (*totemp*), meant to control for agglomeration economies and size effects

ii) the share of employment in manufacturing (*manuf. share*), in order to assess whether comets are associated with specialization in manufacturing

iii) the Herfindahl diversity index (*Herfindahl,* calculated as the sum of the squares of the share over the total of employment of 2-digit SIC sectors), as a proxy of the diversity of the economic structure. This variable can have two opposite effects: on one side, the literature has emphasized the positive effect of diversity on innovation due to Jacobian externalities (e.g., Glaeser et al., 1992; Duranton and Puga, 2005). On the other side, we may suppose that MAR externalities, rather than Jacobian, are more beneficial for the kind of incremental innovation which underlie the development of comets.

iv) log of the number of plants with less then 500 employees (*n. plants <500 emp.*) as these are defined as 'small plants' in the US.

We anticipate, however, that MSA controls are hardly significant in our regressions. This is due to the inclusion of the MSAs' fixed effects, which absorb most of the effect of variables with small variations across time; and to the grouping of patents at the technological category level, which makes the

association with MSA variables weaker.

In some specifications we also include a variable reporting the log of total number of patents developed in the other five technological categories (*nr. pat. oth. cats.*), and the log of total number of patents in the same MSA/category which are neither stars or comets (*nr. oth. pat. cat.*). Finally, we include a number of fixed effects, controlling for technological category and MSA time invariant factors, for time-specific shocks, and for technological category shocks.

4.1 Instrumental Variable Estimation

Estimates of equation 3 can be inconsistent due to reverse causality or omitted variable biases, which cast doubts on the exogeneity of the main variable of interest (the number of star patents). We therefore create two different instrumental variables to deal with the issue.

4.1.1 First instrument

The first instrument is calculated through the following steps:

a) For each period, we calculate the total number of star inventors active in a given MSA and technological subcategory (patents are classified into 6 categories and 36 subcategories). If an inventor developed patents classified into different subcategories, he/she is assigned corresponding weights summing to one, accordingly to the subcategories' shares. If they have been recorded as resident in several MSAs, the modal one is chosen.

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

c) For each MSA, each period, and each subcategory, we multiply the number of inventors in period *n-2* at point a) by the productivity in the respective technological subcategories in period n calculated in b). Subsequently, we sum the outcome by MSA, period, and technological category. The result is the instrumental variable for total number of star patents in period 3-4-5, by MSA and category.

Formally, it can be expressed with the following equation:

$$IV1_{ikt} = \Sigma_s(StarsInv_{ikst-2} \cdot AvPat_{kst}) \tag{4}$$

where i indexes MSAs, t periods, k technological categories, and s technological subcategories within the category k. The first element of the product is calculated at point a), and the second one at point b).

The validity of the IV relies on an assumption of excludability for point a), i.e., once MSA fixed effects are controlled for, the number of star inventors living in a given MSA in period *n-2* (on average ten years before) has no independent

effect on the number of comet patents developed in period n; and on an assumption of exogeneity for b), i.e., the average productivity in the whole US is exogenous to MSA-specific unobserved factors. At the same time, I expect the productivity of stars inventors working in the same subcategories to be correlated, due to sharing a similar competition pressure, regulatory framework, market demand, etc. For example, we assume that the number of star patents developed in the MSA of New York in the year 1994-97 entails an exogenous component due to the interaction of a) the historical presence of R&D labs in semiconductor devices, and b) the US-wide growth in (patent) productivity of the semiconductor devices sector, relatively to other sectors.

The IV strategy is close in spirit to the approach of Bartik (1991) and Blanchard and Katz (1992), among others, who instrumented regional economic growth interacting the lagged sectoral structure of a region with the contemporaneous national sectoral trend. There is, however, a reason of concern about the exogeneity assumption for point b). To the extent that comets in a given MSA are specialized in the same subcategories of stars, the US-wide variation in productivity in a subcategory can be correlated with the error term of equation 3. This in turn will compromise the validity of the instrument. We therefore build a second IV in order to improve the robustness of our estimate.

4.1.2 Second instrument

The second instrument follows a methodology similar to the first one, but the technological subcategories are substituted with the assignees of the patents. The steps are the following:

a) For each period, we calculate the total number of star inventors active in a given MSA and with a given assignee. In 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, we 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, we multiply the number of inventors in period *t-2* at point a) by the average number of patents produced by star inventors sharing the same assignee in period t calculated in b). Subsequently, we 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:

$$IV2_{ikt} = \Sigma_a(StarsInv_{ikat-2} \cdot AvPat_{at}) \tag{5}$$

where i indexes MSAs, t periods, k technological categories, and a the assignees. Again, the first element of the product is calculated at point a), and the second one at point b).

