Ecis

Eindhoven Centre for Innovation Studies

The Identification of Breakthrough Inventions Using Tail Estimators: Counting Superstar Patents

Carolina Castaldi and Bart Los

Working paper 15.05

Eindhoven Centre for Innovation Studies (ECIS),
School of Innovation Sciences,
Eindhoven University of Technology, The Netherlands

The Identification of Breakthrough Inventions Using Tail Estimators: Counting Superstar Patents*

Carolina Castaldi[#] and Bart Los[+]

[#]Eindhoven University of Technology, School of Innovation Sciences, P.O. Box 530, NL-5600 MB Eindhoven, The Netherlands; c.castaldi@tue.nl.

[+]University of Groningen, Faculty of Economics and Business, Groningen Growth and Development Centre, P.O. Box 800, NL-9700 AV Groningen, The Netherlands; b.los@rug.nl

ABSTRACT

This paper introduces a method to identify superstar patents, among all patents granted by the United States Patent Office to inventors residing in the US, with application years between 1976 and 2000. Superstar patents are defined as patents that have attracted so many citations from subsequent patents that they must have acted as a 'focal point' for later successful inventive efforts. The paper explains the relevance and logic of this new identification procedure, offers the methodological details and illustrates current and future applications for economic research on innovation.

* The authors thank Koen Frenken and Bart Verspagen for useful discussions and James Bessen, Bronwyn Hall and Colin Webb for offering constructive comments on peculiarities of the NBER Patent-Citations Database. Previous versions of this paper were presented at the International Schumpeter Society Conference (Rio de Janeiro, July 2008) and at a seminar at Utrecht School of Economics (November 2008). Sponsorship of the European Commission within the FP7-funded EU KLEMS project is gratefully acknowledged.

1. Introduction

The spotlight of research on innovation is being increasingly pointed towards a tiny set of inventions that people have often called 'breakthrough' inventions. The leading role of these inventions has been studied across several research areas. Breakthrough inventions are essential for the development of new technological fields (Fleming, 2000) and new product markets (Christensen and Bower, 1986). They can shape the fates of companies in dynamic fields (Ahuja and Lampert, 2001) and they are also able to rewrite the history of cities and regions by igniting innovation hotbeds (Kerr, 2010). All kinds of research questions concerning breakthrough inventions (such as "How do they originate?", "How do they affect economic performance?", and many more), ultimately rely on the ability of researchers to correctly identify breakthrough inventions in empirical studies. Innovation researchers have long acknowledged that inventions are highly heterogeneous in value, however measured, but it remains challenging to identify the real outliers, those breakthrough inventions that are significantly different from 'run-of-the-mill' inventions. The breakthrough inventions can of course be identified through historical, in-depth case studies of technologies, but systematic indicators for statistical research covering multiple technologies, firms or regions are trickier to derive. Nevertheless, assessing the relative inventive performance of firms, regions and even countries by simple counts of inventions is increasingly being replaced by attempts to count only breakthrough inventions. Unfortunately the methods currently used suffer from a number of shortcomings and a substantial degree of arbitrariness. In this paper, we introduce a database of 'superstar patents' representing breakthrough innovations. The database covers the period 1976-2000 and relates to inventions produced and patented in the United States.

Patent data offer many advantages for measuring new technological discoveries in a consistent and coherent way (Smith, 2005). In our database construction, we rely on the number of citations received to gauge the value of patents (Trajtenberg, 1990, first assessed the validity of using such numbers of 'forward citations' as an indication of value). The need of separating the 'really valuable' patents from the rest has become ever more pressing in view of the surge of strategic patenting with increasing shares of 'junk patents' (Jaffe and Lerner, 2004). We chose the term 'superstar patent' to relate to a stream of literature that has investigated recurring regularities in the statistical properties of a range of variables that all display highly skewed distributions and are characterized by a few superstar units. We start from a statistical regularity about the distribution of patent value that was uncovered by Silverberg and Verspagen (2007). They show that the extreme right tail of the distribution of citations to patents is characterized by a power law, while the frequencies of patents with fewer citations follow a lognormal distribution.

Apparently, a small share of patented inventions can be considered as superstars, just like a few superstar artists and athletes earn disproportionately much more money.[1]

Ultimately, we propose an indicator that produces counts of superstar patents with two key properties, whose relevance will become evident once we explain the details of our approach. First, the criteria for a superstar patent are endogenously determined from the data instead of exogenously fixed according to rules of thumb, as is common in the current literature. This allows us to have specificity according to two key dimensions. The criteria are both technology-specific and time-specific. There is hardly any reason to assume that superstar patents appear in identical proportions across technologies and the criteria should be allowed to vary across application years, since technologies tend to go through innovation cycles with different inventive foci (Utterback and Abernathy, 1975).

Second, the methodology is appropriate to provide rather timely counts, i.e. numbers are available within a reasonable time span after patent publication. Given truncation issues in patent data, citation-based indicators of importance are generally bad at assessing the value of younger patents. We introduce a probabilistic approach that relies on patent characteristics (based on shorter time series), which predicts the likelihood of a patent becoming a superstar rather well.

The organization of the rest of this paper is as follows. In Section 2, we discuss some important properties of patent citation data. Section 3 gives an overview of the literature related to the identification of breakthrough inventions using patent data. Section 4 is devoted to the exposition of our superstar patents identification procedures. In Section 5, we show some properties of superstar patents in the database and compare these with the control group of patents in the database that do not belong to this category. Section 6 concludes with a discussion of current and future applications of superstar patents.

2. Patent Citation Data

In a seminal paper, Griliches (1990) argued that the use of patent counts as innovation output indicators is riddled with problems. One of the most prominent problems is that the actual impact of patents is extremely heterogeneous, both within and across industries or technology fields. For instance, many patents do not relate to a substantial innovation over current practice, but are mainly applied for by the eventual patentee with strategic considerations in mind. In a recent series of papers and books, citation counts have been used as a proxy for the value of

[1] For other examples see, e.g. Clementi and Gallegati (2005) and Castaldi and Milakovic (2007) for personal income distributions, Eeckhout (2004; 2009) and Levy (2009) for the size distribution of cities, and Mitzenmacher (2004) for the size distributions of electronic files.

patents. Jaffe and Trajtenberg (2002), for instance, contains some classic articles in which several indicators of importance were constructed and used to analyze the innovative performance of firms, universities and other research institutes. Jaffe and Trajtenberg also initiated the construction of the NBER Patent-Citations Datafile that includes the data to operationalize their importance indicators for empirical research. The most basic indicator is the unweighted forward citation count: the more citations a patent receives in subsequently granted patents, the more important it is considered to be. This claim was confirmed by Trajtenberg (1990), among others. We also use the number of forward citations (NCITING, in the notation adopted by Trajtenberg et al., 1997) as our point of departure: a patent that is cited more often than another one has had more impact on subsequent technological developments and can therefore be seen as more important.

