Darwinian evolution of culture as reflected in patent records

Andrew Buchanan,[1] Norman Packard,[2] and Mark Bedau[1,2,3*]
[1]Reed College, Portland
[2]ProtoLife Inc, San Francisco
[3]Initiative for Science, Society, and Policy, University of Southern Denmark, Odense
[*]Corresponding author: mab@reed.edu

Abstract

We argue that culture undergoes an evolutionary process, analogous to biological evolution. As evidence, we analyze the first page of all the utility patents issued in the United States over the past thirty three years, which comprise over three million patents. The set of issued patents is regarded as an evolving population. A patent is considered to "reproduce" when it is cited by a new patent, and variability is introduced into the population by the innovations in new patents. When we analyze patent records with statistics that quantify the degree to which the population of patents is shaped by natural selection, and so undergoes Darwinian evolution, we find convincing evidence for Darwinian evolution. Further, we observe that, according to our statistics, the most fit patents cover "door-opening" technologies, which are technologies that enable a broad range of further innovations.

Introduction

We study the evolution of technology as reflected in US patent records. Everyone agrees that technology evolves, but there is controversy about what this means, and especially whether the evolution of technology is "Darwinian" in some interesting sense. By Darwinian evolution, here, we mean that the process of natural selection in a population is a significant factor in explaining how the traits in the population change over time. Natural selection, in turn, is defined as the process by which heritable traits that make members of a population more likely to survive and reproduce tend to spread through the population over time.

It should be noted that our conception of Darwinian evolution is consistent with cultural evolution being simultaneously significantly shaped by many non-Darwinian mechanisms, like random genetic drift, pleiotropy, and epigenesis.

In this paper, we develop methods to address the following two questions:

1. Does natural selection shape the evolution of technology?
2. If so, what kinds of technological innovations especially drive its evolution?

Our aim is both to show the value of the methods, even when applied in new settings and adapted to new contexts, and also to investigate and learn from the first fruits of applying the methods to patent data. In the end, our conclusions will be two: (1) Natural selection does significantly shape the evolution of patented technology, and (2) the statistical evidence corroborates the hypothesis that so-called "door-opening" technologies have been especially important drivers of the evolution of technology.

Patent data

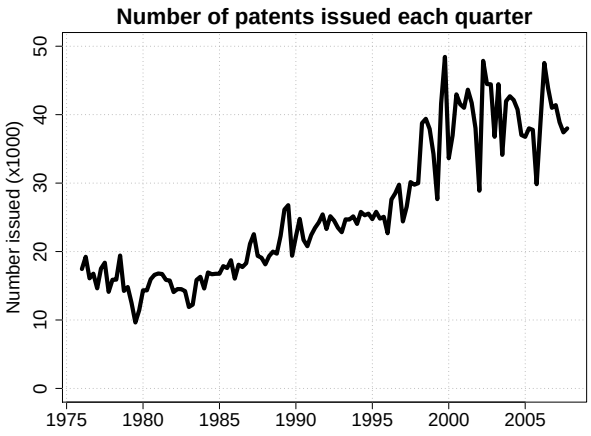

Figure 1: Number of patents issued each quarter, over the thirty three years in our database.

The patent data we mine in this experiment consists records of US patents issued over the last thirty three years, 1976-2009. Figure 1 shows that the rate at which patents have been issued has doubled over the past thirty years.

In this study we focus only on a few key pieces of information in the patent record: patent number, title, issue date, abstract, IPC codes, and references. We use the patent number as a unique identifier for each patent, which is stamped with an issue date.

USPTO patent examiners assign each US patent a handful of IPC codes, designed to classify the invention. In this paper we use IPC codes to measure the degree of similarity and dissimilarity between two inventions.

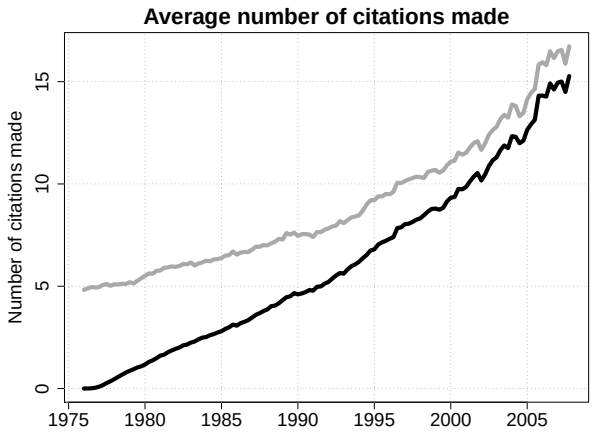

Figure 2: Average number of citations made per quarter; upper curve includes all citations made, lower curve includes only citations made to patents within our database.

