Citations Profile and the Complexity of Innovation*

Guido Cozzi† and Francesco Schettino‡

August 27, 2008

Abstract

Patent or article citations reflect the consequences of a published idea on the discovery of new ideas. We draw a simple theoretical model predicting that the shape of the future citations of an idea can reveal the complexity of its innovative research spillover. We apply this method to the patent forward citations in the US industries.

J.E.L. Classification: O31, O33
Keywords: Patent Forward Citations, Technological Complexity, Skewness.

1 Introduction

The shape of the citations of a patent or of a scientific article is the outcome of the innovative process stimulated by the patent or the article. In this paper we develop a theoretical model of innovation that predicts a simple and natural measure of the potential intertemporal impact of an idea after its publication. Our model seems to generate a realistic dynamics for the patent forward citations of the US industries and provides a simple way of assessing the evolution of the complexity of innovating in the US industries.

This paper contributes to assessing the empirical relevance of the increasing difficulty on R&D (Jones, 2005), by exploiting a detailed data set for the

*With the usual disclaimers, we thank Michele Giammatteo, Luca Spinesi and Francesco Venturini for helpful comments.

†University of Glasgow, g.cozzi@lbss.gla.ac.uk
‡Polytechnic University of Marche, f.schettino@univpm.it

US economy - patent forward citations. In our model - sketched in Section 2 - each new idea can directly inspire only a number of future ideas. The researchers' work on the possible consequences of an idea is getting more and more difficult as the number of potential applications get fished out, but at the same time new avenues of research are opened up over time as a result of present discoveries. This can be studied in terms of the number of forward patent citations, which, as we shall see, seems to behave over time as predicted by our law of motion In Section 3 we inquire on innovation complexity by measuring the skewness of the patent forward citations in the 1963-1999 US data.

2 Theoretical Model

Let us assume a continuous and unbounded time horizon indexed by $t \in R_+$. We imagine that each new idea i, patented at date $t_i > 0$, discloses the possibility of inventing $X_i > 0$ new patentable ideas, which we will call "applications". We will assume that the innovation technology exhibits constant returns to R&D labor. More specifically, letting $x_i(t) \in [0, X_i]$ denote the number of applications of idea i already invented up to time $t > t_i$, the probability per unit time that a research labor unit invents a new implication of idea i is assumed equal to:

$$\beta_i \left[1 - \frac{x_i(t)}{X_i}\right] \left(\frac{x_i(t)}{X_i}\right)^{a_i} \tag{1}$$

where $a_i > 0$ and $\beta_i > 0$ are constants. Hence the innovation process per unit R&D labor is described by an independent Poisson process whose arrival rate is computed from eq.(1). Probability intensity (1) can be interpreted as the product of three factors:

1. The probability, $1 - \frac{x_i(t)}{X_i} \in [0, 1]$, of finding a potential application that at time t has not yet been explored: this is proportional to the fraction of not yet discovered applications of idea i.

2. The probability, $\left(\frac{x_i(t)}{X_i}\right)^{a_i} \in [0, 1]$, that, conditional on finding a potentially new application, the researcher is able to successfully complete it.

3. A fixed scale parameter, β_i, specific of idea i.

Given the amount, $x_i(t)$, of applications found from idea i, the higher a_i the lower the expected flow per unit time of new applications generated

by a unit R&D effort flow. This means that, given cumulated successful experience, it is easier to invent a new product for ideas that have a lower a_i. Hence the spillover parameter, a_i, captures the complexity in researching and developing new ideas from the existing ones. Note that we are here modeling an innovation process in which the ability of completing an application of an idea benefits from the intertemporal spillover of all the previously successful application solving activities stemming from that idea. Hence the social experience in finding and completing useful applications of an idea improves the ability of a researcher - who is lucky enough to be exploring a not yet applied R&D trajectory disclosed by that idea - of developing the full potential of a promising application.

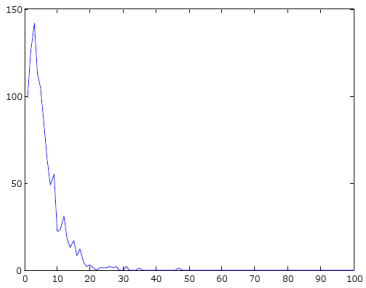
The total number of potential direct inspirations of idea i, X_i, can be interpreted as the full understanding of the innovative content of idea i.

