The Science of Science

Dashun Wang and Albert-László Barabási

Part 3: The Science of Impact

This is the submitted version.
The printed version[1] is published by Cambridge University Press in 2021.

[1] Wang, Dashun, and Albert-László Barabási. 2021. *The Science of Science*. Cambridge University Press. doi:10.1017/9781108610834.

"If I have seen further, it is by standing on the shoulders of giants," wrote Isaac Newton in a letter to Robert Hooke on February 1676 [1]. That oft-quoted line succinctly captures a fundamental feature of science: its cumulative nature. Indeed, scientific discoveries rarely emerge in isolation but tend to build on previous work by other scientists. Scientists throughout the ages have typically acknowledged the provenance of the ideas they built upon. Over time this custom has turned into a strict norm, giving rise to *citations*.

A citation is a formal reference to some earlier research or discovery, explicitly identifying a particular research paper, book, report, or other form of scholarship in a way that allows the reader to find the source. Citations are critical to scientific communication, allowing readers to determine whether previous work truly supports the author's argument. Used appropriately, they help the author bolster the strength and validity of their evidence, attribute prior ideas to the appropriate sources, uphold intellectual honesty, and avoid plagiarism. For the author, citations offer a vehicle to condense and signal knowledge and eliminate the need to reproduce all of the background information necessary to understand the new work.

Lately, however, citations have taken on an additional role: the scientific community has begun using them to gauge the scientific impact of a particular paper or body of work. This is possible, the thinking goes, because scientists only cite papers that are relevant to their work. Consequently, groundbreaking papers that inspire many scientists and research projects should receive many citations. On the other hand, incremental results should receive few or no citations.

Yet as straightforward as this linkage between impact and citations may appear, it is fraught with ambiguity: How many citations constitutes "a lot"? What mechanisms drive the accumulation of citations? And what kinds of discoveries tend to garner more citations? Can we know how many citations a paper will collect in the future? How soon can we tell if a discovery has sparked serious interest? We must also take a step back and ask what citation counts tell us about the validity and impact of the ideas described in the paper. Put simply: Are citations meaningful at all?

We seek quantitative answers to these questions in the next six chapters. We have previously examined a range of patterns underlying individual scientific careers and collaborations in the first two parts of this book. Now that we have understood a lot more about the "producers" of science, be it individuals or teams, it is now time to focus on what they produce. Our first question is therefore, how many papers have we produced? For that, we need to go back to 1949 and pay a visit to Singapore.

Chapter 3.1

Big Science

In 1949, Derek de Solla Price was teaching applied mathematics at Raffles College in Singapore when the college's new library received a complete set of the *Philosophical Transactions of the Royal Society of London*. Published since 1662, *Philosophical Transactions* was the first journal exclusively devoted to science. As the library was still waiting for its future building, de Solla Price kept the volumes at his bedside. He spent the next year reading the journals cover to cover, sorting each decade of volumes into a separate pile. One day, as he looked up from his reading, he made a curious observation: the pile for the first decade was tiny, and the second only slightly taller. But as the decades went on, the height of the piles began to grow faster and faster, accelerating wildly in the most recent decades. Altogether, the 28 piles formed what looked like a classic exponential curve.

It was an observation that sparked a lifetime of passion—in the following decades, de Solla Price would systematically explore how science grew over time. By 1961, he had counted everything he could get his hands on: the number of scientific journals, the number of scientists who contributed to them, and the total number of abstracts in multiple fields [2]. No matter which dimension he charted, the same exponential curve greeted him, confirming his original insight that science was accelerating [2]. "[E]xponential growth, at an amazingly fast rate, was apparently universal and remarkably long-lived," he concluded [3].

But, de Solla Price was quick to realize, such exponential growth is rarely sustainable. We know that from studies on bacterial colonies: the number of bacteria grows exponentially in the beginning, but eventually the nutrients that sustain their growth are exhausted, and the growth saturates. Hence, de Solla Price predicted that the exponential growth of science must represent merely an initial growth spurt. Science must run out of steam and saturate eventually. He even went as far as to predict that the rapid growth in science should taper off shortly after the 1950s [4].

As we show in this chapter, while his observation was real, his prediction was wrong. The exponential growth of scientific research has continued virtually uninterrupted since de Solla Price's time, dramatically reshaping the world along the way.

3.1.1 The Exponential Growth of Science

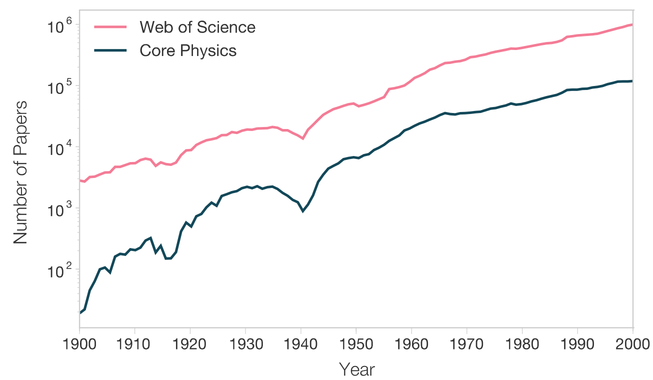


Figure 3.1.1 **The Growth of Science**. The number of papers catalogued in the Web of Science (WoS) published over the past century, illustrates the exponential growth of the scientific literature. It was only disrupted around 1915 and 1945 due to the World Wars. The figure also shows the growth of the physics literature, which follows an exponential growth similar to that followed by science as a whole. After Sinatra *et al.* [5].

Figure 3.1.1 shows the number of publications indexed yearly by Web of Science (WoS), documenting that for over a century the number of published papers has been increasing exponentially. On average, the total number has doubled roughly every 12 years. The figure also shows that science did not saturate after the 1950s as de Solla Price expected. Rather, its exponential growth has continued for the past 110 years, an expansion that was halted only temporarily by the two World Wars. Moreover, if we measure the growth rate of a single discipline, such as physics, we find similar exponential growth (Fig. 3.1.1). The fact that individual disciplines have accelerated in the same fashion suggests that the expansion of science is not simply driven by the emergence of new fields, but is instead a feature that applies across different areas of science.

What has the exponential growth of science meant for science, and for scientists? In this chapter, we will attempt to answer these complex questions.

Box 3.1.1 How many papers are there?

Considering the exponential growth of science over the centuries, you might wonder how many papers has science produced. In 2014, researchers estimated the total number of scholarly documents on the Internet using a "mark and recapture" method [6] (originally developed by ecologists to estimate the population size of wild animals when they cannot count every individual). They found that back then, there were at least 114 million English-language scholarly documents accessible on the web. Obviously, given the size of the Internet, we will never know the precise number of scholarly documents out there. But whatever the estimated number may be at any given point in time, it would quickly become outdated. Indeed, thanks to the exponential growth of science, by March 2018, four years after the 114 million estimated, Microsoft Academic Search has indexed more than 171 million papers [7].