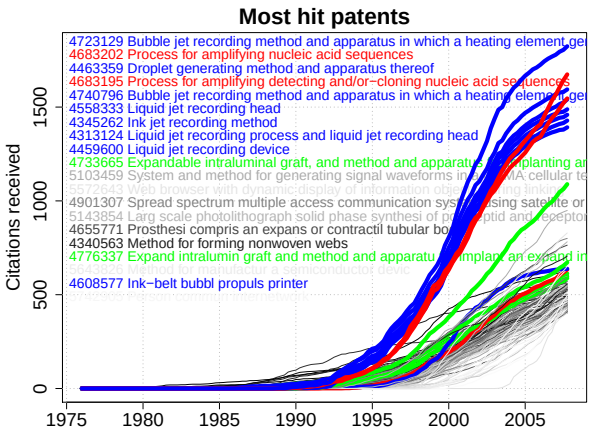

Figure 3: The cumulative number of citations as a function of time. Each curve represents citations accumulated by a particular patent. Only the top 400 patents are shown, ranked by total number of citations that a patent accumulates in the database. The patent numbers and titles of most heavily cited patents are listed, in the same color as the curve for that patent.

The title and abstract are texts that succinctly describe the invention. The central purpose of this paper is to display new ways to mine these texts to visualize and quantify the evolution of technology.

Each patent record is required by the USPTO to cite all of the previous inventions on which it depends. These citations establish an intention's "prior art" and they are added by patent examiners when necessary. Figure 2 shows a three-fold rise in the average number of citations each patent makes. Citations play a pivotal role in our evolutionary analysis of the patent data. We develop a precise formalism for key statistics about citations, and visualize the evolution of technology by highlighting the most heavily cited inventions.

Evolutionary activity

We regard the evolutionary activity of a patent as the cumulative summation of the number of times it is cited. For patent p, $c^t(p)$ is defined as the set of patents issued at time t that cite p, and the cumulative citations of patent p up to time t:

$$C_p^t = \sum_{t'=0}^{t'=t} \sum_{p' \in c^{t'}(p)} f^t(p, p'), \qquad (1)$$

where $f^t(p, p')$ is a counting function, constructed to count contributions of citations to the cumulative sum. The simplest version of a counting function is $f^t(p, p') \equiv 1$, in which case each citation in $c^t(p)$ is counted with equal weight. For this case, C_p^t is illustrated in Figure 3. The counting function $f^t(p, p')$ may be crafted to emphasize or de-emphasize different aspects of the population, as discussed below.

In Figure 3, we overlay the patent number and title (stemmed) for the twenty most heavily cited patents in our data set. In this and all the following plots, we color the cumulative citation waves for certain kinds of patents, as follows: Blue patents concern inkjet printing, red patents concern PCR, and green patents concern stents. These three kinds of technology cover all of the ten most heavily cited patents in our data. Later in this paper we test a hypothesis about why these three kinds of technology proved to be so fecund.

The average behavior of C_p^t, obtained by averaging over all patents issued at each new time t (the time resolution of the data is quarterly), is illustrated in Figure 4. Notice that the curves are roughly straight lines, indicating that patents continue to receive citations at roughly the same rate over their life in the database. Notice also that the slope of the lines increases through the first two decades of in our data, and then it levels off.

Shadow models

In order to determine which aspects of the patent data might be shaped by natural selection, we construct a "shadow patent" system. Shadow patents and real patents exhibit many of the same statistics, by construction. If a real patent is issued, then so is a shadow patent, and if a real patent cites an earlier patent, then so does a shadow patent. Thus, by construction, Figures 1 and 2 are identical for real and shadow patents.

