Darwinian evolution of culture as reflected in patent records

Andrew Buchanan,[1] Norman Packard,[2] and Mark Bedau[1,2,3*]
[1]Center for Advanced Computation, 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 bibliographic information of all utility patents issued in the United States from 1976 through 2007, 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. We analyze patent records with statistics that quantify the degree to which the population of patents is shaped by natural selection, and we find convincing evidence of Darwinian evolution. Further, weighting our statistics by the classification distance between parent and child shows that the most fecund patents are "door-opening" technologies that enable an especially 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 (Jablonka (2002); Benzon (1996)). 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 be increasingly represented in 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 (Jablonka and Lamb (2005); Sperber (1996)).

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 significantly shapes 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.

Our project applies earlier work on evolutionary activity statistics (Bedau and Packard (1992); Bedau et al. (1997, 1998); Bedau and Brown (1999); Rechtsteiner and Bedau (1999); Raven and Bedau (2003)) and significantly expands and develops an earlier similar pilot project (Skusa and Bedau (2002); Bedau (2003)).

Patent data

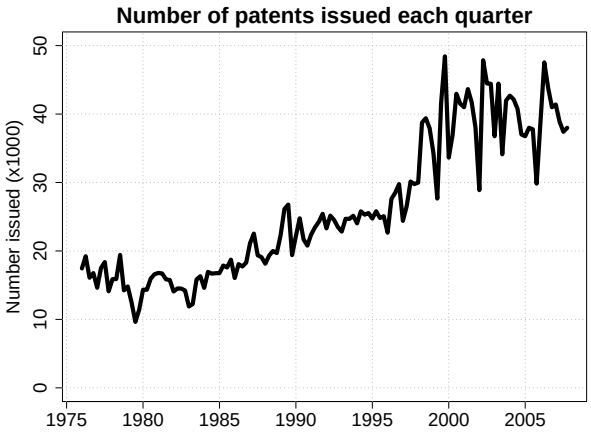

Figure 1: Number of patents issued each quarter, over the thirty years in our database.

The patent data we mine in this experiment consists of records of US patents issued over thirty years from 1976 through 2007. Figure 1 shows that the rate at which patents have been issued has doubled over the past thirty years.

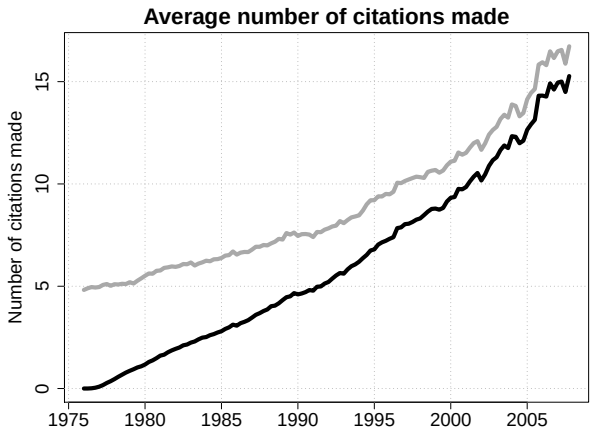

Figure 2: Average number of citations made per quarter; upper curve includes all citations made, lower curve includes only citations made to patents within our dataset.

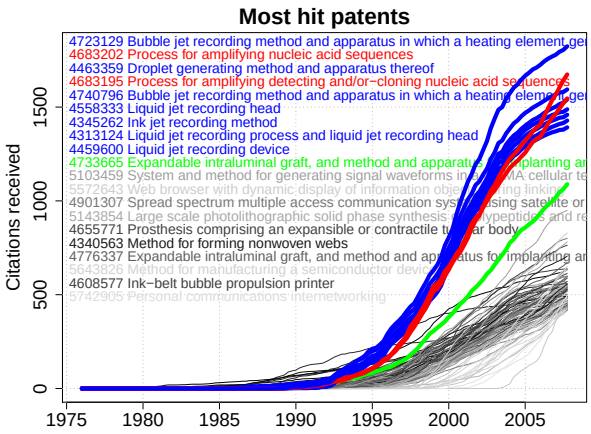

Figure 3: The cumulative number of citations as a function of time. Each curve represents citations accumulated by a particular patent. Only the top 100 patents are shown. Patent numbers and titles are printed in the same color as the corresponding citation curve.