The excludability condition is identical to the one for the first instrument, while the exogeneity assumption is similar: given that stars and comets generally have different assignees (the assignee is very often the employer of the inventor, and comets have, by definition, assignees which less than 50 patents assigned in total - while, on average, assignees of stars have 4010 assigned patents) we assume that the average productivity of an assignee in the whole US (excluding the given MSA) has no independent effect on the productivity of comets of that MSA.

5 Results

Results from the OLS estimation are reported in table 7. All the variables are reported in logs (except from the Herfindahl index and the share of manufacturing employment, which are strictly smaller than one), thus the coefficients can be interpreted as elasticities.

The effect of star patents on comets is always positive and significant, but overall quite small: the coefficient ranges from 0.04 to 0.14. Among the other controls, the total number of patents in the MSA/category that are neither stars or comets has a positive sign, as expected, and its inclusion reduces significantly the stars coefficient. The total number of patents in other categories has, instead, an unexpected negative coefficient, while total employment has a large positive effect, although barely robust. The other coefficients are not significant and this probably depends on the fact that they all are grouped at MSA level, which in turn confirms the convenience of conducting the analysis at category level.

Results from the IV regressions are reported in table **??**. The two instruments are both strong, as evidenced by the first stage regression (table 8), and lead to extremely similar results. Instrumented coefficients are still positive and significant, and are significantly bigger than OLS estimates: now the elasticity of comet to star patents ranges between 0.33 and 0.40. We explain the downward bias of the OLS as originating from negative selection, rather than from reverse causality: from table 5, we know that patents and stars tend to concentrate in cities with different structural characteristics. To the extent that some of the MSA-specific variables are unobserved and correlated with the number of star and comet patents with opposite sign, OLS estimates are downward biased.

A further concern about the specification involves the potential endogeneity of the variable reporting the number of other patents in the same category (*oth. pat. same cat.*). The best option to deal with it would be to instrument that variable as well, but unfortunately a further instrument in unavailable. However, we can see that its inclusion to the IV specifications leaves the main coefficient almost unaffected, as we would expect in a valid IV specification (robust to omitted variable bias).

Table 7: regression of comet patents, OLS

VARIABLES	(1) comets	(2) comets	(3) comets	(4) comets
stars	0.142***	0.0404*	0.0378*	0.121***
	(0.0188)	(0.0207)	(0.0206)	(0.0180)
Oth. pat. same cat.		0.304***	0.282***	
		(0.0403)	(0.0409)	
Pat. other cats.			-0.140*	-0.304***
			(0.0746)	(0.0772)
herfindahl			2.194	1.414
			(3.052)	(3.153)
Plants <500 emp.			0.0341	0.0734
			(0.183)	(0.188)
Manuf. share			0.210	0.0712
			(0.559)	(0.567)
Tot. emp.			0.352	0.554**
			(0.246)	(0.247)
Observations	2113	2113	2113	2113
R^2	0.857	0.865	0.865	0.860

Period, MSA, and tech. cat. dummies included in all specifications
Robust standard errors in parentheses

Table 8: IV - first stage

VARIABLES	(1) stars	(2) stars	(3) stars	(4) stars
IV1	0.242***	0.0811***		
	(0.0188)	(0.0161)		
IV2			0.362***	0.130***
			(0.0248)	(0.0225)
Oth pat. same cat.		0.895***		0.868***
		(0.0471)		(0.0488)
Pat. oth. cats.		-0.175*		-0.163*
		(0.0965)		(0.0963)
herfindahl		-3.508		-2.745
		(4.284)		(4.245)
Plants <500 emp.		-0.0900		-0.0157
		(0.244)		(0.243)
Manuf. share		1.818**		1.650*
		(0.870)		(0.860)
Tot. emp.		-0.305		-0.377
		(0.323)		(0.320)
Observations	2113	2113	2113	2113
R^2	0.76	0.82	0.77	0.82
F-stat	24.70	50.49	27.87	51.58

Period, MSA, and tech. cat. dummies included in all specifications
Robust standard errors in parentheses

Table 9: regression of comet patents, first IV

VARIABLES	(1) comets	(2) comets	(3) comets	(4) comets
stars	0.354***	0.404***	0.373**	0.339***
	(0.0463)	(0.143)	(0.149)	(0.0544)
Oth. pat. same cat.		-0.0703	-0.0499	
		(0.148)	(0.149)	
Pat. other cats.			-0.0576	-0.0510
			(0.0882)	(0.0950)
herfindahl			3.079	3.070
			(3.283)	(3.216)
Plants <500 emp.			0.0550	0.0501
			(0.194)	(0.190)
Manuf. share			-0.412	-0.350
			(0.654)	(0.585)
Tot. emp.			0.422*	0.400
			(0.250)	(0.251)
Observations	2113	2113	2113	2113