A few things should be taken into account before the NCITING-values for two arbitrary patents can be compared directly. First, as is well known (see e.g. Cohen et al., 2000), the propensity to patent innovations differs considerably across technology fields. Hence, patents in technologies with low propensities to patent will generally not be able to attract similar numbers of forward citations as patents in technologies in which virtually all inventions are patented. Second, not all citations are received at once. Verspagen and De Loo (1999) reported that citations to patents issued by the European Patent Office applied for between 1979 and 1997 had been received after 4.67 years, on average. Based on citations to USPTO patents granted during a much longer period, Hall et al. (2002) even find mean lags of up to 16 years. The consequence of the generally long lags is that relatively new patents will often have received fewer citations than older patents. Third, another issue that precludes reasonable comparisons of citation-based indicators across years relates to observed increasing propensities to cite. As Hall et al. (2002) argue, increased computerization of the patent system led to less time-consuming queries by patent examiners, as a consequence of which the citations to patent ratios rose considerably in the 1980s. This implies that NCITING of a patent of 10 years old applied for in 1975 cannot directly be compared to NCITING of patent of the same age applied for in 1995, for example.

To deal with these differences, we base our criteria regarding NCITING on technological category-specific cohorts of patents applied for in a given year.2 The NBER Patent-Citations Datafile contains data on patent citations to utility patents granted by the U.S. Patent Office. The original dataset included all patents granted in the period 1963-1999 and was updated by Bronwyn Hall to cover the years until 2002. The most recent NBER patent citation database now covers the years 1976-2006. For reasons of lags in patent granting procedures, which will be

[2] We use the application year since citations can be collected from the moment that an application for a patent is filed. Also, the application year is closer to the actual time of invention than the grant year.

detailed at a later stage, patents with application years after 2000 cannot be classified into superstar patents or regular patents.

We assigned patents to the technological subcategories defined in Hall et al. (2002), constructed from grouping patent technology classes. The USPTO classifies patents in about 400 main 3-digit patent classes. Hall et al. (2002) aggregated these classes into 36 subcategories. The level of aggregation is relatively high, but there is a trade-off between accounting for more intra-category variability and having enough patents for a given technology and year. If only few patents are present, statistical procedures are unable to discriminate between superstar patents and regular patents. Finally, we excluded all subcategories containing miscellaneous classes of a given category (the last digit of these subcategories is 9), because these subcategories are by definition very heterogeneous in terms of technological characteristics.3

Finally, some specific properties of how citations are registered at USPTO make it problematic to compare citations received by patents from US-based inventors with patents invented by non-US residents. Applicants (and patent examiners) can add citations to USPTO versions of a patent, or to other members of the same 'patent family', for example a patent granted by the European Patent Office (EPO). Citations to non-USPTO family members have not been included in the NBER Patent Citations Datafile, as opposed to USPTO family members. Since US inventors tend to include citations to USPTO versions of patents while non-US inventors also show a home bias in citing, this property of the database renders comparisons of US inventions and foreign inventions invalid (see Webb et al., 2005). For this reason, the analysis and ensuing database only consider patents with a US-based first inventor.

In sum, the final dataset contains all US-invented patents granted by USPTO with application years ranging from 1976 to 2000 and information on citations received until 2006.4 These patents include patents granted to individuals and governments, but more than 75% were awarded to non-governmental organizations (corporations and universities).5 In total, the analysis has been conducted on a database of almost 980,000 patents.

3. Previous Patent-Based Indicators of Breakthrough Inventions

Given that a small fraction of patents are granted to inventions that have a big impact (in economic terms and/or with respect to subsequent technological developments), recent large-

[3] See the appendix for a list of the technological subcategories that have been included.

[4] Self-citations (i.e. citations to previous patents granted to the same organization) have been included. These partly reflect the cumulative nature of invention processes (Dosi, 1982).

[5] See Hall *et al.* (2002, p. 413) for details.

sample studies on the inventiveness of firms, regions and countries have attempted to account only for those highly valuable patents. Given that 'run-of-the-mill' inventions are much more common than what we will label 'superstar' inventions, just counting all patents might well yield misguided conclusions on the degree of innovativeness. Several approaches to focus on the technologically most important subset of patents have been proposed. We review a set of these approaches, in order to clarify their pros and cons and ultimately justify the contribution of our novel methodology (for a comprehensive discussion of indicators of patent value see e.g. Dahlin and Behrens, 2005, and Van Zeebroeck, 2011).

A handful of studies, such as Clark *et al.* (2010) have chosen to focus on 'triadic patents', i.e. inventions for which patent protection has been sought (and found) in Europe, the US and Japan. This is a subset of the entire set of patents granted by USPTO. Since applying for protection at different offices involves more costs and efforts, triadic patents can be viewed as representing inventions that were considered by applicants as relatively important. Similarly, one could focus on the subset of patents applied through the Patents Cooperation Treaty, which implies that applicants filed a patent at multiple national or supranational patent offices (see for instance Usai, 2011).

Most methods follow Trajtenberg (1990) and rank patents by the numbers of citations received (NCITING). Methods differ, however, regarding the rule that separates breakthrough patents from regular patents. In some cases, a rather arbitrary cutoff point in terms of an absolute value for NCITING is set a priori. Schoenmakers and Duysters (2010), for instance, defined all (EPO) patents that received at least 20 citations within the first five years after application as 'radical', irrespective of the year of application and the technology subcategory concerned.

In most cases, the rule involves considering an upper quantile of the citations distribution (see for instance Phene *et al.*, 2006 or Ahuja and Lampert, 2001). Akkermans *et al.* (2009) considered several citation-based indicators of invention importance (NCITING was one of these) for an economy-wide set of industries, and considered the top-5% and top-10% of patents in terms of these indicators as 'radical' inventions. These quantiles were constructed per cohort of patents (characterized by year of application) in an industry-of-origin context.[6] For each cohort, they considered citation windows of maximum length, i.e. they used all citations received up till the year after which the database was truncated. A similar approach was adopted by Singh and Fleming (2010), who selected the patents that belonged to the top 5% for specified technological subcategories and years. In identifying inventions that initiated General Purpose

[6] Akkermans *et al.* (2009) were mainly interested in the performance of industries, and therefore mapped technology categories onto an industry classification. The industry-of-origin refers to the industry where an invention was most likely produced.

Technologies, Hall and Trajtenberg (2006) did not distinguish between technological categories and years of application, but roughly selected all patents that received at least three times the number of citations received by the patent that demarcated the 99th percentile of the entire NCITING-distribution. The cohort-specific technology-specific computation of upper quantiles adopted by Akkermans *et al.* (2009) and Singh and Fleming (2010), among others, recognizes that patenting and citing behavior vary across technologies and over time (see Hall *et al.*, 2002, for an early plea for such an approach), due to differences in propensities to patent, stages of the technological life cycles, etc. (see the previous section).

In summary, we feel that considering a patent that received many more citations than an equally old patent as more important is justified. It is clear, however, that there is no reason other than convenience to consider the most heavily cited 10% of amusement devices-related patents granted in 1976 as equally important as the most heavily cited 10% of organic compounds-related patents in 2000. Endogenously estimating cutoff points, based on statistical properties of empirical distributions for NCITING, appears to us as a preferred option.

A recent empirical approach proposed by Silverberg and Verspagen (2007) provides a way to move in this direction. The basic intuition is that the top cited patents, i.e. the ones falling in the right tail of the distribution of the value of patents, follow a power law distribution. This distribution is intrinsically different from the lognormal distribution that characterizes the numbers of citations to the rest of the patents. An estimate of the cutoff point between the power law tail and the lognormal part can be used as an endogenous rule to draw the line between very important patents (labeled 'superstar patents') and regular patents. In the next section we explain the methodological details of our approach.