In this study we focus only on a few key pieces of information in the patent record: patent number, title, issue date, IPC code, and references. The patent number serves by design as a unique identifier for each patent and we use it as such.

Each US patent is assigned a handful of IPC codes by the inventor and patent examiners at the USPTO, designed to classify the invention. In this paper we use IPC codes to measure the degree of similarity and dissimilarity between two inventions. The IPC codes are also used to control for differences in citation practices in diverse technical fields.

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 are compiled by both patent examiners at the USPTO in and the inventor. Figure 2 shows a three-fold rise in the average number of citations each patent makes over the past thirty years. 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 number of times other patents cite it. For patent p, $c^t(p)$ is defined as the set of patents issued at time t that cite p, and C_p^t as the cumulative citations to patent p up to 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 for the twenty most heavily cited patents in our dataset. In this and all subsequent plots, we color the citation waves as follows: Top inkjet printing patents are blue, top polymerase chain reaction (PCR) patents are red, and the top stents patent is green. All other patents are colored various shades of gray. We focus on inkjet printing, PCR, and stents because all of the ten most heavily cited patents in Figure 3, by a significant margin, are innovations in one of those three areas of technology. Later in this paper we consider what makes those three technologies so fecund.

The average behavior of C_p^t, obtained by averaging over all patents issued at each new time t is illustrated in Figure 4 (the time resolution is quarterly). 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 slopes of the lines increase through the first two decades of in our data and then level 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 makes a citation, then so does a shadow patent. Thus, by construction, Figures 1 and 2 are identical for real and

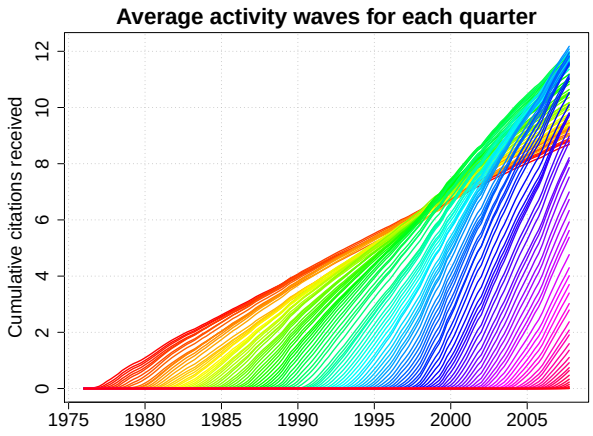

Figure 4: Average number of citations per quarter. Each curve represents the cumulative sum of the citations received of all patents issued in a given quarter.

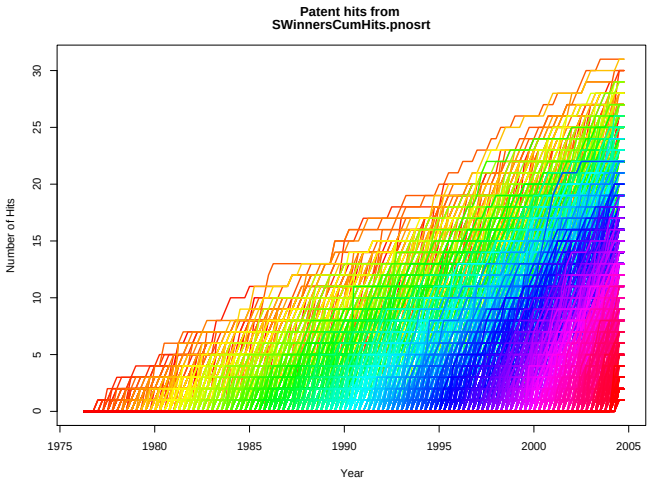

Figure 5: The cumulative number of citations of the most heavily cited patents issued each quarter in a shadow patent model (see text).