3.1.2 The meaning of exponential growth

Any system following an exponential growth must expand at a rate proportional to its current size. Since the scientific literature doubles roughly every 12 years, this means that of all scientific work ever produced, half of it has been produced in the last 12 years. Thus, science is characterized by immediacy [3]: the bulk of knowledge remains always at the cutting edge. Furthermore, with an ever-growing number of scientists, scientists are often contemporaneous with those who have revolutionized their discipline. This is well captured by de Solla Price in his book, *Little Science, Big Science, and Beyond* [3]:

> *"During a meeting at which a number of great physicists were to give firsthand accounts of their epoch-making discoveries, the chairman opened the proceedings with the remark: 'Today we are privileged to sit side-by-side with the giants on whose shoulders we stand.' This, in a nutshell, exemplifies the peculiar immediacy of science, the recognition that so large a proportion of everything scientific that has ever occurred is happening now, within living memory. To put it another way, using any reasonable definition of a scientist, we can say that 80 to 90 percent of all the scientists that have ever lived are alive now."*

To understand what this exponential growth has meant for individuals, imagine a young scientist at the beginning of her career [3]. After years of reading the literature, with guidance from knowledgeable mentors, she has finally reached the frontier of her discipline, and is ready to strike out on her own. If science had stopped growing years ago, she would be voyaging alone into uncharted seas. While happily making discovery after discovery, she would also be quite lonely, having relatively few peers with whom to collaborate or learn from. Hence, in this respect, the continual growth of science is good news, supplying ample like-minded colleagues and interesting ideas, so that scientists can build on the work of one another and explore the world of unknowns together.

But the rapid expansion of science gives rise to another problem. As our young scientist unties the rope and tries to sail away from the harbor, she will find that there are many other boats, big and small, headed in the same direction, captained by peers with the same level of training, ambition, determination, and resources. This competitive environment will have serious consequences for our young scientist throughout her career.

First, try as she might, it will not be possible for her to monitor closely where each boat is headed. While the volume of new knowledge grows exponentially, the time a scientist can devote to absorbing new knowledge remains finite. As such, it is impossible for an individual today to read every paper in her own field.

Perhaps more importantly, each of those boats is hoping to discover something new. Indeed, in science, priority for a discovery has always been of key importance. Our young scientist would want to chart a new water than to sail in the wake of others. But so much competition may affect the individual's odds of making that grand discovery.

Many people have a heroic conception of science, believing that the field is moved forward by a small number of geniuses. But in reality, groundbreaking discoveries are often the culmination of years of hard work by many scientists. A "Eureka" moment happens when someone acknowledges what came before and manages to take the next logical leap. Indeed, Francis Bacon once argued, all innovations, social or scientific, "are the births of time rather than of wit" [8]. In other words, apples fall when they are ripe, regardless of who is around to harvest them. Thus, if a scientist misses a particular discovery, someone else will discover it instead. After all, if we already have a steam engine and a boat, how long will it take for someone to invent a steamboat?

This perhaps underlies one of the most important ways in which science differs from other creative endeavors. If Michelangelo or Picasso had never existed, the sculptures and paintings we admire in art museums would be quite different. Similarly, without Beethoven, we would not have the 5th Symphony, nor the distinctive "Pa pa pa PAM!" However, if Copernicus had never existed, we would not have arrived at an alternative description of the solar system—sooner or later we would have figured out that earth orbits around the sun, and not the other way around.

This has important implications for scientists: Apples may not care who harvests them—but for apple pickers, being the first to reach a ripe apple is critical. If science is a race to be the first to report a discovery, then the exponential growth of science poses a major question: Did the increasing competition make it more difficult to practice science?

The answer is not obvious. After all, as spectacular as the recent growth of science has been, the historical record suggests that pressures to digest and create new knowledge are not new. For example, in 1900, after being scooped by Pierre and Marie Curie on a radioactivity paper, Ernest Rutherford wrote: "I have to publish my present work as rapidly as possible in order to keep in the race." [9] (See Box 3.1.1). Or consider this confession [3]:

> "One of the diseases of this age is the multiplicity of books; they doth so overcharge the world that it is not able to digest the abundance of idle matter that is every day hatched and brought forth into the world."

This passage by Barnaby Rich dates back to 1613, half a century before the publication of the first scientific journal. Clearly the experience of feeling overwhelmed by the quantity of existing knowledge—and daunted by the prospect of adding to it—predates our generation. So, is practicing science really getting harder? To find out, let's break down the steps required to become a successful scientist, examining how each step has evolved in this age of big science.

Box 3.1.2 The birth of "Letters" Letters are a common publication format in many journals, such as *Nature* or *Physical Review Letters*. But the title may seem deceiving; these "letters" are usually complete research papers, rather than anything resembling a letter. So where does the name come from? It traces back to Ernest Rutherford, whose pragmatic choice of communicating his research shaped the way we disseminate scientific knowledge today [10].

At the beginning of the twentieth century, being the first to make a discovery was already an important goal; however, very few journals published frequently enough to allow for the rapid dissemination of new discoveries [11]. One such journal was *Nature*, which was published weekly and was read by European (mainly British) scientists, Rutherford's primary competition and audience. He cleverly started using the "Letters to the Editor" section—a column traditionally dedicated to comments on someone else's work—as a way to rapidly communicate his own discoveries and establish priority. If you read Rutherford's "letters" as far back as 1902, you will find that they look more like scientific papers than short commentaries. This Trojan Horse approach proved so successful that it was soon adopted by Rutherford's students and collaborators, notably Otto Hahn, Niels Bohr and James Chadwick. By the 1930s Otto Frisch, Lise Meitner, and the "Copenhagen school" of physicists working with Bohr had adopted it, eventually morphing the letters section into the format we know today.

3.1.3 Is it getting harder to become a scientist?

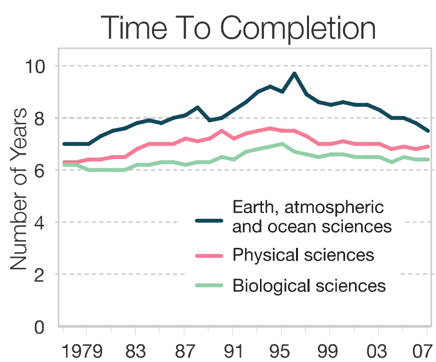


Figure 3.1.2 **How long does it take to get a PhD in the United States?** In 2007, it took a median of 7.2 years to complete a science or engineering PhD. After Cyranoski *et al.* [12].

Becoming a scientist these days requires a few essential steps, from getting the appropriate degree, to finding employment that allows the individual to pursue science. The rapid growth of training opportunities and that of the scientific workforce have affected both of these steps.

Obtaining a PhD

Figure 3.1.2 plots the time it takes to complete a science PhD in the United States, showing that the process today takes only slightly longer than it did 40 years ago. Despite a slight downward trend in the last two decades, the average time to degree in the life sciences and engineering remains six to eight years. In fact, the five years quoted by most Ph.D. programs, remains a dream for most candidates—less than 25% of students manage to complete degrees within this period [13, 14]. Instead, 55% of candidates require seven or more years to complete their degree. And these statistics only count the individuals who do obtain a degree, obscuring the fact that 40% to 50% of candidates who begin their doctoral education in the US never graduate [13, 15].

It certainly seems easier than ever to begin doctorate training, given the rapidly growing number of programs. But, why does it take so long to acquire a PhD? Multiple measures show that, once enrolled in the program, it is getting harder to produce the dissertation required for graduation.