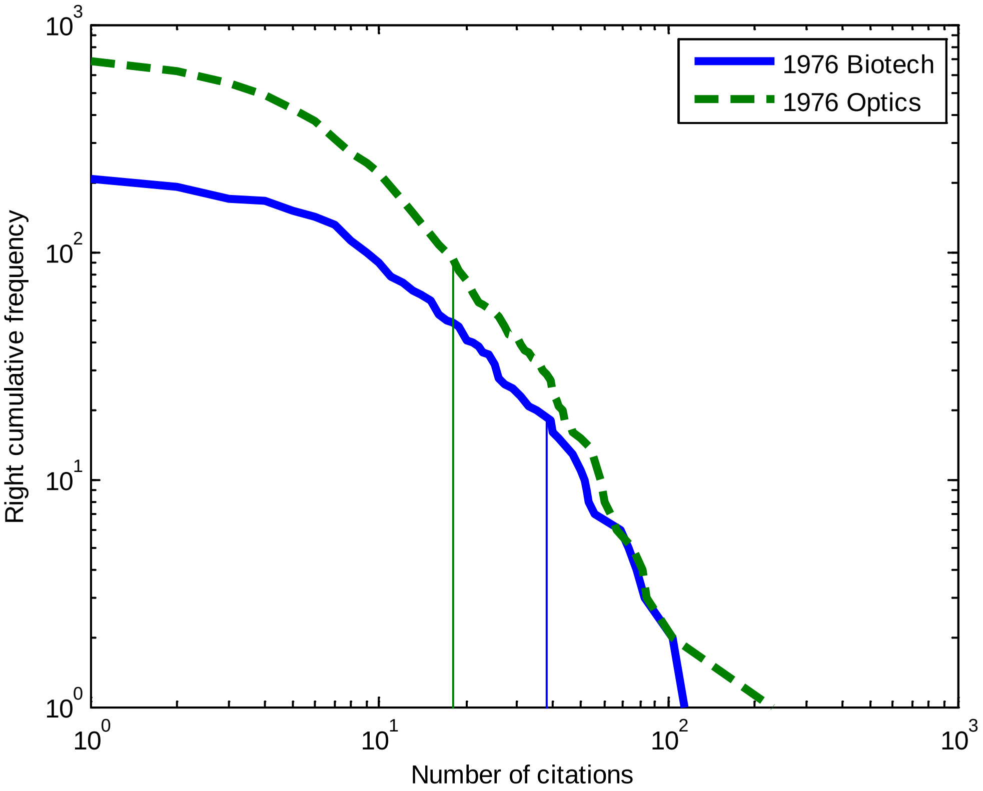
4. Methodology

In this section, we give a detailed account of the procedure that we follow to single out superstar patents. As mentioned in the previous section, we order the patents that were assigned to a technology category in a given year on the basis of citation-based scores.

4.1 Counting superstars

The point of departure of Silverberg and Verspagen (2007) is the by now uncontested finding that returns to innovation activity are highly skewed. Only about one in every four innovation projects yields a positive return and only a few projects generate the big chunk of total returns to investment in R&D. The specific skew statistical distribution that is most appropriate to describe the empirical distribution has been a topic of debate. Traditional goodness-of-fit tests suggest that lognormal distributions do a good job, but more thorough examination shows that Pareto distributions are superior in matching the observed frequency distributions in the right tail of the returns distribution, i.e. the frequencies for the most valuable innovations (see, e.g., Scherer et al., 2000). This phenomenon can also be observed for frequencies of numbers of forward citations of patents.

Figure 1: Fat tails in Numbers of Forward Citations



Source: Authors' computations on NBER Patent-Citations Datafile, citations received in 1976-2006. Estimated cutoff points between lognormal distributions and Pareto distributions (vertical lines) obtained by Drees-Kaufman-Lux procedure.

To illustrate this, we generated Figure 1 by ordering all patents with application year 1976 assigned to subcategory 12 ("biotech"), according to the numbers of citations they received in the period 1976-2006 and adopting an identical procedure for patents with the same application year but assigned to subcategory 22 ("optics").[7] The numbers of citations are depicted along the horizontal axis. The numbers of patents with an equal or higher number of citations than the value depicted on the horizontal axis are indicated along the vertical axis. Since both axes have a logarithmic scale, a Pareto distribution appears as a straight, downward sloping line. Exponential distributions (such as the lognormal) show curvature. For both technologies, a mixed distribution seems to fit the observed frequencies in Figure 1 more accurately than one type of distribution over the whole range. For less-

[7] We selected these categories and year for ease of exposition only.

cited patents, lognormal distributions fit the ever more steeply declining curves better. The rightmost parts of the curves are approximately linear, reflecting Pareto distributions. The numbers of patents in the tail differ significantly for the two groups of patents (18 for Biotechnology; 82 for Optics), as do the proportions of patents in the tail in all patents (8.6% for Biotechnology; 11.9% for Optics). This illustration strengthens the need for a technology-specific approach.

Silverberg and Verspagen (2007) present extensive evidence that a mix of lognormal and Pareto distributions can be used to describe observed frequency distributions for a variety of indicators of patent importance, such as patent valuations obtained by surveys among inventors and data on actual revenues from patents. As Silverberg and Verspagen argue in a related paper (Silverberg and Verspagen, 2005), breakthrough inventions come about in a different way than run-of-the-mill inventions. They derive part of their argument from Dosi's (1982) observation that radical innovations evoking a 'technological paradigm' are almost always followed by swarms of more incremental innovations. Rosenberg (1969) already suggested that radical innovations work as 'focusing devices' for the subsequent incremental innovations. Given that a dominant design is slowly emerging in such cases, the degree and nature of uncertainty surrounding innovation processes change over time. Changes in behavior by potential innovators caused by the changing environment they face could well yield different statistical distributions that govern innovation and citation frequencies.[8] Curves like for "Biotechnology" in Figure 1 suggest that a limited number of breakthrough inventions took place in 1976, which received numbers of citations that would never have been attained if NCITING would have been lognormally distributed for the entire sample.

A branch of statistics, so-called Extreme Value Statistics (see for instance Coles, 2001), is concerned with the properties of fat-tailed distributions. Results from Extreme Value Statistics allow us to estimate the numbers of citations that correspond to the cutoff point. We call a patent a superstar patent if it received the cutoff point number of citations or more.

Following Silverberg and Verspagen (2007), we estimate the cutoff point using an estimator for the essential parameter of the Pareto distribution. If the tail follows this distribution F(x) = 1-$x^{-\alpha}$, a maximum likelihood estimator of the parameter α can be obtained using the Hill estimator (Hill, 1975). Such estimator has a very simple expression. Given the rank-order statistics of the sample $X_{(1)} \geq X_{(2)} \geq \ldots \geq X_{(n)}$, the Hill estimator of the inverse of α is obtained as:

$$\hat{\gamma} = \hat{\alpha}^{-1} = (1/k)\sum_{i=1}^{k}\left(\ln X_{(i)} - \ln X_{(k+1)}\right)$$

[8] See Sanditov (2005) for similar arguments based on generalized Polya urn processes.

Note that the parameter alpha reflects the magnitude of the negative slope of the straight line characterizing the Pareto distributions in Pareto-plots like Figure 1.