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 of patents being issued and the number of citations being made), we simulate shadow patents and observe typical maximal citation levels. If the most cited real patents have significantly more citations that the most cited shadow patent, then the real citation levels are not statistical fluctuations.

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 striking fecundity of heavily cited patents is no accident. It is not mere noise. Rather, there must be something special about the meaning or content of heavily cited patents that makes them so fecund.

Super 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: inkjet printing, PCR, and stents.

Inkjet printing: The Japanese company, Canon, holds a spate of patents on inkjet printing that have been very heavily cited. Although originally developed for putting ink on paper, the fundamental innovation behind inkjet printing actually involves the ability to extremely precisely position extremely small bits of matter ("ink"). Beside traditional inks, for the original printing applications, the printed materials now also 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 inkjet printing technology being used to print batteries, clocks and flexible video screens, among other things.

PCR: Polymerase chain reaction is one of the cornerstones of contemporary biotechnology. Patented (number 4683202) 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. It is now a fundamental tool in a wide range of biotech applications. 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 inserted angioplasty balloon. This procedure allows some blocked coronary arteries to be repaired without open-heart surgery, allowing much simpler and safer treatment. Citations to this patent indicate that it opened the door to a wide range of minimally invasive blood vessel therapies. Stents have been in the news recently because of patent litigation between

Boston Scientific and Johnson and Johnson, and because of controversy about the merits of drug-coated stents.

Eliminating data biases and artifacts

The definition of 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. The canonical way in which C_p^t will be modified is through the definition of the counting function $f^t(p, p')$. We will see how modified counting functions will enable biases and artifacts to be compensated for explicitly. Generally, these modifications may contain a parameter that must be chosen for a certain level of compensation; for this reason these modified counting functions may be regarded as heuristic, rather than fundamental.

A simple example of such an artifact is evident from Figure 2, in which the number of citations grows with time. This leads us to expect that patents issued later would accumulate citations more rapidly than patents issued earlier. Patents are more likely to cite (relatively) recent patents, and over time the number of citations made increases, thus favoring later patents.

A normalization to adjust for this effect uses the counting function

$$f^t_{\text{rate}} = \frac{R^{t'}/N^{t'}}{R^t/N^t}, \tag{2}$$

where N^t is the total number of patents issued at time t, and R^t is the total number of citations made by patents issued at t, and t' is the (arbitrary) baseline time point in the dataset. The total number of citations made must be equal to the total received so $\sum_t \sum_p R_p^t = \sum_t \sum_p C_p^t$. The effect of this normalization is to value all citations in terms of the baseline citation rate, similar to adjusting historical prices for inflation. Because patents at the beginning of the dataset make one third as many citations as those at the end, their citations are given three times as much weight. Then, the adjusted cumulative citation sum, $C^t_{\text{rate }p}$, is computed from equation (1) using $f^t(p, p') \equiv f^t_{\text{rate}}$.

The dynamics of $C^t_{\text{rate }p}$ 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 super 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

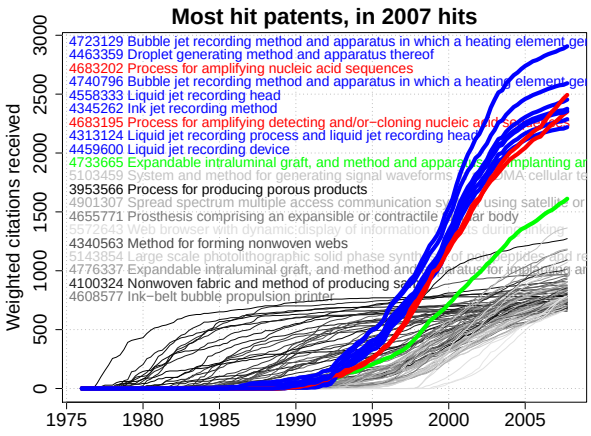

Figure 6: Normalization by relative rate of citation due to changes in the number of citations that are being given over time. Activity is valued in terms of most recent citation rates.