Consider, for example, the length of the thesis itself: the average length of biology, chemistry and physics PhD theses nearly doubled between 1950 and 1990, soaring from 100 pages to nearly 200 pages in just four decades [16]. The number of references contained in a paper has been also rising over the years [17], indicating that research papers today build on a larger body of previous knowledge than ever before. Finally, the bar for publishing that dissertation—often a key milestone for aspiring scientists—has also grown higher: Comparing biology papers published in three major journals in the first six months of 1984 with the same period in 2014, researchers found that the number of panels in experimental figures (the charts and graphs published alongside papers) rose two- to four-fold [18], indicating that the amount of evidence required for a successful publication has increased significantly.

These statistics make clear how hard it has become to earn a PhD. But completing one's degree is only the first step toward a successful career in science.

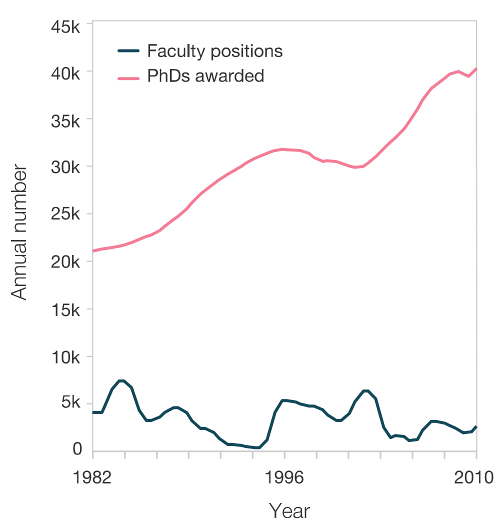


Figure 3.1.3 **The academic pipeline.** Since 1982, almost 800,000 PhDs have been awarded in science and engineering (S&E). These graduates competed for about 100,000 academic faculty positions. The number of S&E PhDs awarded annually gradually increased over this time frame, from 19,000 in 1982 to 36,000 in 2011. Yet the number of faculty positions created each year has remained stable or even fallen slightly. In 2010, the 40,000 PhD candidates competed for roughly 3,000 new academic positions. After Schillebeeckx *et al.* [15].

Securing an Academic Job

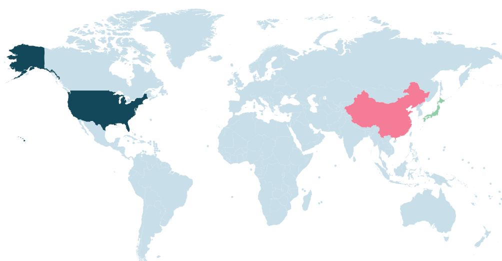
Price's observation that "80 percent to 90 percent of the scientists who ever lived are alive now" memorably illuminates the exponential growth of the scientific workforce. Yet not all parts of the academic pipeline have expanded equally quickly. For example, the number of science doctorates earned each year grew by nearly 40% between 1998 and 2008, reaching 34,000 across the 34 countries that constitute the Organization for Economic Co-operation and Development (OECD) [12]. In the same three decades, however, the number of faculty positions in those nations has remained largely unchanged and even fallen slightly (Fig. 3.1.3) [15].

While not everyone with a PhD aims to secure an academic position, the vast majority of PhDs today do seem to prefer an academic career to alternatives in industry, government, or non-profits. In a world-wide survey conducted by *Nature* in 2017, nearly 75% of the 5,700 doctoral students polled preferred a job in academia to these non-academic alternatives [19]. For them, the trend shown in Fig. 3.1.3 may seem disheartening, showing a steadily growing PhD pool competing for a fixed number of opportunities.

Box 3.1.3 Academic pipelines around the world. The **United States** produced an estimated 19,733 PhDs in the life and physical sciences in 2009 [12], a number that has continued to increase ever since. But the proportion of science PhDs who secure tenured academic positions has been dropping steadily, and industry has not been able to fully absorb the slack. In 1973, 55% of US doctorates in the biological sciences had secured tenure-track positions six years after completing their PhDs, and only 2% were still in a postdoc or other untenured academic position. By 2006, the fraction of graduates who had secured a tenure-track position within six years dropped to 15%—but 18% were now holding untenured positions.

The number of PhD holders in **mainland China** is also going through the roof, with some 50,000 students graduating with doctorates across all disciplines in 2009, more than twice the number of US degree recipients. However, thanks to China's booming economy, most Chinese PhD holders quickly find jobs in the private sector.

Japan, too, has sought to get its science capacity on par with the West. In 1996, the country's government announced an ambitious initiative to increase the number of Japanese postdocs to 10,000, and stepped up PhD recruitment to meet that goal. Yet, the government gave little thought to where all those postdocs would find jobs once their fellowship ends. Even though the government once offered companies a subsidy of around 4 million yen (US$47,000) to hire these well-educated individuals, the country still has 18,000 former postdocs who are still unemployed.



One consequence of increasing PhD production is a booming number of postdoctoral fellowships, representing short-term, non-tenure track research positions. As Fig. 3.1.4 illustrates, the number of researchers in US postdoctoral positions has more than tripled since 1979, with growth particularly prominent in the life sciences. Although postdocs drive research in many disciplines, they are often poorly compensated. Five years after receiving a PhD, the median salaries of scientists in tenure-track positions or industry far outstrip those of postdocs.

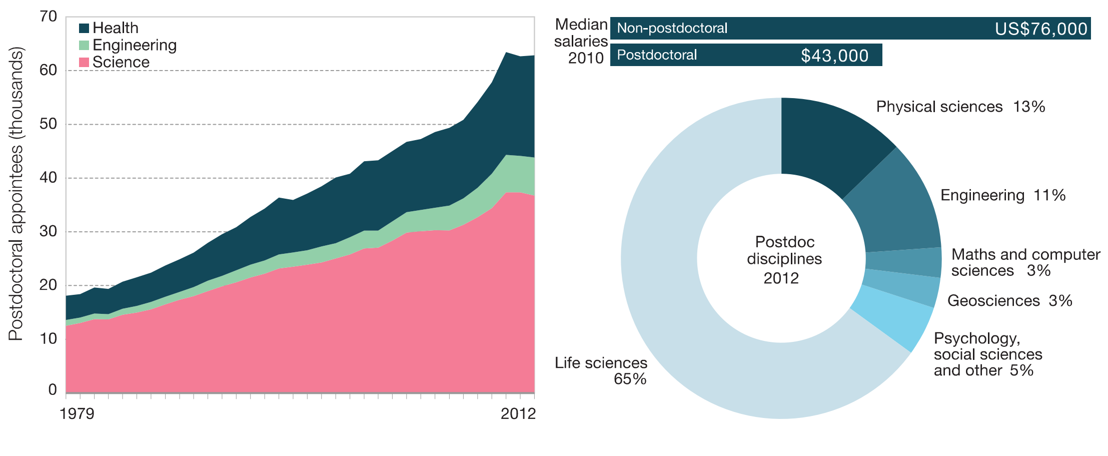


Figure 3.1.4 **The Postdoc Pileup.** The number of researchers in US postdoctoral positions has more than tripled since 1979. The vast majority of postdocs are in the life sciences. Across fields, median postdoc

salaries are outstripped by those who choose non-postdoc careers, when measured up to five years after receiving a PhD. After Powell [20].