According to equation (1), it is when we know all possible consequences of an idea that, in the case we forget the details of one of them, we are able to re-invent it with probability one. Unfortunately, the most able researcher finds her/him-self fishing in an empty pond, due to the fishing out effect. Conversely, when only a small percentage of the potential consequences of an idea have been found, the initial idea is only partially assimilated by the researcher's mind, which means that most of its complex implications remain obscure: this justifies the low ability of the R&D workers of finding one of its consequences. Hence equation (1) states that the probability of completing the second promising application if $X_i = 2$ is much higher than the probability of completing the second promising application if $X_i = 200$.

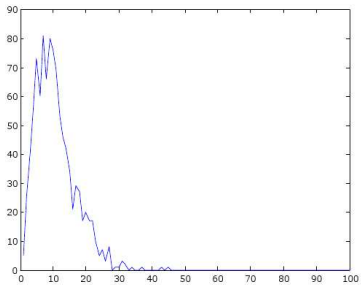
The figures in (1) are constructed by running[1] our citations stochastic process under the following assumptions: $\beta_i = 200$; $X_i = 1000$, and $a_i = .1$, .5, 1, and 1.5.

The North-West quadrant depicts a typical citations profile of a paper or a patent exhibiting high initial spillover. Complexity gets higher the more we read clockwise. The later the intertemporal spillover reaches its maximum the more complex the idea. This is a regular feature of our model. We can immediately see from the figures that the more difficult the idea spillover the more left skewed the generated citations profile.

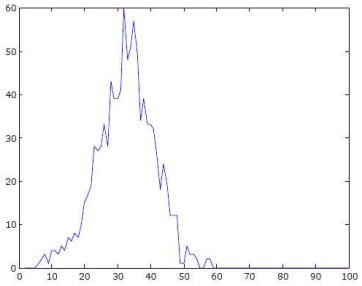
[1]The Matlab m-files used to generate these figures are available from the authors upon request.



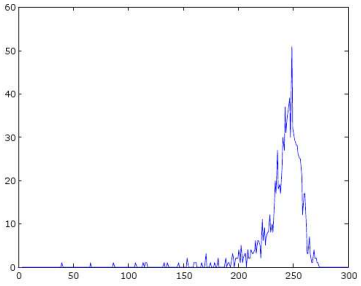
(a) Very Simple Idea: $a_i = .1$



(b) Simple Idea: $a_i = .5$



(c) Complex Idea: $a_i = 1$



(d) Highly Complex Idea: $a_i = 1.5$

Figure 1: Simulation Outcomes

2.1 Deterministic Approximation

If a mass $L_{Ri} > 0$ of research labor are independently researching new applications of idea i, the law of large numbers implies that the flow per unit time of new applications, $\dot{x}_i(t)$, behaves deterministically according to the following nonlinear ordinary differential equation:

$$\dot{x}_i(t) = \frac{1}{X_i^{1+a_i}} L_{Ri}\beta_i \left[X_i - x_i(t)\right] x_i(t)^{a_i}. \tag{2}$$

Notice that, as usual in economics, we have approximated a discrete variable - the (integer) "number of discovered applications of idea i as of time t" - with a continuous real function of time, $x_i(\cdot)$, to be read as the "mass of discovered applications of idea i as of time t". In this interpretation, X_i denotes the maximum mass of potential applications of idea i.

Dropping time indexes, twice differentiating (2) leads us to the following:

Lemma 1 *If $a_i \in [0,1]$, $\dot{x}_i$ is a strictly concave function of x_i for all $x_i \in [0, X_i]$. If instead $a_i > 1$, $\dot{x}_i$ is strictly convex for all $x_i \in [0, \frac{a_i-1}{a_i+1}X_i]$, whereas it is strictly concave for all $x_i > \frac{a_i-1}{a_i+1}X_i$. Moreover, for all $a_i > 0$, $\dot{x}_i$ is an increasing function of x_i for all $x_i \in [0, \frac{a_i}{a_i+1}X_i]$, and it is decreasing in x_i for all $x_i > \frac{a_i}{a_i+1}X_i$. Hence $\dot{x}_i$ is maximized by $x_i = \frac{a_i}{a_i+1}X_i$, where it is equal to:*

$$\max_{x_i \in [0,X_i]} \dot{x}_i = L_{Ri}\beta_i a_i^{a_i} \left(a_i + 1\right)^{-(1+a_i)}. \tag{3}$$

It is important to note that the marginal productivity of R&D labor, $MPL_{Ri} = \beta_i \left[1 - \frac{x_i}{X_i}\right] \left(\frac{x_i}{X_i}\right)^{a_i}$, implied by (2) satisfies, at any date $t > 0$, the following properties:

1. Given the number of already invented applications, MPL_{Ri} is constant in the number of R&D workers searching for a new application of a given idea i.