However, the same does not necessarily hold for Figure 3. When shadow patents choose *which* patents to cite, they do so *randomly* and with equal probability from the pool of earlier patents. To test the hypothesis that heavily cited real patents are heavily cited just by chance (given the number

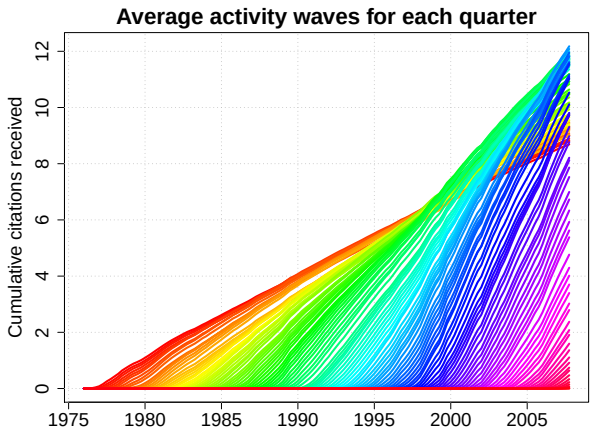

Figure 4: Average number of citations per quarter. Each curve represents the cumulative sum of the citations received of all patents issued in a given quarter.

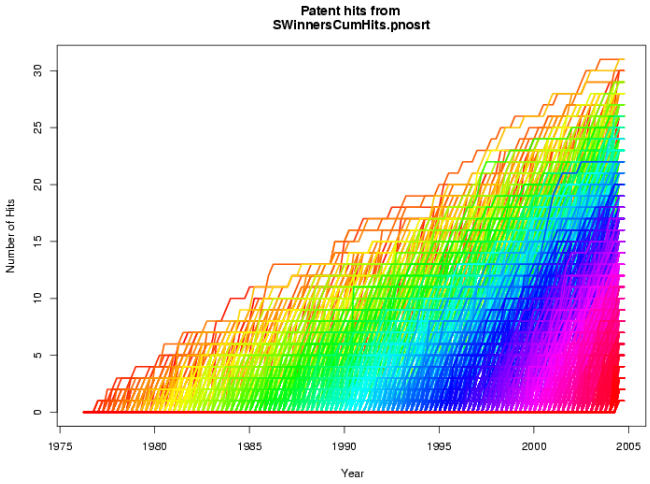

Figure 5: Waves of the most heavily cited patents issued each quarter in a shadow patent model (see text).

of patents being issued and the number of citations being made), we simulate the shadow patents and observe typical maximal citation levels. If the most cited real patents have an order of magnitude more citations that the most cited shadow patent, that implies that high citation counts are no statistical fluctuation of a random shadow citation process.

Figure 5 shows the cumulative citations of the most heavily cited shadow patents issued each quarter. Comparison of the y-axis in Figures 3 and 5 shows that heavily cited real patents get orders of magnitude more citations than any shadow patent. We conclude that the fecundity of heavily cited patents strongly indicates that the patent's specific content explains why it is so heavily cited.

Star patents

The significant rise of evolutionary activity, measured by raw cumulative citation counts C_p^t, over shadow model activity is itself evidence of the process of Darwinian evolution, driven by selection of the fittest.

Further insight may be gained by examining particular high-fitness patents, to create narratives that may contribute to our intuition about the evolutionary process. Studying the patents in Figure 3 reveals that the most heavily cited patents typically involve one of the following three innovations: ink-yet printing, PCR, and stents.

Ink-jet printing. Although originally developed for putting ink on paper, the fundamental innovation behind ink-jet printing actually involves the ability to extremely precisely position extremely small bits of matter ("ink"). Beside traditional inks, the materials printed include skin cells (so skin grafts can be printed), DNA or RNA primers (on microarray chips), and metals. Depositing successive layers of materials means that we can print certain arbitrary three dimensional structures. One now reads about ink-jet printing technology being used to print batteries, clocks, flexible video screens, among other things.

PCR. PCR, or the polymerase chain reaction, is one of the foundations of contemporary biotechnology. Patented in 1987 by Kary Mullis of Cetus Corporation (one of the first biotech firms), PCR makes it possible to rapidly make millions of copies of an arbitrary DNA sequence. This method has been extensively modified to achieve many different kinds of genetic manipulations. In 1993 Mullis received the Nobel Prize in Chemistry for his work on PCR.

Stents. Stents are man-made tubes that are used to hold open conduits in the body, such as coronary arteries partially occluded with plaque. In 1986 Julio Palmaz patented a stent that could be expanded within a blood vessel by an angioplasty balloon that inserted into a blood vessel. This procedure allows some blocked coronary arteries to be repaired without open-heart surgery, involving much simpler and safer procedure. Stents have been in the news recently because of a thicket of patent litigation between two healthcare giants, Boston Scientific and Johnson and Johnson, and because of recent controversy about the merits of drug-coated stents.