The value of the Hill estimator is a function of *k*, the number of observations included in the tail.

The slope parameter of the Pareto-distribution is initially estimated for an extremely small subsample, which contains the most highly-cited patents only (to be found in the extreme right of Figure 1). The estimated slope is obtained by means of the well-known Hill-estimator. Next, the subsample is extended with the most cited patent that did not belong to the initial subsample and the Hill-estimator is computed again. This procedure is repeated for a successively growing subsample of well-cited patents. As long as these growing subsamples remain drawn from a Pareto distribution indeed, the estimated slopes will remain relatively stable. This changes, however, as soon as patents are added that are well-cited, but belong to the lognormally distributed part of the set of patents. This can be easily visualized with the aid of a so-called Hill plot: the sequence of estimated slopes starts to show a saw-toothed pattern, and each added patent causes a swing in the estimated slopes. The Hill plot can be used to get an idea of the value at which the Hill estimates stabilizes. For very low values of k the estimates will be highly fluctuating. If the underlying distribution is Paretian, the Hill estimates will stabilize at a certain value. But if the distribution is not overall Paretian, including observations from the central part of the distribution will decrease the validity of the estimator. A method is then needed to estimate the 'optimal' value of the parameter k.

Lux (2001) provides an overview of various methods proposed to estimate the parameter k. In the computationally convenient procedure adopted by Drees and Kaufmann (1998), the length of the right tail is first set to one observation. Next, the most likely length is found by examining the fluctuations in the value of the Hill-estimator when adding more observations to the tail. Such fluctuations emerge if Hill-estimators are applied to distributions that are not Pareto. If a predetermined threshold value is exceeded by the fluctuation, an estimate for k is found. We use a slightly modified version of this Drees-Kaufmann estimator, proposed by Lux (2001): in this version the stopping rule is modified with a higher threshold so that the tail includes fewer observations from the central part of the distribution. An additional modification from our side was to include in the tail also all patents with the same number of citations of the ones automatically included in original estimation.

Silverberg and Verspagen (2007) did not study the distributional properties of the stochastic estimators for *k* and α, nor the properties of the resulting counts of superstars. The statistical properties of these estimators cannot be derived analytically, but it is well-known (see, e.g. Clauset et al., 2009) that it can be hard to tell a lognormal and a Pareto distribution apart. We followed the suggestion of Clauset et al. (2009) to construct confidence intervals by means of bootstrapping methods. We use the by now fairly standard bootstrapping procedure of drawing a large number of pseudo-samples of a size equal to the real sample, by drawing from the observations with replacement (Efron and Tibshirani 1986, 1993). For each of these pseudo-samples the cutoff-points are estimated by means of the Drees-Kaufmann-Lux (DKL) procedure described above. Next, these are ordered. 90% confidence intervals are constructed by determining the values of the estimators for the 5th percentile and the 95th percentile of the corresponding values found for the pseudo-samples.

5. Tail estimators: Patterns across Technology Fields

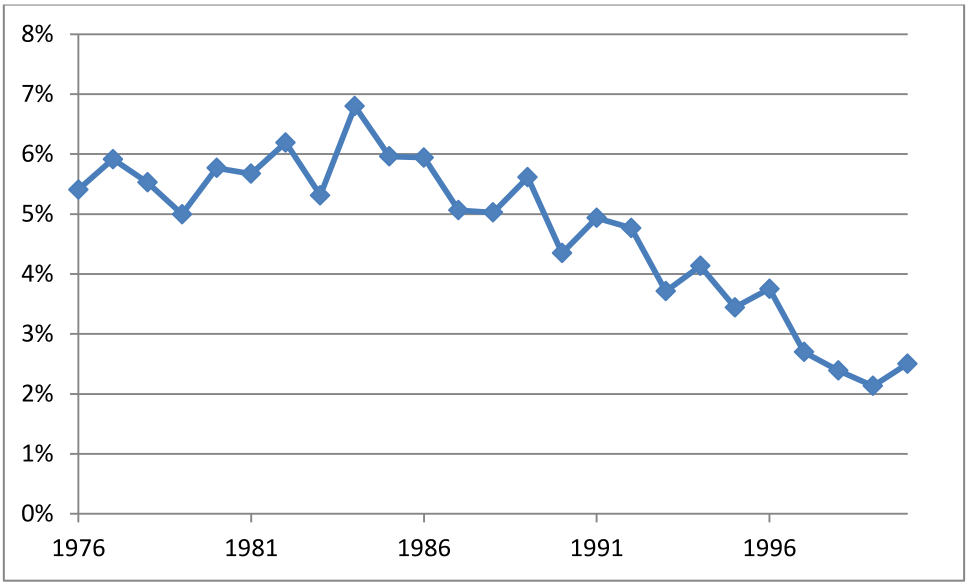
In this section we present results from our estimation procedures. We applied the Drees-Kaufmann-Lux procedure to sets of patents sharing the same application year and assigned to the same technological category. We selected patents that received at least one citation. This is done because the Hill estimator involves taking the logarithm of the variable at hand, so positive values only can be used. The share of superstars in each cohort is the ratio of the number of superstar patents, i.e. those patents receiving more citations than the estimated cutoff point, relative to the total number of patents. We perform two general checks to explore the validity of the results from the tail estimators.

As a first graphical exploration, Figure 2 shows the evolution in time of the estimated shares of superstars, averaged across categories. At least until 1985, the estimates remain remarkably stable. After 1986, we observe a clear monotonic decrease in the average shares.

This decrease could indicate a real phenomenon related to a decrease in technological opportunities or to a structural change in the use of patenting, but a more likely candidate explanation is found in the already discussed truncation problem of patent data. Patents applied for in the later years can receive much less citations than older patents. Hall et al (2002) show that mean citations received decline from around 1986 (Figure 8 in their paper). We partially solved this problem by estimating cut-off points for cohorts (instead of pooling patents for all years), which implies that patents in young cohorts require fewer citations to be superstar than patents from older cohorts (everything else equal). But if superstar patents tended to receive more citations far after their application year, the distribution of citations received for recent patents would look more log-normal and the tail would not appear in any significant way. Hall and

Trajetenberg (2005) already show that highly cited patents in the NBER Database have significantly higher citation lags. We check if their claim applies also when highly cited patents are obtained using the adjusted DK routine. It turns out that the half-life of superstar patents is generally higher than the one of less important patents (results not shown here).9 The half-life is calculated as the median age of all patents citing the group of patents in the tail for each category and year. We can then claim that superstars are patents that receive citations for a longer time span. Hence, when relying on a relatively short citation window, superstar patents cannot be properly distinguished from regular patents. Looking at Figure 1, we can claim that only for patents applied until 1985 the superstar ones can safely be distinguished. Our proposed solution to this truncation problem is to adopt a probabilistic approach for younger patents, which we discuss in the next section.

Figure 2: Shares of superstar patents over time[*]



*All indicated proportions are computed as weighted averages of the technology-specific shares obtained from the tail estimator.