Period, MSA, and tech. cat. dummies included in all specifications
Robust standard errors in parentheses

Table 10: regression of comet patens, second IV

VARIABLES	(1) comets	(2) comets	(3) comets	(4) comets
stars	0.350***	0.388***	0.362***	0.335***
	(0.0421)	(0.121)	(0.125)	(0.0482)
Oth. pat. same cat.		-0.0537	-0.0389	
		(0.125)	(0.126)	
Pat. other cats.			-0.0603	-0.0551
			(0.0852)	(0.0899)
herfindahl			3.050	3.044
			(3.247)	(3.197)
Plants <500 emp.			0.0543	0.0505
			(0.193)	(0.189)
Manuf. share			-0.391	-0.343
			(0.622)	(0.578)
Tot. emp.			0.420*	0.403
			(0.249)	(0.249)
Observations	2113	2113	2113	2113

Period, MSA, and tech. cat. dummies included in all specifications
Robust standard errors in parentheses

5.0.3 Robustness tests

We run a series of robustness tests to corroborate the validity of our estimates. In 11, we report the estimates of 3 applying a Poisson model to different selection of the sample: the OLS one, the OLS one plus the observation with zero comets and stars, and the whole sample. In the specifications with the larger samples we had to exclude the variable measuring the number of other patents in the same category (*oth. pat. cat.*), as the maximum likelihood function could not be maximized otherwise.

Results show that the coefficient of star patents is always positive and highly significant, and substantially unaffected by the different sample selections. Furthermore, its size is comparable with the OLS one. We therefore exclude sample selection biases in our OLS estimations, due to either the exclusion of zeros or the threshold of 25 patents.

Table 11: Poisson regressions

	(1)	(2)	(3)	(4)	(5)
VARIABLES	comets	comets	comets	comets	comets
Sample	OLS	OLS + zeros	OLS	OLS + zeros	All
stars	1.00025*	1.00026*	1.00072***	1.00072***	1.00073***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Oth. pat. same cat.	1.00047***	1.00046***			
	(0.0001)	(0.0001)			
Herfindahl	28.4197	29.7385	26.8579	25.6403	8.2029
	(75.7794)	(77.0970)	(72.2427)	(66.9334)	(12.1439)
Plants <500 emp	1.00026***	1.00028***	1.00028***	1.00030***	1.00036***
	(0.00009)	(0.00009)	(0.0001)	(0.0001)	(0.0001)
Manuf. share	1.72634	2.14841	1.67935	2.06135	1.6081
	(0.8932)	(1.0743)	(0.8817)	(1.0461)	(0.5873)
Tot. emp.	0.9999	0.9999	0.9999	0.9999	0.9999
	(0.00004)	(0.00004)	(0.00004)	(0.00004)	(0.00004)
Observations	2113	2277	2113	2277	7614

Period, MSA, and tech. cat. dummies included in all specifications
Robust standard errors in parentheses

As a second robustness test, we include in our OLS specifications a set of MSA-specific linear time trends. Results, reported in 12, are similar to the previous ones, although IV coefficients are slightly smaller.

Finally, the last robustness test is the exclusion of the sixth category, which includes all the patents not classifiable under the other five categories. Results are identical and slightly less precise. We do not report them for brevity but they are available from the author upon request.

Table 12: Poisson regressions

VARIABLES	(1) Comets	(2) Comets	(3) Comets	(4) Comets
OLS/IV	OLS	OLS	IV	IV
stars	0.0504**	0.0505**	0.245**	0.230*
	(0.0235)	(0.0236)	(0.0985)	(0.132)
Oth. pat. same cat.	0.294***	0.294***	0.0873	0.0749
	(0.0434)	(0.0435)	(0.106)	(0.0866)
Pat. oth. cat.				-0.127
				(0.315)
Herfindahl		7.032		6.577
		(7.381)		(6.765)
Plants <500 emp		0.108		0.308
		(0.337)		(0.321)
Manuf. share		-0.252		-0.581
		(1.171)		(0.989)
Tot. emp.		0.205		0.249
		(0.511)		(0.453)
Observations	2113	2113	2113	2113

Period, MSA, and tech. cat. dummies included in all specifications
Robust standard errors in parentheses

6 Conclusions

Two main conclusions emerge from the analysis: i) once controlling for the overall concentration of the patenting activity, stars and comets are associated with cities with different structural characteristics, and ii) the activity of star inventors is highly beneficial to the activity of comet inventors. More research is needed to clarify both the points: about the first one, in order to better identify the characteristics of cities associated with concentrations of the two categories of inventors; regarding the second, to investigate the channels through which the spillovers take place. [TBC]

6.0.4 Policy implications

The policy recommendations are not clear-cut. On one side, given the strong effect of stars on the productivity of comets, the attraction of stars to a city may be highly beneficial to the local economic environment: stars will benefit comets, which in turn will 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 employment and value added is located elsewhere, they may be highly beneficial in the light of the aforementioned indirect effect.