Eliminating artifacts

Definition of the evolutionary activity in terms of the raw cumulative citation counts C_p^t as described above may suffer from artifacts in the data that are not related to evolutionary selection of the fittest, which effect evolutionary activity aims to capture. This leads to variations in the definition of activity, obtained by modifying C_p^t to counter these effects through a process of normalization.

A simple example of such an artifact is evident from Figure 2, in which the number of citations grows with time. This fact alone would lead us to expect that patents issued later would accumulate citations more rapidly than patents

issued earlier. A simple normalization to adjust for this effect uses the counting function

$$f^t_{\rm prior} = \frac{N^t}{C^t},$$

where N^t is the total number of patents issued up to time t, and $C^t = \sum_p C^t_p$ is the cumulative total number of citations, so that the weight of a citation is boosted if there are more patents it could cite, and lowered if there are more total citations. Then, the adjusted cumulative citation sum, $C^t_{{\rm prior}\ p}$, is computed from equation (1) using $f^t(p,p') \equiv f^t_{\rm prior}$.

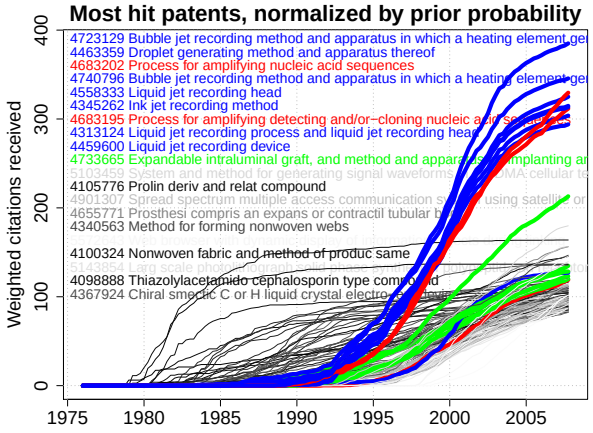

Figure 6: Normalization by prior expected probability of being cited, due to changes in (i) the number of patents that could be cited, and (ii) the number of citations that are being given. Both (i) and (ii) change over our data base by a factor of three (data not shown). The artifact of IPC citing rate does not substantially change the top stories (rank correlation XXX).

The dynamics of $C^t_{{\rm prior}\ p}$ is illustrated in Figure 6. Notice that this normalization significantly boosts the citation counts for earlier patents, as expected. Notice also that the same ten patents involving inkjet printing, PCR, and stents still occupy the top ten positions in the graph. Thus, although normalizing by prior expected probability of being cited does significantly change which patents are judged to be technology stars, the narrative of technology evolution being most strongly driven by innovation in inkjet printing, PCR, and stents.

Different IPC classifications are known to have average citation rates that vary by orders of magnitude. These skewed IPC citation distributions might be thought to create further artifacts in our cumulative citation statistics. We can test that hypothesis by introducing a new counting function, $f_{\rm IPC}$, to normalize by the total number of citations by patents in a given category.

The IPC classification of a patent has five levels, $I(p) = (c_1, ...c_5)$, where each c_i may be thought of as an integer labeling different categories. So, to define the new counting function, we first define the categories of interest to be all possible values of the first two category coordinates, $\mathbf{c} = (c_1, c_2)$. The total number of patents in a given category is

$$N^t_{\mathbf{c}} = \sum_p \delta(c_1 - I(p)_1)\delta(c_2 - I(p)_2),$$

and the the total number of citations in the category is

$$C^t_{\mathbf{c}} = \sum_p \sum_{p' \in c^t(p)} \delta(c_1 - I(p')_1)\delta(c_2 - I(p')_2),$$

where $\delta(x) = 1$ if $x = 0$ and 0 otherwise. So we can define $f_{\rm IPC}$ to be a function that depends only on the citing patents:

$$f^t_{\rm IPC}(p') = \sum_{\mathbf{c}} \frac{N^t_{\mathbf{c}}}{C^t_{\mathbf{c}}}\delta(c_1 - I(p')_1)\delta(c_2 - I(p')_2).$$

E.g., a patent in category A1 issued in 1990 has it's outgoing citations halved in weight because A1 patents issued in 1990 made twice as many citations on average as A1 patents from 1976 (chosen as the arbitrary baseline rate). Also, the outgoing citations of a patent in category C1 issued in 1990 count for less, because those patents made more citations on average than A1 patents from 1976.