Such bleak prospects for landing a full-time university job means that many scientists have to seek alternatives outside academia. But how well do they fare once they leave science? To answer this question, researchers collected data in eight American universities and linked anonymized Census data on employment and income to administrative records of graduate students at these universities [21]. Focusing on those PhD students who graduated between 2009 and 2011, they found that almost 40% had entered private industry, with many getting jobs at large, high-wage establishments in the technology and professional services industries. Top employers included electronics, engineering and pharmaceutical companies. These results show that PhDs who leave science tend to navigate toward large firms and earn a median salary of more than $90,000. Related research also suggests that scientists in the private sector experience very low unemployment rates overall [22]. According to the National Science Foundation's Survey of Doctorate Recipients, just 2.1% of individuals with doctorates in science, engineering or health in the United States were unemployed in 2013, while the overall national unemployment rate for people aged 25 and older was 6.3%.

Together, these data show that for the voyagers, the ocean of knowledge and opportunities *is* getting increasingly crowded. As the number of PhDs grows and the number of available university posts remains constant, competition intensifies, leading to a soaring number of postdocs, and more people leaving academia to pursue careers elsewhere. But it is important to note that this trend is not necessarily detrimental to society, nor to the scientists who pursue a wide range of careers. All signs so far indicate that PhDs who leave science do tend to find fulfilling jobs. And while they may not be creating knowledge in the scholarly sense, their work often leads to critical patents, products, and innovative solutions. As such, the loss for science is often a gain for the society.

3.1.4 Is science running out of gas?

As we stated earlier, apples fall when they are ripe—that is, scientific breakthroughs are possible once a critical mass of knowledge has been amassed. But with more and more people trying to harvest ripe apples, has the remaining fruit on the tree become harder to reach? More specifically, do new discoveries require more effort now than they did in the past?

Tracking changes in productivity can help us answer this question. Economists define productivity as the amount of work-hours required to produce an output, like building a car or printing a book. In science, it corresponds to making a new discovery, measuring the amount of effort required to write a research paper. As we have shown, both scientific discoveries and the scientific workforce have been growing exponentially for decades. But depending on which has grown faster, the average productivity per scientist may be either rising or falling. And if scientific productivity has indeed decreased, that would imply that more labor is now required to achieve any breakthrough—in other words, that science may be becoming more difficult.

Consider a related example from the computer industry. In 1965, Intel co-founder Gordon Moore noticed that the number of transistors per square inch on integrated circuit boards (now called microchips) had doubled every year since their invention. He predicted that this trend would continue into the foreseeable future. Moore's prediction proved correct: For nearly half a century, the chip density has indeed doubled approximately every 18 months, representing one of the most robust growth phenomena in history. But this growth required the efforts of an ever-increasing number of researchers [23]. Indeed, the number of individuals required to double chip density today is more than 18 times larger than the number required in the early 1970s. Therefore, even though processing power continues to grow exponentially, when it comes to producing the next generation of microchip, individual productivity has plummeted; every new release requires unprecedented amounts of manpower. Similar patterns have been observed in a wide array of industries, from agricultural crop yields to the number of new drugs approved per billion dollars spent on research [23]. This means that in these industries, the exponential growth of progress hides the fact that "apples" are indeed getting harder and harder to find.

In science, while the number of publications has grown exponentially, large-scale text analysis of physics, astronomy, and biomedicine publications revealed that the number of unique phrases in article titles has grown only linearly [24]. This suggests that the cognitive space of science—approximating the number of distinct scientific ideas in existence—may be growing much slower than scientific outputs.

So, is science running out of gas? As shown in this chapter, the exponential increase of scientific publications is indeed accompanied by an exponential growth in the number of scientists. But, if we compare the two growth rates across a large variety of disciplines, ranging from computer science to physics and chemistry to biomedicine, we find that the former is often comparable with the latter [25]. Indeed, as we discussed in Ch. 1.1, in science, individual productivity has stayed relatively stable over the past century, and even increased slightly in recent years.

As scientists look back upon the 20th century, we marvel at the discoveries and inventions that our predecessors have made, from internal combustion engines to computers to antibiotics. The tremendous progress in these areas could imply diminishing returns, as the sweetest, juiciest apples are already harvested. Yet this does not seem to be the case. The data suggests that we are poised to discover and invent even more in the next twenty years than we have in all of scientific history. In other words, even after a century of breakneck progress, science today is fresher and more energetic than ever.

How can science continue to run tirelessly after a century of exponential growth? Because unlike a car or a colony of bacteria, science runs on ideas. While a car will eventually run out of gas, and bacteria will run out of nutrients, ideas are resources that grow the more they are used. Existing ideas give birth to new ones, which soon begin to multiply. So, while our ability to further improve internal combustion engines, computers, or antibiotics may indeed have diminished, we now look forward to new advances in genetic engineering, regenerative medicine, nanotechnology, and artificial intelligence—fields that will once again revolutionize science and our society, opening up whole new chapters beyond our wildest imagination.

Chapter 3.2

Citation Disparity

On the fiftieth anniversary of Eugene Garfield's *Science Citation Index (SCI)*, *Nature* teamed up with Thomson Reuters to tally the number of papers indexed by *SCI*, counting 58 million in all [26]. If we were to print just the first page of each of these papers and stack them on top of each other, the pile would reach to the top of Mt. Kilimanjaro (Fig. 3.2.1).

While this heap is certainly impressive, even more remarkable is the disparity in scientific impact that the mountain hides. If we order the pages based on the number of citations each paper received, placing the most cited ones on top and working downward, the bottom 2,500 meters—nearly half of the mountain—would consist of papers that have never been cited. At the other extreme, the top 1.5 meters would consist of papers that have received at least 1,000 citations. And just a centimeter and a half at the very tip of this mountain would have been cited more than 10,000 times, accounting for some of the most recognizable discoveries in the history of science.