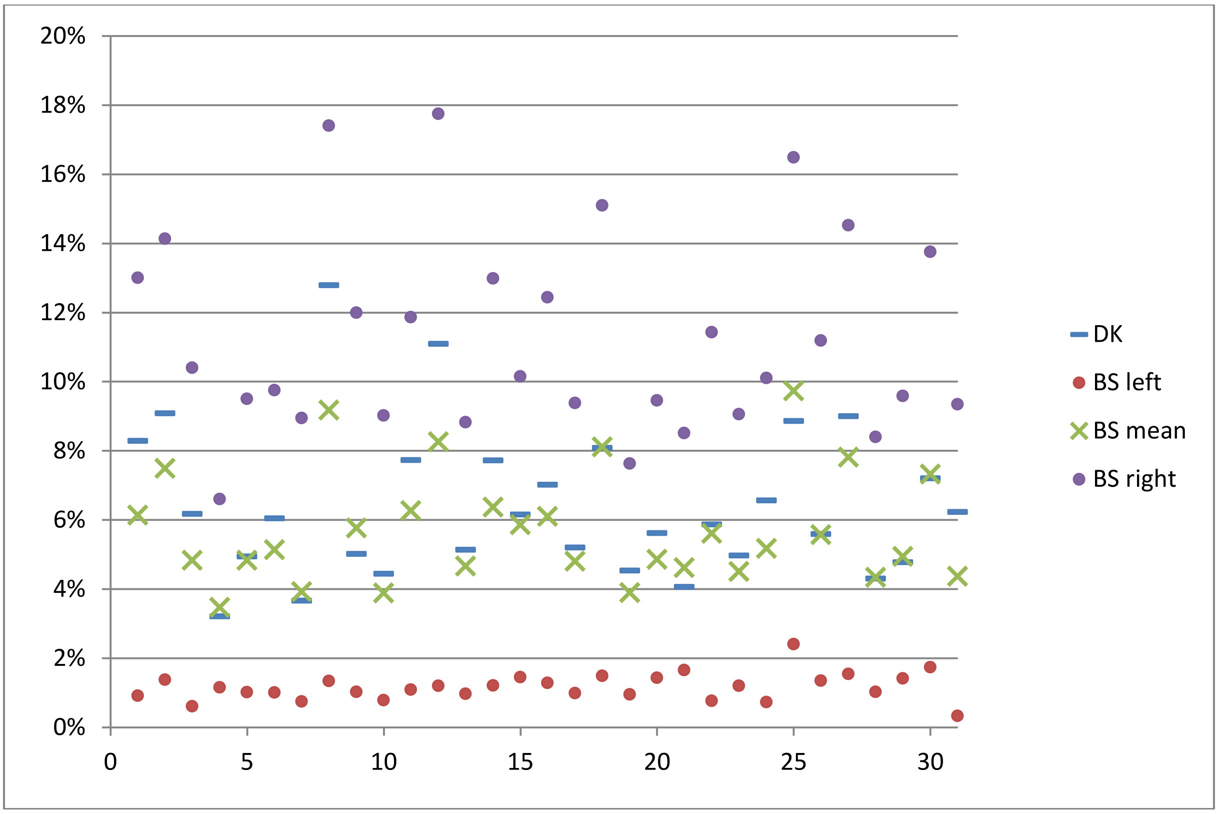
A second explorative analysis we perform is studying the variability of the estimates across technologies. We calculated bootstrapped estimates of 90% confidence intervals and included the mean bootstrap estimate in order to compare it with the point estimate from the DKL

[9] An example: It took regular patents in the 1980-cohort of the category Information Storage about 7 years to receive 50% of the citations it had accumulated in 2006, whereas it took the superstar patents as many as 12 years to achieve this.

procedure. Figure 3 shows the estimated shares of superstars per technology, averaged across the years 1976-1985 (which, based upon the truncation problems just discussed, are the only meaningful period to consider here). The graph indicates a high cross-sectional variability of the estimates. For a number of technologies and years, the confidence intervals are so wide that it proves difficult to engage in any meaningful discussion of inter-technology differences of the point estimates. On the other hand, we consider the fact that DKL estimates and bootstrap mean are generally close to each other as a positive property of the DKL procedure adopted by Silverberg and Verspagen (2007). Nevertheless, relying on point estimates would undermine the stability of the results. To overcome this problem, we opt for taking the bootstrap mean estimates to base our definition of superstar patents, given that these means are less affected by randomness.

In sum, the two explorative checks highlighted that point tail estimators cannot be directly used to reliably identify superstar patents. Next, we illustrate the method that overcomes the two problems of the deterministic DKL point estimates discussed above.

*Figure 3: Shares of Superstar Patents by Category (averages)**



*All indicated shares are computed as weighted averages of the year-specific shares, 1976-1985.

5. A probabilistic approach

For younger patents, we do not know for sure whether they will become superstars, but we can estimate the probability of this event. These estimates are based on patterns of citation reception of older patents. We feel it can be justified to decide on the superstarness of patents after 20 years, since vast majorities of old patents in most technology categories had received at least 80% of their total citations (up to 2006) in that timeframe. To adjust for the high variability of the point estimates from the DKL estimator, we generate estimates of the number of patents belonging to the tail of the citation distribution for 1000 boostrap samples and then average the numbers to obtain a bootstrapped mean cutoff point, i.e. threshold number of citation needed to be classified as a superstar patent. This implies that we have samples of superstar patents and regular patents for the years 1976-1985, purely based on the bootstrap mean cut-off point estimates originating from numbers of citations received after 20 years. We also have information about the history of citations these old patents received, which we summarize in a set of variables supposed to predict the eventual superstarness of a patent. Formally, this means estimating a set of logistic regressions where the dependent variable is a binary variable indicating whether the patent has become a superstar after 20 years, and the independent variables are instead patent characteristics based on a shorter citation window a, which we consider to vary between a=5 and a=20. The regressions, which are estimated for technology-specific samples of patents applied in the years 1976-1985, are specified as follows.

We estimate logistic regressions for all 1976-1985 patents on technology-specific samples, for *a*=5 to *a*=20. To make sure that differences in patenting and citing behavior (over time) would not affect the regression results, all explanatory variables were expressed as standardized deviations from the cohort-specific medians.

$$\ln\frac{p_{ak,i}}{1-p_{ak,i}} = \alpha_{ak} + \beta_{ak,1} ncit_i + \beta_{ak,2} frec_i + \beta_{ak,3} GEN_i + \varepsilon_i$$

Where:

$p_{ak,i}$ is the probability that the patent eventually represents a superstar patent for technology category *k*.

$ncit_i$ represents the logged number of citations received in a citation window of *a* years. A patent that already receives a larger number of citations is expected to have a higher chance of becoming a superstar patent. The variable is standardized across patents with the same application year. Since the variable is highly skewed we take the z-score of the logarithmic value

$frec_i$ indicates the proportion of citations received in the most recent half of the window considered and is meant to take into account the fact that superstar patents receive citations for a longer time span.

GEN_i is the measure of generality introduced by Hall et al. (2002) and calculated in this case based on the technological categories. A high generality indicates that the patent received citations from many different categories. This property may be seen to increase the chance of receiving citations in general and thus the chance of eventually becoming a superstar patent (Hall and Trajtenberg, 2005).

We check the predictive validity of the logistic regressions for the 1976-1985 patents. Table 1 reports the estimated coefficients for a selection of categories and a selection of citation windows. The goodness of fit of the model is assessed though a pseudo R^2, extensively used for logit regressions (Hosmer et al, 2013).