2. Given the fraction $\frac{x_i}{X_i}$ of already invented applications, MPL_{Ri} is decreasing in a_i.

3. The higher a_i, the closer to 1 the fraction of discovered applications on total potential applications $\frac{x_i}{X_i} = \frac{a_i}{a_i+1}$ that maximizes the per unit time flow, $\dot{x}_i$, of new applications.

Our assumed law of motion can be tested by observing the density of the distribution of the patents that cite a given initial patent: according to our theory it should have a bell form. Moreover, point 2 suggests a new

way of testing increasing complexity: if over time the mode of the density of the distribution of the patents citing a given idea/patent moves toward the upper tail of the distribution, then we can conclude that complexity has been increasing. This can be easily measured by considering the skewness of the distribution of observed citations.

3 Data, Methodology and Results

In this section we employ the patent forward citations as a measure of the R&D spillover of patents, using NBER patent citations dataset (1963-1999, see [2]). It is important to remark that after a patent has been granted it generates an observable realization of consequences determined by two principal kinds of elements:

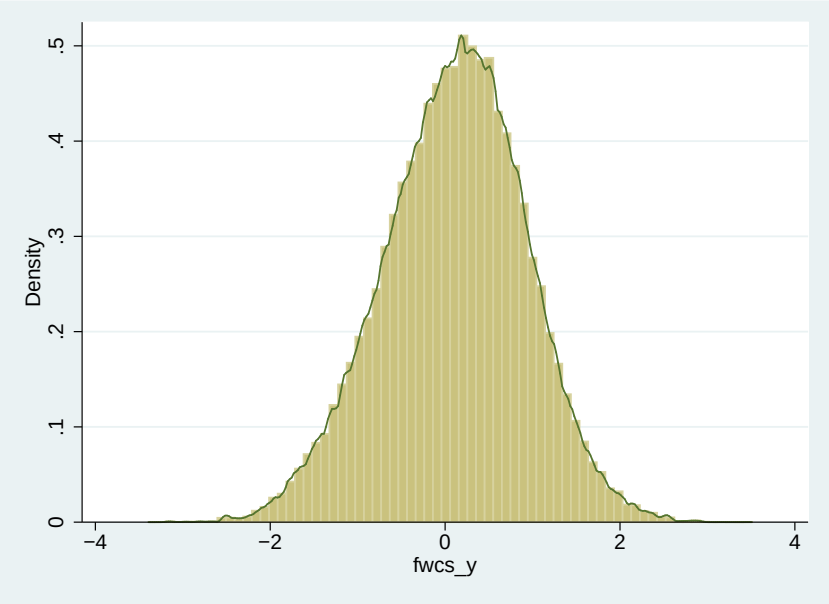
- (Endogenous) The frequency and the duration of the forward citations (forward citations lags). This is the patent scientific relevance, i.e. the more important, the more it is cited[2];
- (Exogenous) The limited length of the patent dataset that produces truncation problem in the dataset;

The second item has a twofold effect on our data, because the spillover-life has two principal components: the first is the number of each patent forward citations, the second is the mean forward citations lag. Clearly, the older is the patent application, the higher level of mean forward citations lag it will have; moreover the older the granted patent, the higher the number of citations that it received. Once we solved the second problem, normalizing each data for the yearly mean of citations received (see [2]), we standardize, normally, the mean forward citations lag, to compare, in terms of patent-life, our data from 1963 to 1999.

Once the patent features are standardized, we can observe the spillover life of each patent, by industry and year. These statistics are crucial to empirically evaluate the theory laid down in Section 2: in fact we can compare the complexity of each ideas and, as a consequence, the number of fertile spillovers by analyzing the distribution of citation received. When a patent is granted - or filed, as in European Patent Office (EPO) case - each inventor

[2]See, among others, [4] and [6]).

Figure 2: Total Frequency Distribution of Citations received by patent applied in 1965-1990



can study on the new information embodied in this document; thus it gives a number of spillovers over a number of years, measurable at the end of what we may call 'direct spillover life'. Thus, each inventor can utilize the patent documentation as fundamental source for his/her new ideas[3]. But, after a new inventor files her/his application, based on the previous patent, to the US patent office[4], any other inventor examining the new patented idea observes, as backward citation, the previous patent and thus, in turn, can cite it in her/his own future patent(s). In this way an increasing cumulative spillover dynamics follows that stops once the old patent finishes its spillover potential. The forward citations frequency distribution in Figure 3 interprets and confirms our idea.