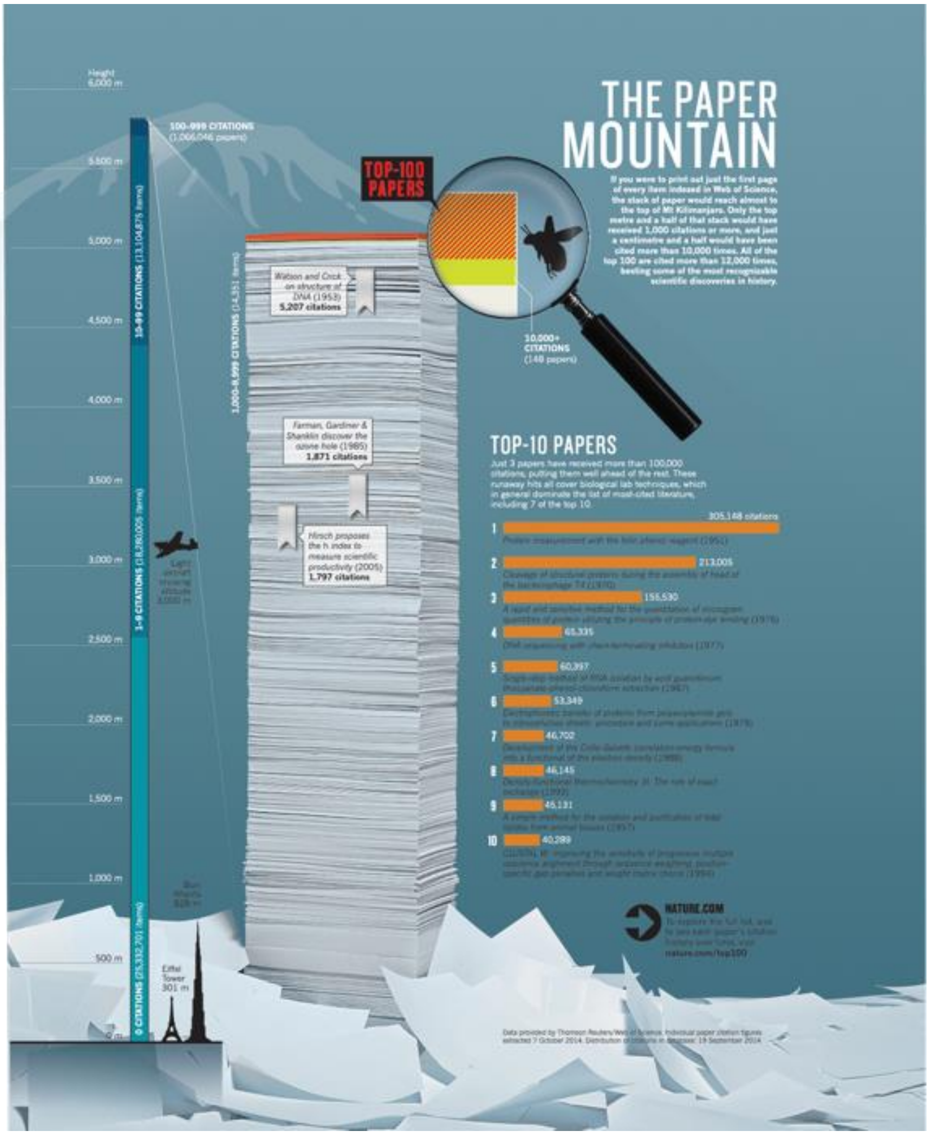


Figure 3.2.1 **The Mountain of Science**. If you print out just the first page of each research paper indexed in Web of Science by 2014, the stack would reach almost to the top of Mt Kilimanjaro. Only the top meter and a half of that stack would have received 1,000 citations or more, and just a centimeter and a half would have been cited more than 10,000 times. The top 100 papers have all been cited more than 12,000 times, including some of the most recognizable scientific discoveries in history. After Van Noorden [26].

3.2.1 Citation distribution

The difference in impact among papers can be captured by a citation distribution, $P(c)$, representing the probability that a randomly chosen paper has c citations. de Solla Price was the first to compute this distribution [27], relying on citation data manually curated by Garfield and Sher in the early 1960s [28]. For this he counted by hand the number of papers that had been cited once, twice, three times, and so forth.

He found that most papers had very few citations, but that a handful of papers had many more. Plotting the data, he realized that the citation distribution can be approximated by a power-law function,

$$P(c) \sim c^{-\gamma}, \tag{1.1}$$

with citation exponent $\gamma \approx 3$.

Most quantities in nature follow a normal or Gaussian distribution, commonly known as a bell curve. For example, if you measure the height of all your adult male acquaintances, you will find that most of them are between five and six feet tall. If you draw a bar graph to represent their heights, it will have a peak somewhere between five and six feet, and will decay quickly as you move away from the peak in either direction. After all, it is very unlikely that you know an eight-foot or a four-foot adult. Many phenomena, from the speed of molecules in a gas to human IQ, follow a similar bell curve.

Power laws like (1.1), however, belong to a different class of distributions, often called "fat-tailed distributions." The key property of these fat-tailed distributions is their high variance. For example, as Fig. 1.2 shows, the millions of papers with only a few citations coexist with a tiny minority of papers that have thousands of citations (See Box 3.2.1 The 80/20 rule.) If citations were to follow a bell curve, we would never observe such highly cited papers. Indeed, imagine a planet where the heights of the inhabitants follow a fat-tailed distribution. On such a planet, most creatures would be quite short, under one foot—but we would also occasionally encounter two-mile-tall monsters walking down the street. The strangeness of this imaginary planet highlights the stark difference between the fat-tail distribution followed by citations and the bell curves frequently seen in nature (Fig. 3.2.2).

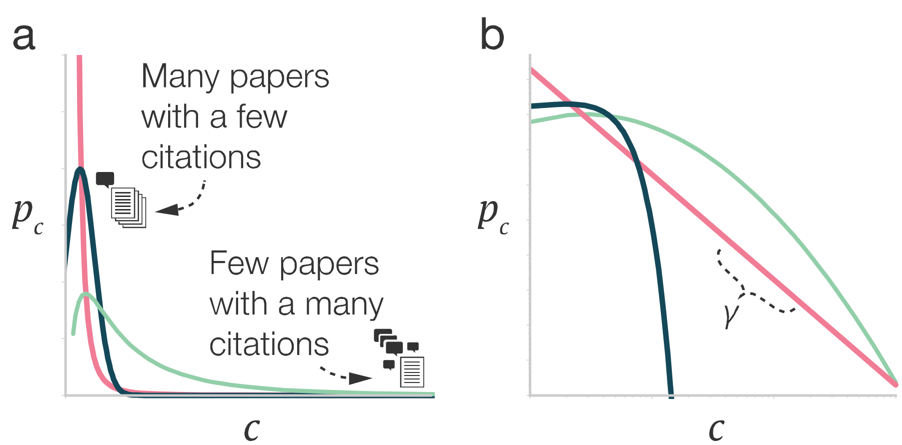


Figure 3.2.2 **Illustrating normal, power-law, and lognormal distributions**. (a) Comparing a power law and a lognormal function to a normal distribution on a linear-linear plot. (b) The same comparison shown on a log-log plot, helping us see the fundamental difference between the normal and fat-tailed distributions in the high citation regime. A power law follows a straight line on a log-log plot, its slope indicating the citation exponent γ. Sometimes, it can be difficult to tell lognormal and power laws apart, as they appear similar on a log-log plot.

Box 3.2.1 The 80/20 Rule

Power law distributions are frequently seen in the context of income and wealth. Vilfredo Pareto, a 19th century economist, noticed that in Italy a few wealthy individuals were earning most of the money, while the majority of the population earned rather small amounts. Looking more closely, he concluded that incomes follow a power law [29]. His finding is also known as the 80/20 rule: He observed that roughly 80 percent of money is earned by only 20 percent of the population.

Versions of the 80/20 rule apply to many situations with fat-tailed distributions [30-32]. For example, in business, 20% of sales reps typically generate 80% of total sales; on the World Wide Web, 80 percent of links lead to 15 percent of webpages; and in hospitals, 20% of patients account for 80% of healthcare spending.

3.2.2 Universality of citation distributions

There are remarkable differences in citations between disciplines. For example, biology papers regularly collect hundreds or even thousands of citations, while a highly influential math paper may struggle to collect a few dozen. These differences are illustrated in Fig. 3.2.3a, which depicts the distribution of citations for papers published in 1999 across several disciplines. As these plots show, the probability of an aerospace engineering paper gathering 100 citations is about 100 times smaller than the odds of a developmental biology paper reaching the same threshold. These systematic differences indicate that simply comparing the number of citations received by two papers in different disciplines would be meaningless. The less-cited aerospace engineering paper may have reported defining discoveries in that field, whereas a more cited biology paper may have reported a merely incremental advance within its discipline.