Figure 7 shows a plot of $C^t_{{\rm IPC}\ p}$, defined by equation 1, with $f^t(p,p') \equiv f^t_{\rm IPC}(p')$. This figure shows that the skewed IPC citation distribution strongly affects the cumulative citation values. Comparison with Figure 6 shows that the cumulative citations for PCR (red) patents have been significantly raised, while those for inkjet printing (blue) have been significantly lowered, as have stent patents (green). Nevertheless, those same three narratives still play a dominant role in driving technological innovations.

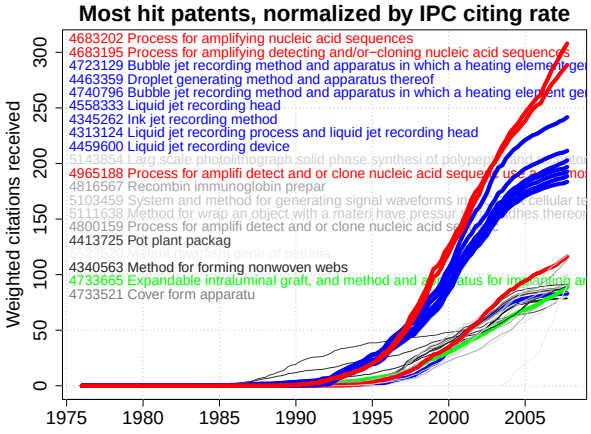

Figure 7: Normalization by IPC citing rate: hits are weighted based on the average number of citations made by patents in that (level 2) category at that time.

Another important effect present in the data is that some patents are cited by subsequent patents that are closely related, and that often have the same assignee. We refer to this as "self-citation" because of the effective redundancy. This effect is not surprising; if a company makes an innovation, it is motivated to build on that innovation and to patent related further developments. The effect may, however skew the citation statistics, creating an artificially large citation count for patents followed by a rash of followup patents. A simple normalization to adjust for this effect uses the following counting function that discounts self-citations:

$$f_{\text{self}}(p, p') = \alpha,$$

when p and p' have the same assignee, and $(p, p') = 1$ otherwise, with $\alpha < 1$. Then, the adjusted cumulative citation sum, $C^t_{\text{self}(p,p')}$, is computed from equation (1) using $f^t(p, p') \equiv (N^t/C^t) f_{\text{self}}(p, p')$, where we have include normalization with respect to overall size of the patent pool and the set of citations, as described above for f^t_{prior}.

Figure 8 shows a plot of $C^t_{\text{self}}(p, p')$ for $\alpha = 0.25$ (other values of α produce similar results). This normalization dramatically reshuffles the relative impact of the top patents. One dramatic effect is the drop in inkjet printing patents (blue). It turns out that the vast majority of those patents are assigned to the same company, Canon, and subsequent Canon inkjet printing patents heavily cite earlier Canon inkjet printing patents. By contrast, discounting self-citations especially boosts stent patents, increasing their representation in the top twenty patents by 500%.

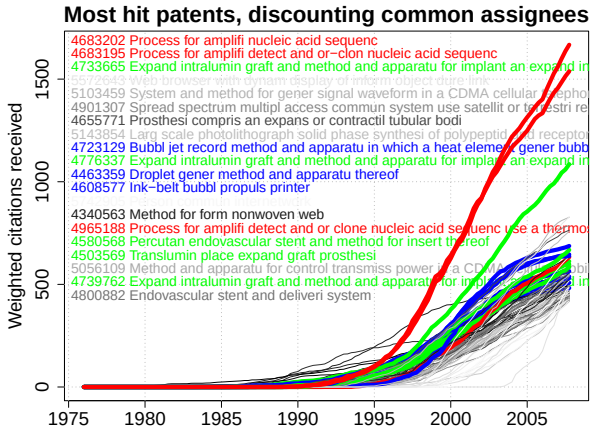

Figure 8: Discounting for self-citations. Notice that the ranking of star patents significantly changes, but PCR (red), ink-jet printing (blue), and stents (green) remain stars.