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_{IPC}, to normalize by the mean number of citations made 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 citations made by patents in the category at time t is

$$R^t_{\mathbf{c}} = \sum_{p' \in p} r(p')\delta(c_1 - I(p')_1)\delta(c_2 - I(p')_2),$$

where $\delta(x) = 1$ if $x = 0$ and 0 otherwise and $r(p')$ is the number of citations made by p'. So we can define f_{IPC} to be a function that depends only on the citing patent:

$$f^t_{\text{IPC}}(p') = \sum_{\mathbf{c}} \frac{R^{t'}_{\mathbf{c}}/N^{t'}_{\mathbf{c}}}{R^t_{\mathbf{c}}/N^t_{\mathbf{c}}}\delta(c_1 - I(p')_1)\delta(c_2 - I(p')_2). \tag{3}$$

E.g., a patent in category A01 issued in 1976 has its outgoing citations doubled in weight because A01 patents issued in 1976 made half as many citations on average as B02 patents from 2007 (chosen as the arbitrary baseline rate). In this way the contributions to evolutionary activity of different categories and different times are equalized.

Figure 7 shows a plot of $C^t_{\text{IPC }p}$, defined by equation (1), with $f^t(p, p') \equiv f^t_{\text{IPC}}(p')$. 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.

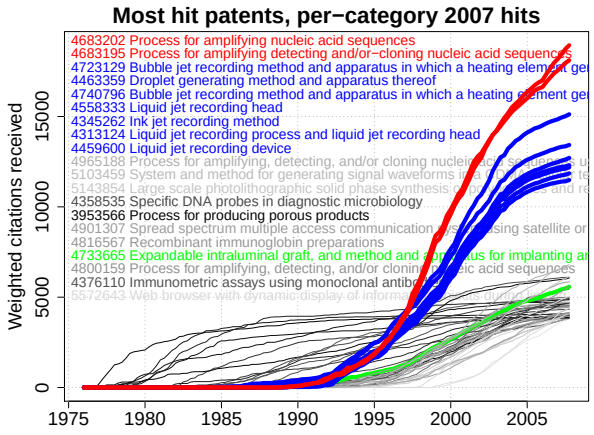

Figure 7: Normalization by mean outgoing citation rate for individual IPC categories (first two levels). This rate varies over time. Contribution to activity is weighted based on the mean number of citations made by patents in that (level 2) category at that time.

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. It is not surprising that citation counts can become inflated due to self-citations; if a company makes an innovation, it is motivated to build on that innovation and to patent further developments. However, this might create an artificially large citation count for some patents that all derive from the same source. A simple normalization to adjust for this effect uses a counting function that discounts self-citations, as follows:

$$f_{\text{self}}(p,p') = \begin{cases} \alpha & \text{if } p \text{ and } p' \text{ have the same assignee} \\ 1 & \text{otherwise} \end{cases}$$

with $\alpha < 1$. Then, the adjusted cumulative citation sum, $C^t_{\text{self}(p,p')}$, is computed from equation (1) using $f^t(p,p') \equiv f_{\text{rate}}(p,p') f_{\text{self}}(p,p')$, where we include normalization with respect to changing mean citation rates, as described above for f^t_{rate}.

Figure 8 shows a plot of $C^t_{\text{self}}(p,p')$ for $\alpha = 0.33$ (other values of α produce similar results). This normalization reshuffles the relative impact of the top patents. One effect is the dramatic drop in inkjet printing patents (blue). Those patents cover inventions developed at Canon, and numerous subsequent Canon patents cite their earlier inventions as prior art. However, relatively few other groups cite Canon's inkjet printing patents. By contrast, the PCR and stent patents virtually unaffected in both relative and absolute terms.