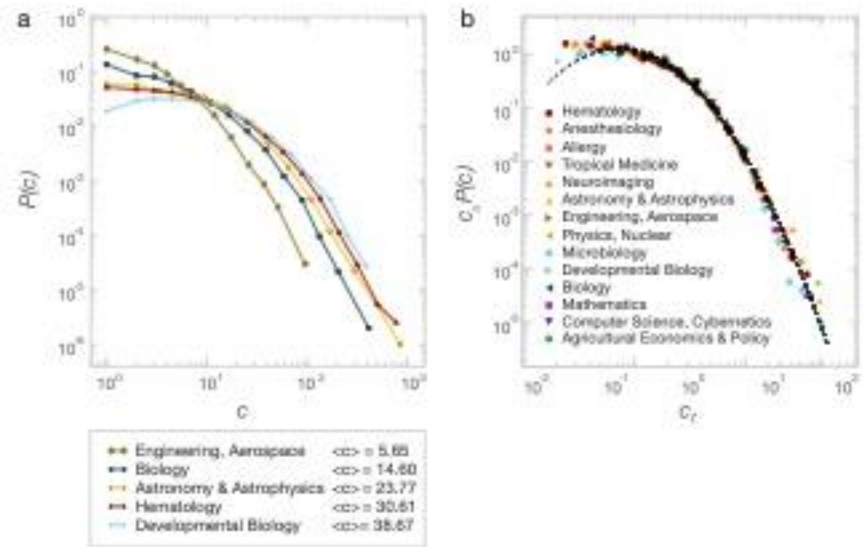


Figure 3.2.3 **Universality of citation distribution. a**. The distribution of citations for papers published in 1999, grouped by discipline. The panel shows the probability $P(c)$ of a paper receiving exactly c citations for several scientific disciplines, illustrating that in some fields, like developmental biology, highly cited papers are more common than in engineering. **b**. Rescaled citation distribution, illustrating that in fact all curves shown in A follow the same distribution, once rescaled by $\langle c \rangle$, the average number of citations in the same field and year. The dashed line shows lognormal fit (1.2) to the data. After Radicchi *et al.* [33].

To get a better idea of how influential a particular paper is, we can compare it to the average article in its discipline. Dividing a paper's citations by the average number of citations for papers in the same field in the same year gives us a better measure of relative impact. When the raw citation counts are normalized in this way, we find that the distribution for every field now neatly follows a single universal function (Fig. 3.2.3b). The finding that the curves that were so visibly different in Fig. 3.2.3a now collapse in a single curve offers two important messages:

(i) Citation patterns are remarkably universal. Whether you publish in math, the social sciences or biology, the impact of your work *relative* to your own discipline has the same odds of being mediocre, average, or exceptional as the research happening in every other building on campus (Fig. 3.2.2B).

(ii) The universal curves are well approximated by a lognormal function

$$P(c) \sim \frac{1}{\sqrt{2\pi}\sigma c} exp\left(\frac{-(\ln c-\mu)^2}{2\sigma^2}\right), \quad (1.2)$$

Box 3.2.2 A variety of functions capture citation distributions

As illustrated in Fig. 3.2.2, when plotted, fat-tailed distributions like the lognormal (1.2) or power-law function (1.1) can appear quite similar to each other. But how do we know which function offers the best fit? Various studies have argued that a variety of functions could capture the citation distribution—from power laws [31, 32, 34, 35] and shifted power laws, [36] to lognormals, [33, 35, 37-42] to other more complicated forms [43-47]. How well these distributions suit the data often depends on which corpus of papers researchers analyzed, the publication year, the selection of journals, scholars, and their departments, universities, and nations of origin, to name a few factors. The mechanisms governing the emergence of different forms for $P(c)$ is an active research topic and will be discussed in the next chapters.

The universality of the citation distributions offers a simple way to compare scientific impact across disciplines. Let's focus on two hypothetical papers: Paper A is a computational geometry paper published in 1978, which has collected 32 citations to-date; paper B is a biomedical paper published in 2002 with 100 citations. Although it may feel like an apples-to-oranges comparison, the relative citation function helps us compare the impact of the two papers. To do so, we first take all computational geometry papers published in 1978 and count their average citations. Similarly, we take all biomedical papers published in 2002, and calculate their corresponding average. Comparing the raw citation counts of paper A and B to the averages in their respective field and year provides a metric of *relative* impact for each, allowing us to make an unbiased comparison between the two.

In Table A1 (in Appendix), we calculate the average citation counts up to 2012 for all subject categories in 2004. As the table reveals, the more papers in a particular subject area, the more citations its average paper tends to receive—likely because there are more opportunities to be cited in that field. For instance, the biological sciences have some of the highest average citation counts, mostly in the 20s, whereas those in engineering and math have some of the lowest. There can be large differences even within the same subject. For example, biomaterials, a hot subfield of material science averages 23.02. Yet, the characterization and testing of materials, another subfield of material science, averages just 4.59. Interestingly, the number of papers in the two areas are not that different (2,082 vs. 1,239), suggesting that biomaterials papers are often cited in other disciplines.

Box 3.2.3 The Widening Citation Gap

The Occupy Wall Street Movement, born during the 2008 economic crisis, highlighted the fact that in the US, 1% of the population earns a disproportionate 15% of total income. This signals a profound income disparity, rooted in the fact that the income distribution follows a power-law distribution. Yet the debate about the 1% is less about the magnitude of its wealth than its trend over time: Income disparity has been skyrocketing for the last several decades (Fig. B3.2.4).

Given that citations follow fat-tailed distributions, could there be a similar rise in impact disparity in science? To answer this, we looked at the most-cited 1% of papers published in *Physical Reviews* journals in every year since 1893, measuring what fraction of total citations those papers received in the year following publication [48]. As Fig. B3.2.4 shows, impact disparity in the physical sciences is indeed present, and has been rising steadily over the past century.

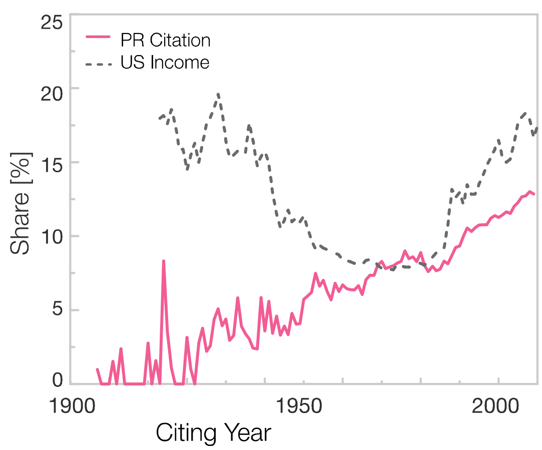


Figure B3.2.4: **The top 1% of Science.** The fraction of income earned by the top 1% of the population in US between 1930 and 2010, and the share of citations received by the top 1% most cited papers in *Physical Reviews*. We used the *Physical Reviews* (PR) corpus of 463,348 papers published between 1893 and 2009 to determine the share of citations each paper gains the following year. After Barabasi, Song, and Wang [48].

3.2.3 What do citations (not) capture?

Many of the things you can count, don't count. Many of the things you can't count, do count.