These regressions led to parameter estimates that were very heterogeneous across technologies and ages a. The intercepts can be interpreted as the odds that a patent that has not received any citation within a years eventually becomes a superstar. As expected, this estimate sharply decreases with a: if a patent has not received any citation in the first 15 years, than chances are very low that it will become a superstar. The number of citations received is the main predictor of superstarness. Also, its estimated effect increases with the width of the citation window, as expected. Frec and GEN also play a role, in the expected direction, but with varying degrees of significance across technologies.

The explanatory power of the model (as measured by McFadden's pseudo-R2) is good. As expected, it depends on the citation window since the predictive power of the model increases when more years of citation history are available, but the goodness of fit remains reasonable for the smallest citation window.

Reassured about the predictive power of our models for the 1976-1985 patents, we then extrapolate the results for the more recent years. We use the estimated coefficients to estimate the odds that patents applied for after 1985 will eventually become superstar patents. For each patent we obtain values for the independent variables included in the logistic models based on the largest citation window available, thus a=19 for 1986 patents, a=18 for 1987 patents, until a=5 for 2000 patents[10]. For each cohort, we only use the variables whose estimated coefficient in the relevant regression was statistically significant. Next, the odds are transformed into probabilities. We do not consider individual patents, but focus on groups of patents. We opt for a

[10] Note that we do not use the last two years of the citation data: the NBER database contains information on all granted patents, but we use the application year, which means that because of the granting lags, patents with application years 2005-2006 are a much smaller subset than all granted patents.

procedure in which a patent with a 0.2 probability of becoming superstar counts as 0.2 superstar patents and 0.8 regular patents. This implies that we add probabilities. Hence, two more recent patents with a 0.5 probability of becoming a superstar patent count as "heavily" as a one early patent that we unconditionally attributed to the class of superstar patents.

Table 1: Estimates of the logistic regression model predicting superstarness (selected technologies and selected citation windows).

		const	ncit	frec	GEN	R^2
Biotech	*a=5*	-2,22**	0,71**	0,02	0,03**	0,14
		(-29,76)	(9,31)	(0,66)	(3,39)	
	a=6	-2,32**	0,87**	0,02	0,04**	0,19
		(-26,16)	(10,35)	(0,36)	(3,39)	
	a=7	-2,50**	0,99**	0,09**	0,03**	0,22
		(-26,62)	(12,56)	(2,05)	(2,27)	
	a=10	-2,92**	1,37**	0,21**	0,02	0,32
		(-25,62)	(16,59)	(3,81)	(1,07)	
	a=15	-3,77**	2,80**	-0,44**	0,05	0,52
		(-22,66)	(20,21)	(-5,31)	(1,14)	
	a=19	-4,78**	3,89**	-0,58**	0,05	0,61
		(-19,35)	(18,66)	(-5,96)	(0,74)	
Optics	*a=5*	-3,37**	0,71**	0,00	0,00	0,11
		(-39,15)	(10,37)	(0,16)	(0,13)	
	a=6	-3,44**	0,79**	-0,01	0,01	0,14
		(-35,36)	(10,34)	(-0,26)	(0,92)	
	a=7	-3,65**	0,89**	0,04	0,01	0,19
		(-35,21)	(12,06)	(0,98)	(1,08)	
	a=10	-4,03**	1,20**	0,01	0,03	0,28
		(-29,92)	(14,75)	(0,27)	(1,71)	
	a=15	-5,52**	2,37**	0,06	-0,06**	0,48
		(-22,81)	(19,01)	(0,91)	(-1,76)	
	a=19	-8,69**	4,36**	-0,01	-0,10**	0,65
		(-17,57)	(17,08)	(-0,14)	(-1,84)	

**: significant at at least 5%

6. Key results of the probabilistic approach

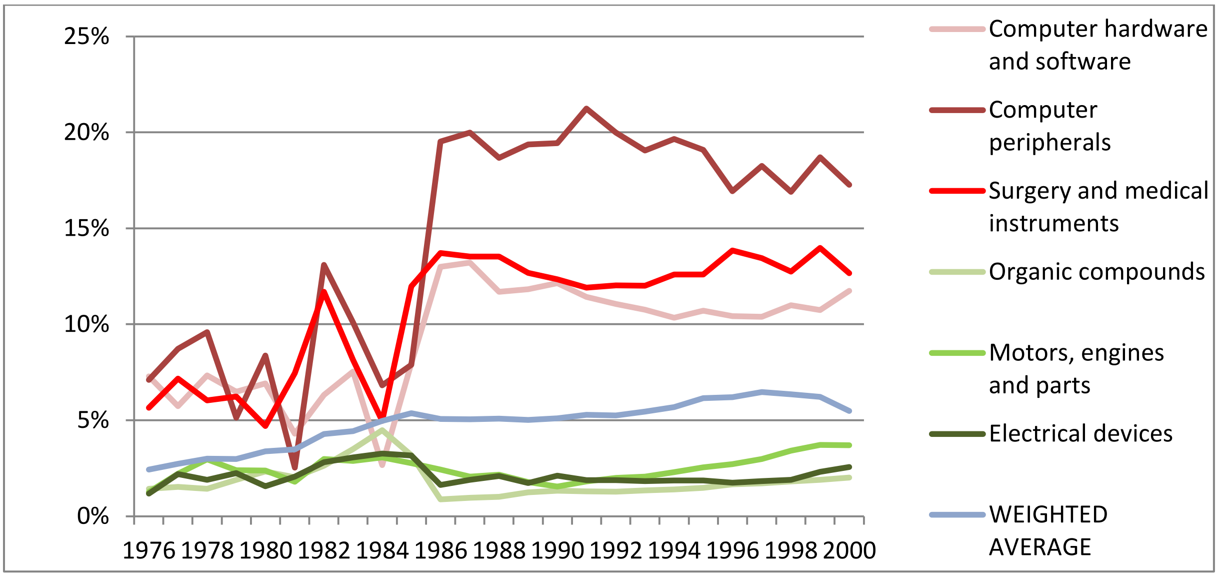
We describe here the main patterns emerging from our estimated counts of superstar patents.

Figure 4 shows the trend in the shares of superstar patents for a number of technologies: we selected the three technologies (Computer hardware and software, Computer peripherals and Surgery and medical instruments,) with the highest shares of superstars in 1976, and the two technologies with the lowest shares (Electrical devices, Motors, engines and parts, and Organic compounds). We also include the overall trend indicated by the weighted average across all technologies.

First of all, we can observe that the truncation problem we had encountered from the point estimations has now been solved. The estimated shares that we obtain from our probabilistic method for the years 1986-2000 do not show any dramatic decline. Rather, they are relatively stable and even show a moderate increasing trend.

Second, the shares of superstar patents differ substantially both across technologies and in time, confirming that measuring breakthroughs with fixed quantile definition is grossly at odds with the empirical evidence.

Figure 4: Shares of Superstar Patents generated by the probabilistic method, selected technologies and weighted average across technologies[*]



Third, while the average trend seems to indicate overall increasing technological breakthroughs, these are clearly concentrated in a few technological fields where radical new solutions have emerged in the last decades. Many of these fields are the core technologies of the ICT revolution, but other science-based fields like Drugs and Biotech demonstrate a high level of dynamism. Note that our estimates indicate that for these fields a 5% rule for defining breakthroughs would significantly underestimate the actual share, which in some case goes up to 20%. Instead, the shares of superstar patents are much lower for more mature fields. Still, even in these fields, breakthroughs also emerge. Thanks to our technology-specific indicators, we are able to capture these breakthroughs as well, which would have likely been overshadowed by the more evident breakthroughs in the more dynamic fields.