On the other side, we know that stars and comets are concentrating in

different places, which might imply that attracting stars where comets are might not be a successful policy, as stars in 'comets' places' may be less productive. [TBC]

References

Acs ZJ, DB Audretsch and MP Feldman, 1992, Real Effects of Academic Research: Comment, American Economic Review, 82:1

Acs ZJ, DB Audretsch, 1990, Innovation and small firms, the MIT Press Real Effects of Academic Research: Comment

Audretsch DB, 2002, The dynamic role of small firms: evidence from the US, Small Business Economics, 18:1-3

Balasubramanian N, J Sivadasan, 2008, What Happens when Firms Patent? New Evidence from US Economic Census Data, working paper

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

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

Greenstone M, R Hornbeck, E Moretti, 2008, Identifying agglomeration spillovers: evidence from million dollar plants, NBER Working Paper n. W13833

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

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.

Peri G, 1995, 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

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

Zucker, L G. and MR Darby, 2007, Star Scientists, Innovation and Regional and National Immigration NBER Working Paper Series, w13547,

Appendix A: Data

Patents' data come from the NBER/USTPO database described in Hall et al, 2001. To the original dataset we added the inventors' unique identifier developed by Trajtenberg et al (2006) and the standardized assignee name available in the Prof. Bronwyn H. Hall website. We are aware that the latter is not always 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).

We eliminated 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 Web Services. In case more authors are listed for the same patents and they live in different city, the city of residence of the first author is chosen; the procedure is standard in patent literature and Carlino et al. (2006) show that the approximation is substantially innocuous. The geocoding operation was successful for the 1,161,650 patents, which correspond to the 97% of the database. We then assigned cities to counties using the ArcGis spatial join tool, and subsequently counties into MSAs (1993 definition). Those counties which are not included in the MSAs dataset are reported singularly - the geographical units are therefore a mix of counties and MSAs (for simplicity in the paper we do not distinguish between the two entities and call all the spatial units 'MSAs'). This is a sensible choice to the extent that small counties not included in the MSAs definition do not exhibit strong commuting flows and are therefore self-contained functional entities. To our knowledge, is the first time that patents' data are geocoded (almost) entirely, without disregarding small counties.

Other County and MSA specific variables for employment and industrial structure are calculated from the County Business Pattern dataset, while data on establishments' births come from Company Statistics. Both the databases are freely available from the US Census webpage.

Figure 1: Share of stars by MSAs, kernel density

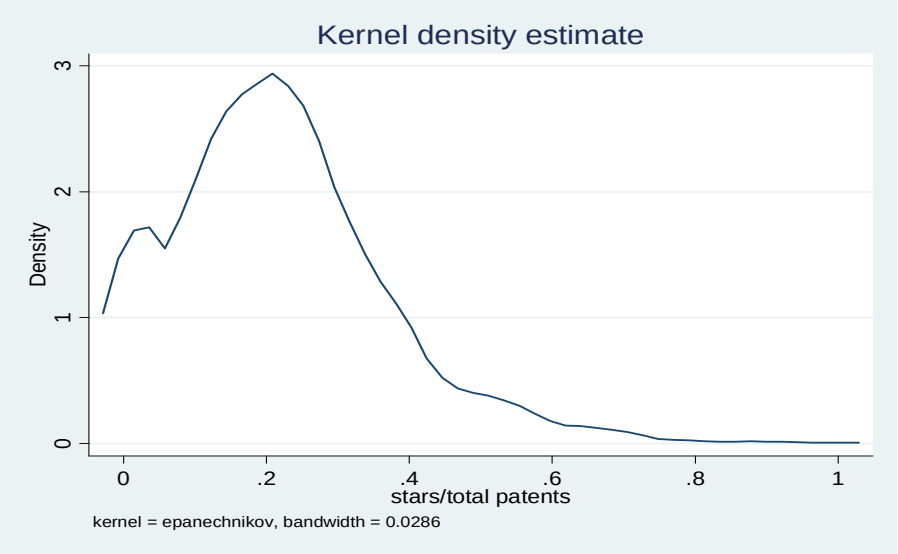

Figure 2: Share of comets by MSA, kernel density