— Albert Einstein

Citations and citation-based indicators are often used to quantify the impact or the quality of a paper. These metrics play an important role both in science policy and in the evaluation of individual scientists, institutions, and research fields. For example, citation counts are often discussed at hiring and tenure decisions, as well as in prize recommendations. Furthermore, many countries consider them in decisions pertaining to grants and other resources. Given their widespread use, it is easy to forget that citations are merely a *proxy* for impact or scientific quality. Indeed, there are many ground-breaking scientific discoveries that have received relatively few citations, even as less important papers amass hundreds. Consider some of the reasons why this could be:

- Review papers that summarize the state of a field or a topic, tend to be more frequently cited than regular articles, yet are often viewed as minor contributions to science than original research [49].
- Sometimes citations seek not to build upon what came before, but rather to criticize or correct previous papers. These should be viewed as "negative citations"—yet citation counts do not distinguish between supporting and critical references [50].
- Some citations may be "perfunctory," simply acknowledging that other studies have been conducted without contributing to the main narrative of the discovery. Indeed, manual examinations of 30 articles in *Physical Review* on theoretical high energy physics from 1968 to 1972 suggests that the fraction of such citations could be substantial [51]. This means, while each citation is counted the same, they have different roles in advancing science.

All of which prompts the question: Are better-cited papers really better? To what degree do citations correctly approximate scientific impact, or what working scientists perceive as important? A number of studies have explored these questions. Overall, they find that citations correlate positively with other measures of scientific impact or recognition, including awards, reputation [52], peer ratings [53-57], as well as the authors' own assessments of their scientific contributions [49]. For example, a survey of the 400 most-cited biomedical scientists asked each of them a simple question [11]: Is your most highly cited paper also your most important one? The vast majority of this elite group answered yes, confirming that citation counts do constitute a valuable metric for a paper's perceived significance.

But would other researchers in the field agree? If we picked two papers in the same field, showed them side-by-side to researchers who work in that area, and asked them to say which is more relevant to their research, would they reliably pick the paper with more citations?

Close to 2,000 authors from two large universities performed precisely this exercise. However, the results were not straightforward [58]. The scholar's selection typically depended on whether or not he was asked to judge his own work. If both papers offered had been authored by the researcher, he systematically picked his better cited paper as the more relevant one. If, however, the author was asked to compare his paper with a paper by someone else, he overwhelmingly preferred his own paper—even if the other option was one of the most cited papers in their field, and even if the difference in impact amounts to several orders of magnitude.

These results suggest that, just as someone else's baby—no matter how adorable—can never compare to your own, scientists have a blind spot when asked to compare their work to that of others, which can skew their perception of even the most pivotal work.

Taken together, these surveys seem to tell two conflicting stories. On one end, scientists are not immune to biases, especially when it comes to their own work. But at the same time, citations could be a meaningful measure, since they largely align with the perceptions of experts. This suggests that, despite their shortcomings, citations likely play a critical role in gauging scientific impact.

But there is another, more fundamental reason why citations matter: There is not one scientist in the world who can single-handedly demand that a paper amass citations [59]. Each of us decides on our own whose shoulders to stand upon, citing the papers that inspired our work, and those which were key to developing our ideas. When all of those individual decisions are combined, what emerges is the collective wisdom of the scientific community on a paper's importance. It is not enough for one scientist to decide that a paper is great. Others in the community must also agree, with each additional citation acting as an independent endorsement. In other words, scientific impact is not about what *you* think; it's about what *everyone else* thinks.

Box 3.2.4 Patent Citations

Citations can also help quantify the importance of inventions. When a patent is issued by the patent office, it includes citations to prior patents, as well as to the academic literature. Patent citations can be added by inventors, their attorneys or by patent examiners. Just like in science, patents with the most citations tend to describe important inventions [60, 61]. But patent citations go a step further, conveying not just an invention's importance, but also its economic impact. For example, studies have found that more highly-cited medical diagnostic imaging patents produce machines that attract more demand [62] and that companies with highly-cited patents have higher stock market values [63]. More importantly, the link between the commercial value of a patent and its citations is positive and highly nonlinear. One survey polled patent owners 20 years after their invention, asking them how much they should have asked for their patent knowing what they know today. The responses revealed a dramatic, exponential relationship between the economic value of a patent and its citations [64]: A patent with 14 citations, for example, had 100 times the value of a patent with 8 citations.

Chapter 3.3

High Impact Papers

Why are most papers rarely cited, and how do a few lucky ones turn into runaway successes? These questions present a deep puzzle. In every new paper, the authors carefully select which work to cite, based on the topic of the manuscript and their own familiarity with the literature. Yet somehow these many personal decisions result in a highly stable citation distribution, capturing impact disparities that transcend disciplines. How do citation superstars emerge? What determines which papers are highly cited, and which are forgotten? And why are these universal citation distributions universal, independent of the discipline?

In this chapter we show that, perhaps somewhat counterintuitively, citation superstars and the universality of the citation distribution emerge precisely *because* citations are driven by individual preferences and decisions. Although our individual choices differ widely, the behavior of the scientific community as a whole follows highly reproducible patterns. As such, despite the seemingly countless factors that govern each paper's impact, the emergence of exceptionally impactful papers can be explained by a handful of simple mechanisms.

3.3.1 The 'Rich-Get-Richer' Phenomenon

No scientist can read the million or so scientific papers published each year. So we typically discover papers of interest while reading other papers, and looking at the work they cite. This leads to a curious bias: the more widely cited a paper already happens to be, the more likely it is that we will encounter it through our reading. And since we typically only cite work that we read, our reference lists frequently becomes populated with highly cited papers. This is an example of a "rich-get-richer" phenomenon—similar to the Matthew effect encountered in Chapter 1.3—the higher a paper's citation count, the more likely it is to be cited again in the future.

As simple as it seems, the rich-get-richer mechanism alone can explain much of the citation disparity among scientific publications, and for the universal, field-independent nature of citation distributions. This was formalized in a model first proposed by de Solla Price in 1976 (Box 3.3.1), sometimes called the *Price model* [65, 66], which incorporates two key aspects of citations:

1. **The growth of the scientific literature.** New papers are continuously published, each of which cite a certain number of previous papers.
2. **Preferential attachment.** The probability that an author chooses a particular paper to cite is not uniform, but proportional to how many citations the paper already has.

As discussed in more detail in Appendix A2.1, the model with these two ingredients (growth and preferential attachment) predicts that citations follow a power-law distribution, hence can explain the empirically observed fat-tailed nature of citations.

Box 1.3.1. The rich-get-richer effect has been independently discovered in multiple disciplines over the last century, helping to explain disparities in city and firm sizes, abundance of species, income, word frequencies, and more [65, 67-73]. The best-known version was introduced in the context of complex networks [30, 74], where the term preferential attachment was proposed by the Barabasi-Albert model [73] to explain the existence of hubs in real networks. In sociology, it is often called "the Matthew effect," as discussed in Chapter 1.3 and also called cumulative advantage by de Solla Price [65].

This analysis leads us to a critical takeaway. Clearly there are myriad factors that contribute to the impact of a paper, some of which will be considered in later chapters. But the accuracy with which Price's model captures the empirically observed citation distribution demonstrates that these additional factors don't matter if our only goal is to explain the origin of the fat-tailed distribution of citations. Growth and preferential attachment, together leading to a rich-get-richer effect, can fully account for the observed citation disparity, pinpointing the reason why citation superstars emerge (see Appendix A2.2 for the origins of preferential attachment). It also explains why citation distributions are so universal across widely different disciplines: While many factors may differ from one discipline to the other, as long as the same preferential attachment is present, it will generate a similar citation distribution, independent of any disciplinary peculiarities.

3.3.2 First-Mover Advantage

Price's model has another important takeaway: The older a paper is, the more citations it should acquire. This phenomenon is called the *first-mover advantage* in the business literature. That is, the first papers to appear in a field have a better chance of accumulating citations than papers published later. Then, thanks to preferential attachment, these early papers will retain their advantage in perpetuity.