Table 2 reports indicators of the relative ability to generate superstar patents for all technological fields. We look at the overall changes in the period considered: given the volatility of the shares of superstar patents, we consider 3-year moving averages for the first (1977) and last point in time (1999). The fields are sorted by the first indicator, namely the share of the number of superstar patents generated in 1999 relative to the same share in 1977. Thereby, we can label the first block of fields, characterized by positive ratios above 4, as 'Technologies in emergence'. These are technological fields where the numbers of superstars having been substantially increasing in the period considered. Instead, 'Technologies in demise' are fields where the ratios, all below 1, point to a halt in the absolute production of superstars. By way of comparison, we also report the weight of each field in the total production of superstars, as indicated by the percentage of total superstar patents that the field produces. Those dynamic fields with growing numbers of superstars are also fields that account for most of the superstars. And their weight has also been increasing over time (as indicated by the last ratio reported in the table). This confirms that the overall increasing trend in the average shares of breakthroughs has much to do with the contribution of a few key fields.

Table 2: Comparison of technological fields in terms of trends in the production of superstar patents (all figures based on 3-year moving averages for the years 1977-1999).

	Fields	Superstar patents in 1999/ superstar patents in 1977	Percentage of total superstars in 1999	Percentage of total superstars in 1999/percentage of total superstars in 1977
Technologies in emergence	Computer peripherals	23,81	9,3%	5,47
	Semiconductor devices	10,04	7,9%	2,31
	Computer hardware and software	9,88	13,2%	2,27
	Surgery and medical instruments	9,57	14,6%	2,20
	Information storage	8,89	5,0%	2,04
	Drugs	7,36	8,0%	1,69
	Communications	6,28	8,6%	1,44
	Biotechnology	4,19	1,7%	0,96
	Coating	3,51	2,4%	0,81
	Amusement devices	3,29	1,9%	0,76
	Power systems	3,28	2,2%	0,75
	Furniture, house fixtures	3,18	1,8%	0,73
	Nuclear and X-rays	3,14	1,6%	0,72
	Electrical lighting	2,90	1,6%	0,67
	Transportation	2,83	2,6%	0,65
	Receptacles	2,67	1,7%	0,61
	Measuring and testing	2,66	1,8%	0,61
	Earth working and wells	2,23	1,3%	0,51
	Pipes and joints	2,21	0,9%	0,51
	Optics	2,19	1,2%	0,50
	Motors, engines and parts	2,16	1,8%	0,50
	Electrical devices	2,15	1,6%	0,50
	Apparel and textile	2,11	1,2%	0,49
	Agriculture, husbandry, food	2,10	1,7%	0,48
	Materials processing & handling	1,49	1,7%	0,34
	Gas	1,46	0,6%	0,34
	Resins	1,18	1,3%	0,27
Technologies in demise	Metal working	1,00	0,7%	0,23
	Organic compounds	0,87	0,7%	0,20
	Agriculture, food, textiles	0,73	0,4%	0,17
	Heating	0,58	0,3%	0,13

7. Conclusions

The challenge that we took up in this paper was the one of developing valid indicators to measure breakthrough inventions, a tiny set of technological advances with a tremendous impact on innovation and economic opportunities. While our contribution is here an empirical one, we feel that our indicators are crucial for the ability of researchers to test theoretical hypotheses about breakthrough inventions. In fact, while innovation studies have by now developed a solid understanding of how new discoveries build upon previous ones in an evolutionary fashion, much more attention has been devoted so far to processes of path-dependence (Castaldi and Dosi, 2006) rather than to the processes of path-creation (Garud and Karnøe, 2001), i.e. the processes by which new breakthrough inventions emerge and start novel technological trajectories. A host of research questions (like: which organizations are most likely to generate breakthroughs? Is the ability to generate breakthroughs always crucial for economic survival of a region?) comes to mind.

Our methodology is already being applied in research on the geography of innovation, where geographical units are compared in terms of their specialization in breakthrough innovations. Castaldi et al (2015) test theories from evolutionary economic geography that predict that regions specialized in more unrelated inventive activities are able to generate more breakthroughs than other regions. Li et al (2016) offer a cross-country analysis of eco-patenting where one of the definitions of breakthroughs involves the identification of superstar patents.

There are several other applications of our estimated counts of superstar patents. First, they can be used to map life cycles of technologies (see Haupt et al., 2007). Second, they allow investigating the origins of breakthroughs, an emerging question (see Nemet and Johnson, 2012) which bears important policy implications. From a policy perspective, if breakthrough inventions really shape the fates of companies and entire cities and regions, then it becomes crucial to timely identify those inventions that really matter from an economic point of view.

We recognize several extensions of our work. One natural extension would be to consider multiple countries, using PATSTAT data. A more original extension would be to apply our methods to scientific publications data. Highly cited scientific publications can be taken as proxies of the scientific excellence of the knowledge institutions in a region (Tijssen et al., 2002). Similarly to patents, the value distribution of scientific publications is highly skewed (Seglen, 1992), with a few publications receiving most citations. This warrants the application of our identification strategy in search of ‘superstar publications’. Such an application will of course require adjusting our methodology to the specific characteristics of publications data, including for instance different dynamics of citation patterns.

References

Akkermans, D.H.M., C. Castaldi and B. Los (2009), Do 'Liberal Market Economies' Really Innovate More Radically than 'Coordinated Market Economies'? Hall & Soskice Reconsidered, *Research Policy*, *38*(1), 181-191.

Ahuja, G. and C.M. Lampert (2001), Entrepreneurship in the large corporation: a longitudinal study of how established firms create breakthrough inventions, Strategic Management Journal, 2 (6-7), 521–543.

Castaldi, C. and G. Dosi (2006), The Grip of History and the Scope for Novelty. Some Results and Open Questions on Path Dependence, in A. Wimmer and R. Koessler, eds., Understanding Change, Palgrave.

Castaldi, C., K. Frenken and B. Los (2015), Related variety, unrelated variety and breakthroughs: an analysis of US state-level patenting. *Regional Studies*, 49(5), 767-781.

Castaldi, C. and M. Milakovic (2007), Turnover Activity in Wealth Portfolios, *Journal of Economic Behaviour and Organization*, 63, 537-552.

Christensen, C. and J. Bower (1996), Customer power, strategic investment, and the failure of leading firms, Strategic Management Journal, 17, 197-218.

Clauset, A., C.R. Shalizi and M.E.J. Newman (2009), "Power-Law Distributions in Empirical Data", *SIAM Review*, 51, 661-703.

Clementi, F. and M. Gallegati (2005), Power law tails in the Italian personal income distribution, *Physica A: Statistical Mechanics and its Applications* 350.2: 427-438.

Cohen, W.M., R.R. Nelson and J.P. Walsh (2000), Protecting Their Intellectual Assets: Appropriability Conditions and Why U.S. Manufacturing Firms Patent (or Not), *NBER Working Paper 7552* (Cambridge MA: NBER).

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer-Verlag, London.

Dahlin, K.B. and D.M. Behrens (2005), When is an invention really radical? Defining and measuring technological radicalness, *Research Policy*, vol. 34, pp. 717-737.