We may also combine any or all these normalizations, aiming to obtain the cleanest possible picture of which technologies most strongly drive innovation in the evolution of technology. When we do so, we see that the three top stories (PCR, ink-jet printing, and stents) remain dominant among the most fecund technologies. It is striking that, while our efforts to reduce artifacts in cumulative citation counts does significantly change the relative ranking of our stories, the same three stories consistently remain significant. This similarity may be quantified by computing the rank correlation. A scatter plot of ranks for different normalizations is shown in figure 9, and the table of rank correlations is given below.

Figure 9: XXX need figure of rank scatter plot. XXX also table of rank correlations XXX

Door-opening innovations

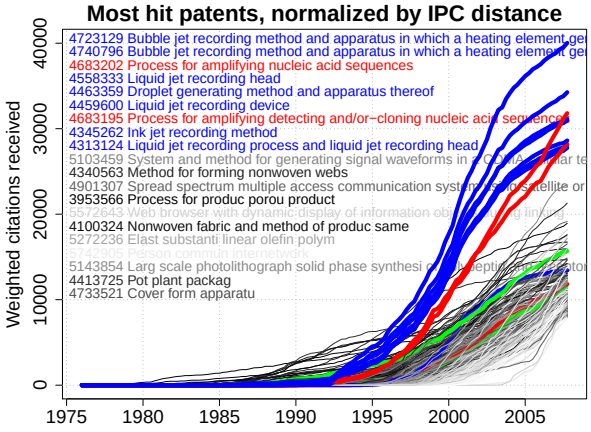

Figure 10: Weighting citation counts by the exponential of IPC distance, so that citations by patents in distant IPC categories count much more. This rewards door-opening innovations and penalize innovations that merely spur further innovations of the same type.

A crucial aspect of biological evolution seems to be the ability of biological innovations to "open doors" to entire new universes of innovation, e.g., through the creation of new modes of interaction and new ecological niches, on all scales from molecular to macro-population. We may ask if such phenomena are present in the evolution of the patent population.

To formulate the question quantitatively, we use IPC categories to quantify the evolutionary impact of a patent in terms of the breadth of different kinds of patents that cite it. The intuition is that if a patent is cited by patents from very similar IPC categories, then it has relatively narrow impact. By contrast, if a patent that is cited by patents in radically different IPC categories, then is has a much broader impact and is opening doors to more kinds of innovations. This intuition may be quantified by weighting the citation count more heavily for more distant IPC categories.

Specifically, if $I(p)$ is the IPC vector $(c_1, ..., c_5)$, with c_1 being the coarsest grain IPC resolution, and c_5 being the

finest grain resolution, we define the IPC distance between two patents as

$$d_{\mathrm{IPC}} = 5 - n_{\mathrm{IPC}},$$

where n_{IPC} is the maximum integer such that $I(p)_i = I(p')_i$ for all $i \leq n_{\mathrm{IPC}}$. Then we may create a counting function that weights by this distance, exponentiating it to emphasize the effect:

$$f^{t'}_{\mathrm{EIPC}}(p, p') = 1 + 2^{d_{\mathrm{IPC}}}.$$

Now, we can compute $C^t_{\mathrm{EIPC}\ p}$ from equation (1), using $f^t(p, p') \equiv f^t_{\mathrm{EIPC}}(p, p')$.

A plot of $C^t_{\mathrm{EIPC}\ p}$ is shown in Figure 10. Note that PCR, inkjet printing and stents remain significant innovations, indicating that they are all likely to be door-opening innovations. (If they were not, then weighting by IPC distance would drastically lower the relative citation levels of PCR, inkjet printing, and stent patents. But instead those patents remain stars. So, they must be door-opening innovations.

Conclusion

Our results show that technology undergoes a Darwinian evolutionary process, analogous to biological evolution. The set of issued patents can be viewed as an evolving population of "organisms" that reproduce when they are cited by later inventions. In the end, we can treat an invention fecundity (number of hits) as its fitness, for its fecundity directly measures the patent's impact on the composition of future populations. The dramatically higher citation counts for the most cited real versus shadow patents shows that high fecundity cannot be explained merely as a statistical fluctuation of a process that cites patents irrespective of their content. This comparison with a no-selection null hypothesis is convincing evidence for Darwinian evolution of technology.

Further, we have adapted our statistics to highlight "door-opening" technologies, i.e., those that enable a broad range of further kinds innovations. It turns out that these statistics corroborate the hypothesis that the patent stars are door-opening technologies.

Acknowledgements

Thanks to Devin Chalmers, Cooper Francis, and Noah Pepper for stimulating discussions about how to quantify the evolution of technology.