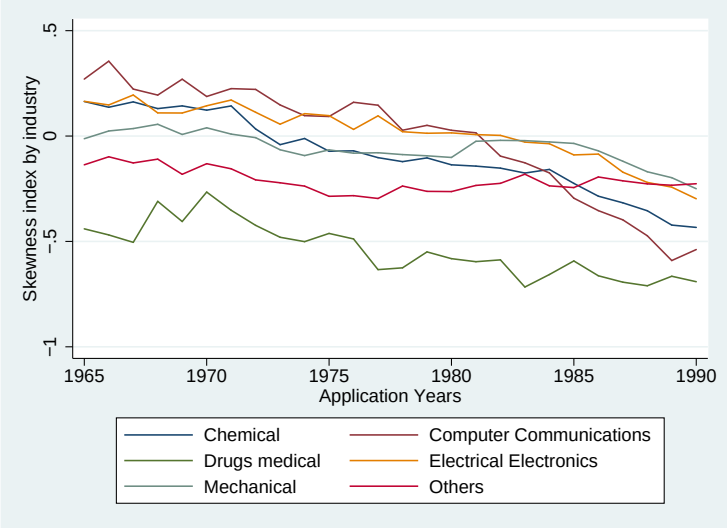
In this figure, the left tail of the distribution measures on the y axis the frequency of citations received by each patent, and on x axis the forward citation lags[5]. This part of the graph depicts the relative number of 'fertile'

[3]See [1] and [5]. on the relevance of 'patent literature' source in the invention process.

[4]For the EPO there exist differences in terms of backward citations (see, among others [6]).

[5]Labelled $fwcs\_y$, standardized as we said above.

Figure 3: Skewness index by Industry and Year



spillovers given by a patented idea; conversely, the right tail (decreasing frequencies) shows the 'non-fertile' spillovers. Thus, using the relative number of citations received by each patent in the left tail, we can give an approximation of its quality. At the same time, we can evaluate the complexity of this patented idea. In fact, the higher the relative number of citations received in the first years (left tail), the more the inventions that have been using this patented idea as a source, without exhausting it too much. Thus, using the skewness Pearson index[6], we evaluate these concepts (see figure 3 and table 1).

In each industry the historical trend of the skewness index is decreasing: this suggests - according to our theoretical model - an increasing complexity of the innovative activity, offering support to the assumptions of the semi-exogenous growth theory (Jones, 2005).

Notice that our theory suggests that the Drugs and Medical industry is characterized by the highest complexity whereas the - relatively younger - Computer and Software industry seems the least complex.

[6]We have used the typical version of Pearson's skewness index: $S_k = 3(mean - median)/(\sigma)$

Table 1: Skewness Index by Industry (1965-1990)

Sector	**Skewness Value**
Chemical	-0.08
Computer Communications	-0.12
Drugs medical	-0.53
Electrical Electronics	0
Mechanical	-0.055
Others	-0.21

4 Conclusions

This paper has suggested a model for the mechanics of the production of ideas by means of ideas. Unlike the existing literature on technological spillovers, we focus on the direct citations of each idea. The citations of an idea are viewed as the direct applications of it in the discovery of new ideas. We assumed that the probability of finding new direct implications of an idea is higher the higher the share of its undiscovered applications and the higher the relative experience in finding them. Our model replicates the bell curve profiles of the patent forward citations, as extracted from the US data. Our theory also allows to relate the complexity of innovation to the shape of the citations profile. In particular, the more skewed to the left the citations profile the higher the complexity of innovation. By aggregating patents by U.S. industries according to their Pearson skewness indexes, we suggest that complexity has constantly been increasing over the 1963-1999 period.

References

[1] Grossman, M. and Helpman, E. (1991), *Innovation and Growth in the Global economy*, Cambridge, MA: MIT press;

[2] Hall B.H., Jaffe A. B., Trajtenberg M. (2001), The *NBER* Patent Citations Data File: Lessons, Insights And Methodological Tools, *NBER Working Paper* 8498;

[3] Jones, C. (2005), Growth and Ideas, *Handbook of Economic Growth*, in P. Aghion and S. Durlauf (ed.s) *Handbook of Economic Growth*, Elsevier.

[4] Schankermann, P. and Lanjouw (2004), Patent Quality and Research Productivity: Measuring Innovation with Multiple Indicators, *The Economic Journal*. Vol. 114, pp. 441-65.

[5] Schettino F. and Sterlacchini A. (2007), European Patenting and the Size of Inventors, *Working Papers 308*, Università Politecnica delle Marche (I), Dipartimento di Economia.

[6] Van Pottelsberghe de la Potterie, B. and Van Zeebroeck (2007), A Brief History of Space and Time: the Scope-Year Index as a Patent Value Indicator Based on Families and Renewals, *CEPR Discussion Papers 6321*, C.E.P.R. Discussion Papers. 1290-1310.