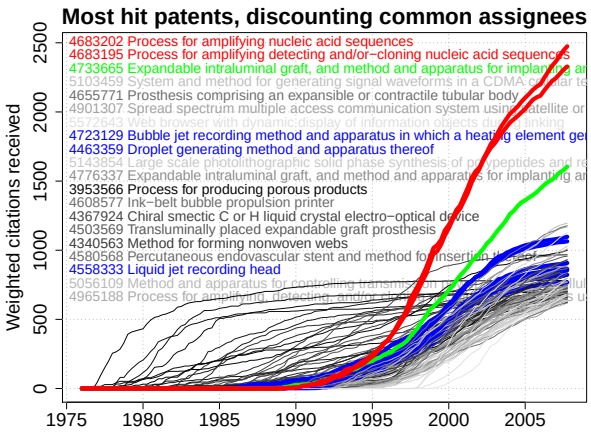

Figure 8: Discounting for self-citations. Notice that the ranking of superstar patents significantly changes, but PCR (red), inkjet printing (blue), and stents (green) remain superstars.

We may 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, inkjet 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 stories consistently remain significant.

Door-opening innovations

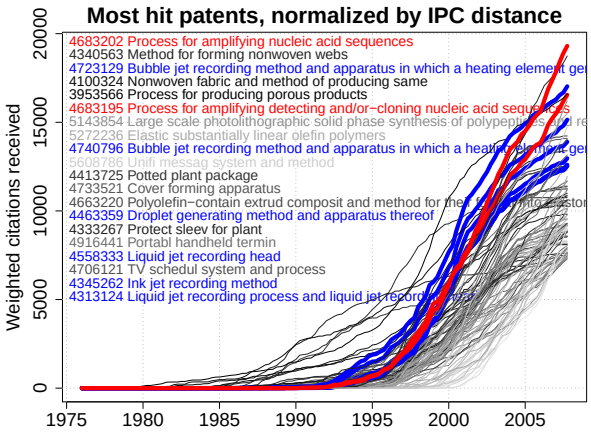

Figure 9: Weighting citation counts by the exponential of IPC distance, so that citations by patents in distant IPC categories count much more. This rewards door-opening innovations and penalizes innovations that merely spur further innovations of the same type.

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. Door-opening innovations contrast with inventions that represent "incremental progress," in which new innovations have similar IPC classifications to their ancestors. We may ask if door-opening innovations are important players in the evolution of the patent population.

Our cumulative citation statistics may be modified to address the question of how and whether door-opening patents are present in the dataset, and in particular, whether they are present in the stars that emerge. The modification to address this question takes substantially the same form as the modifications discussed above for eliminating biases and artifacts in the data: define a new counting function that emphasizes, or accentuates the property being investigated. Such use of a counting function is heuristic, in the sense that there is typically not a fundamental formulation, but rather a range of possibilities, corresponding to the testing of a range of different hypotheses.

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 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:

$$f^{t'}_{\mathrm{IPCd}}(p, p') = 2^{d_{\mathrm{IPC}}}. \quad (4)$$

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

A plot of $C^t_{\mathrm{IPCd}\ p}$ is shown in Figure 9. Note that PCR and inkjet printing remain significant innovations, indicating that they are all likely to be door-opening innovations. The argument is this: If those inventions were not door-opening but instead represented incremental progress, then weighting by IPC distance would drastically lower their relative citation levels. But instead those patents remain superstars. So, they must be door-opening.

Stents do not appear among the top hundred patents with this weighting. This suggests that while significant, stents are not door-opening to the extent that inkjet printing and PCR are. Intuitively this makes sense, stents are a more specialized type of invention. The difference between stents and the other superstars is also apparent in other normalizations where it trails the other superstars.

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's fecundity (evolutionary activity) as its fitness, for its fecundity directly measures the patent's impact on the composition of future populations.

We interpret cumulative citation count as evolutionary activity, that is, as direct evidence of the dynamics being produced by a Darwinian evolutionary process driven by differential selection. The dramatically high citation counts for the most cited patents show that high fecundity cannot be explained merely as a statistical fluctuation. This comparison with a no-selection null hypothesis embodied in the shadow patents is convincing evidence for Darwinian evolution of technology.