Dosi, G. (1982), "Technological Paradigms and Technological Trajectories", *Research Policy*, vol. 11, pp. 147-162.

Drees, H. and E. Kaufmann (1998), Selecting the Optimal Sample Fraction in Univariate Extreme Value Estimation, *Stochastic Processes and their Applications*, vol. 75, pp. 149-172.

Eeckhout, J. (2004), Gibrat's law for (all) cities, American Economic Review, 94:1429-1451.

Eeckhout, J. (2009), Gibrat's Law for (All) Cities: Reply,American Economic Review, 99: 1676-1683.

Efron, B. and Tibshirani, R. (1986), Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy, *Statistical Science*, vol. 1 (1), pp. 54-77.

Efron, B. and Tibshirani, R. (1993), *An Introduction to the Bootstrap*, (London: Chapman & Hall).

Fleming, L. (2000), Recombinant uncertainty in technological space, *Management Science*, 47 (1): 117-132.

Garud, R., Karnøe, P. (2001), Path creation as a process of mindful deviation, in: Garud, R and Karnoe, P.(eds.), Path Dependence and Creation (London: Lawrence Earlbaum Associates), 1–38

Granstrand, O. (1999), *The Economics and Management of Intellectual Property* (Cheltenham UK: Edward Elgar).

Griliches, Z. (1990), Patent Statistics as Economic Indicators, *Journal of Economic Literature*, vol. 28, pp. 1661-1707.

Hall, B.H., A.B. Jaffe and M. Trajtenberg (2002), The NBER Patent-Citations Data File: Lessons, Insights, and Methodological Tools, in: Jaffe, A.B. and M. Trajtenberg, *Patents, Citations & Innovations* (Cambridge MA: MIT Press), pp. 403-459.

Hall, B.H. and M. Trajtenberg (2005), Uncovering GPTs with Patent Data, in C. Antonelli, D. Foray, B. H. Hall, and E. Steinmueller, *Festschrift in Honor of Paul A. David*, Edward Elgar.

Haupt, R., M. Kloyer and M. Lange (2007), Patent indicators for the technology life cycle development, *Research Policy*, vol. 36, pp. 387-398.

Hill, B.M. (1975), A simple general approach to inference about the tail of a distribution, *Annals of Statistics*, 3, pp.1163-73.

Hosmer Jr, D. W., S. Lemeshow, and R.X. Sturdivant (2013), *Applied logistic regression*, Wiley.

Jaffe, A.B. and J. Lerner (2004), *Innovation and its Discontents* (Princeton NJ: Princeton University Press).

Jaffe, A.B. and M. Trajtenberg (2002), *Patents, Citations & Innovations: A Window on the Knowledge Economy* (Cambridge MA: MIT Press).

Kerr, W.R. (2010), Breakthrough Inventions and Migrating Clusters of Innovation, Journal of Urban Economics, 67, 46-60.

Li, D., F. Alkemade and G. Heimeriks (2016), *The emergence of clean-tech innovation: relatedness to country's knowledge base*, working paper.

Levy, M. (2009), Gibrat's law for (all) cities: Comment, The American Economic Review, 99.4: 1672-1675.

Lux, T. (2001), The Limiting Extremal Behaviour of Speculative Returns: An Analysis of Intra-Daily Data from the Frankfurt Stock Exchange, *Applied Financial Economics*, vol. 11, pp. 299-315.

Mitzenmacher, M. (2004), A brief history of generative models for power law and lognormal distributions, *Internet mathematics*, 1(2), 226-251.

Nelson, R.R. and S.G. Winter (1982), *An Evolutionary Theory of Economic Change*, The Belknap Press, Harvard University: London.

Nemet, G. F., and E. Johnson (2012), Do important inventions benefit from knowledge originating in other technological domains?, *Research Policy*, 41(1), 190-200.

Rosenberg, N. (1969), The Direction of Technological Change: Inducement Mechanisms and Focusing Devices, *Economic Development and Cultural Change*, vol.18, pp. 1-24.

Sanditov, B. (2005), Patent Citations, the Value of Innovations and Path-Dependency, CESPRI Working Paper 177, Bocconi University Milano.

Scherer, F.M., D. Harhoff and J. Kukies (2000), "Uncertainty and the Size Distribution of Rewards from Innovation", *Journal of Evolutionary Economics*, vol. 10, pp. 175-200.

Smith, K. (2005), Measuring Innovation, in: Fagerberg, J., Mowery, D., Nelson, R. (eds), The Oxford Handbook of Innovation, Oxford University Press, New York, pp. 148-177.

Silverberg, G. and B. Verspagen (2005), A Percolation Model of Innovation in Complex Technology Spaces, *Journal of Economic Dynamics and Control*, 29, 225-244.

Silverberg, G. and 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, pp. 318-339.

Tijssen, R. J .W., Visser, M.S., Van Leeuwen, T.N. (2002), Benchmarking international scientific excellence: Are highly cited research papers an appropriate frame of reference? Scientometrics, 54, 381-397.

Trajtenberg, M. (1990), A Penny for Your Quotes: Patent Citations and the Value of Innovations, *RAND Journal of Economics*, vol. 20, pp. 172-187.

Trajtenberg, M., R. Henderson and A.B. Jaffe (1997), University vs. Corporate Patents: A Window on the Basicness of Innovation, *Economics of Innovation and New Technology*, vol. 5, pp. 19-50.

Usai, S. (2011), The Geography of Inventive Activity in OECD Regions, *Regional Studies*, 45, 711-731.

Utterback, J.M. and W.J. Abernathy (1975), A Dynamic Model of Process and Product Innovation, *OMEGA*, vol. 3, pp. 639-656.

Verspagen, B. and I. de Loo (1999), Technology Spillovers between Sectors and over Time, *Technological Forecasting and Social Change*, 60, 215-235.

Webb, C., H. Dernis, D. Harhoff and K. Hoisl (2005), Analysing European and International Patent Citations: A Set of EPO Patent Database Building Blocks, OECD Science, Technology and Industry Working Paper 2005/9, OECD, Paris.

Van Zeebroeck, N. (2011), The Puzzle of Patent Value Indicators, *Economics of Innovation and New Technology*, 20, 33-62.

Appendix

Category Classification

Nr.	Description	Sub-category code in Hall *et al.* (2002)
1.	Agriculture, food, textiles	11
2.	Coating	12
3.	Gas	13
4.	Organic compounds	14
5.	Resins	15
6.	Communications	21
7.	Computer hardware and software	22
8.	Computer peripherals	23
9.	Information storage	24
10.	Drugs	31
11.	Surgery and medical instruments	32
12.	Biotechnology	33
13.	Electrical devices	41
14.	Measuring and testing	42
15.	Nuclear and X-rays	43
16.	Power systems	44
17.	Semiconductor devices	45
18.	Materials processing and handling	46
19.	Professional and scientific instruments	51
20.	Metal working	52
21.	Motors, engines and parts	53
22.	Optics	54
23.	Transportation	55
24.	Agriculture, husbandry and food	61
25.	Amusement devices	62
26.	Apparel and textile	63
27.	Earth working and wells	64
28.	Furniture, house fixtures	65
29.	Heating	66
30.	Pipes and joints	67
31.	Receptacles	68