In addition to the population-level conclusion based on cumulative citation rates across the entire population of patents, the conclusion is reinforced by examining individual patents that are "stars," in the sense that they have exceptionally high numbers of citations. The narratives for the star patents are intuitively consistent with the interpretation of the patent population as undergoing Darwinian evolution.

The cumulative citation count on which this conclusion is based can be adjusted, to account for biases inherent in the data. We have discussed various such adjustments, and we find that the evidence for Darwinian evolution is consistently and strongly present over all versions of adjustments we have examined. The decisions for making the adjustments are delicate, and can have a substantial effect on the particular patents that emerge as stars, and on the narratives that accompany them. Some of the difficulties are inherent in the data, e.g., its finiteness, and consequently the absence of citations to the latest patents in the dataset.

Further, heuristic adjustments to our cumulative citation count statistics may be made to emphasize or uncover certain structure in the data. We have used one such adjustment, exponential boost of citations that cross IPC boundaries, to discover which patents appear to be issued for "door-opening" technologies, i.e., those that enable a broad range of further kinds innovations in areas different from the original area the patent was issued in. Applying these statistics largely corroborates the hypothesis that the patent superstars 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.

References

Bedau, M. A. (2003). Objectifying values in science: A case study. In Machamer, P. and Wolters, G., editors, *Science, Values, and Objectivity*, pages 190–219. University of Pittsburgh Press, Pittsburgh, PA.

Bedau, M. A. and Brown, C. T. (1999). Visualizing evolutionary activity of genotypes. *Artificial Life*, 5:17–35.

Bedau, M. A. and Packard, N. H. (1992). Measurement of evolutionary activity, teleology, and life. In Rasmussen, C. L. C. T. D. F. S., editor, *Artificial Life II*, pages 431–461. Addison-Wesley, Redwood City, CA.

Bedau, M. A., Snyder, E., Brown, C. T., and Packard, N. H. (1997). A comparison of evolutionary activity in artificial evolving systems and in the biosphere. In Husbands, P. and Harvey, I., editors, *Proceedings of the Fourth European Conference on Artificial Life*, pages 125–134. MIT Press, Cambridge, MA.

Bedau, M. A., Snyder, E., and Packard, N. H. (1998). A classification of long-term evolutionary dynamics. In Adami, C., Belew, R., Kitano, H., and Taylor, C., editors, *Artificial Life VI*, pages 228–237. MIT Press, Cambridge, MA.

Benzon, W. (1996). Culture as an evolutionary arena. *Journal of Social & Evolutionary Systems*, 19(4):321–365.

Jablonka, E. (2002). Between development and evolution: How to model cultural change. In Wheeler, M., Ziman, J., and Boden, M. A., editors, *The evolution of cultural entities*, pages 27–41. Oxford University Press, New York.

Jablonka, E. and Lamb, M. J. (2005). *Evolution in four dimensions: genetic, epigenetic, behavioral, and symbolic variation in the history of life*. MIT Press, Cambridge, Mass.

Raven, M. J. and Bedau, M. A. (2003). General framework for evolutionary activity. In Banzhaf, W., Christaller, T., Dittrich, P., Kim, J., and Ziegler, J., editors, *Advances in Artificial Life, 7th European Conference ECAL 2003*, volume 2801 of *Lecture Notes in Artificial Intelligence*, pages 676–685, Berlin. Springer.

Rechtsteiner, A. and Bedau, M. A. (1999). A generic model for quantitative comparison of genotypic evolutionary activity. In D. Floreano, J.-D. Nicoud, F. M., editor, *Advances in Artificial Life*, pages 109–118. Springer.

Skusa, A. and Bedau, M. A. (2002). Towards a comparison of evolutionary creativity in biological and cultural evolution. In Bedau;, R. S. M. A. and Abbass, H. A., editors, *Artificial Life VIII*, pages 233–242. MIT Press, Cambridge, MA.

Sperber, D. (1996). *Explaining culture: a naturalistic approach*. Blackwell, Cambridge, Mass.