This prediction was put to the test by analyzing citations patterns of papers published in several subfields, like network science and adult neural stem cells [75], finding a clear first-mover effect, whose magnitude and duration are quite similar to that predicted by Price's model. To appreciate the size of this effect, let's look at the field of network science, born at the end of the 1990s. The first 10% of papers published in this field received an average 101 citations, while the second 10% received just 26 on average. Since the second batch was published immediately after the first, the difference is likely not due to the fact that the second batch had less time to collect citations.

Scientists tend to treat the first papers as founders of a field, which may explain their high citation count. The first researcher to bring a problem to the attention of the scientific community deserves credit for it, regardless of whether all of the original results retain their relevance.

Occasionally, however, prominent latecomers do take over. Consider, for example, the Bardeen-Cooper-Schrieffer (BCS) paper [76], which introduced the first widely accepted theory of superconductivity. It was a relatively late contribution to the field of superconductivity. Yet by explaining a wide range of puzzles plaguing the field with a single elegant theory, it quickly became the defining paper, and the citations followed. Hence, the BCS paper is a living testament to the fact that the first-mover principle is not absolute, prompting us to ask: If only the rich are allowed to get richer, how can a latecomer succeed?

To answer this question, we must recognize that preferential attachment is not the only mechanism driving citation counts. As we will see next, there are other mechanisms that determine which papers loom large in the scientific canon, and which fade into obscurity.

3.3.3 The Fit Get Richer

Despite its simplicity, Price's model omits a crucial factor that does influence the accumulation of citations: not all papers make equally important contributions to the literature. Indeed, there are huge differences in the perceived novelty, importance, and quality of papers. For example, some papers report highly surprising discoveries that alter the prevailing paradigm, and others, like the BCS paper, bring clarity to long-standing puzzles in a large and active field—but these papers coexist with other publications that merely rehash old ideas or proffer half-baked theories. Papers also differ in their publication venues, the size of the audience they speak to, and the nature of their contribution (i.e., review papers and method papers tend to be cited more than regular research papers). In other words, papers differ in their inherent ability to acquire citations, each being characterized by some set of intrinsic properties that will determine its impact relative to the pack. We will call this set of properties *fitness*, a concept borrowed from ecology and network science [77].

Price's model assumes that the growth rate of a paper's citations is determined solely by its current number of citations. To build upon this basic model, let's assume that citation rate is driven by both preferential attachment and a paper's fitness. This is called the *fitness model* [77, 78], or the Bianconi-Barabási model, which incorporats the following two ingredients (See Appendix A2.3 for more detail).

- **Growth**: In each time step, a new paper *i* with a certain number of references and fitness η_i is published, where η_i is a random number chosen from a distribution $p(\eta)$. Once assigned, the paper's fitness does not change over time.
- **Preferential Attachment**: The probability that the new paper cites an existing paper is proportional to the product of paper *i*'s previous citations and its fitness η_i.

Here a paper's citation is not just dependent on its existing citations, captured by the preferential attachment mechanism we have discussed earlier. It also depends on its fitness, indicating that between two papers with the same number of citations, the one with higher fitness will attract citations at a higher rate. Hence, the presence of fitness assures that even a relatively new paper, with a few citations initially, can acquire citations rapidly if it has greater fitness than other papers.

Box 1.3.2: The origin of the lognormal citation distribution. As we discussed in Ch. 3.2, several recent studies have indicated that citation distributions sometimes are better-fitted by lognormal distributions. While a lognormal form is inconsistent with the Price's model (which predicts a power law), it can be explained by the fitness model [77]. Indeed, if every paper has the same fitness, thanks to preferential attachment, citations will follow a power law distribution. If, however, papers differ in their fitness, then the underlying distribution of fitness parameter $p(\eta)$, from which we draw a number to assign a paper its fitness, determines the shape of citation distribution. For instance, if $p(\eta)$ follows a normal distribution, which is a natural assumption for bounded quantities like fitness, the fitness model predicts that the citation distribution should follow a lognormal function [40]. As we show in the next chapter (Ch 3.6), where we estimate the fitness of a paper, the fitness distribution is indeed bounded, like a normal distribution. This suggests that if we wish to explain the empirically observed citation distributions, we cannot ignore the fact that papers have different fitness.

Differences in fitness explain how a latecomer can overtake established citation leaders. Indeed, a strong paper may appear late in the game but nevertheless grab an extraordinary number of citations within a short time frame. Unlike Price's model (which predicted that citations of all papers grow at the same rate), the fitness model predicts that the growth rate of citations is proportional to the paper's fitness, η. Thus, a paper with higher fitness acquires citations at a higher rate, and given sufficient time, it will leave behind the older papers with lower fitness.

But what does it take to write a high fitness paper?

Chapter 3.4

Scientific Impact

If we assume that citations approximate a paper's scientific impact, then the fat-tailed shape of the impact distribution implies that most papers have, unfortunately, no impact at all; indeed, only a very small fraction of the literature affects the development of a field. As the preceding chapter showed, high fitness is crucial for certain ideas to make a large impact. But what predicts high fitness? And how can we amplify the scientific impact of our work? In this chapter, we will focus on the role of two different factors—one is internal to a paper and the other is external—novelty and publicity.

3.4.1 The link between novelty and scientific impact

While many qualities can affect the 'fitness' of a paper, one, in particular, has attracted much attention: novelty. What exactly is novelty, and how do we measure it in science? And does novelty help or hurt a paper's impact?

Measuring Novelty

As discussed in Chapter 3.1, new ideas typically synthesize existing knowledge. For example, inventions bring together pre-existing ideas or processes to create something original (Fig. 3.4.1) [79]. The steamboat is a combination of a sailing ship and a steam engine, and the Benz Patent-Motorwagen, the first automobile in the world, combined a bicycle, a carriage, and an internal combustion engine. Even the smartphone in your pocket is simply a combination of many pre-existing parts and features: memory, digital music, a cell phone, internet access, and a lightweight battery.

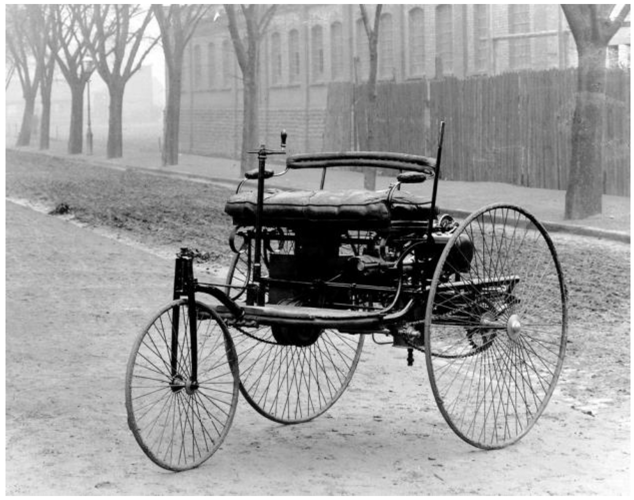


Figure 3.4.1 **New ideas are often an original combination of existing ones.** The Benz Patent-Motorwagen ("patent motorcar"), built in 1885, is regarded as the world's first production automobile. The vehicle was awarded the German patent number 37435, for which Karl Benz applied on 29 January 1886. The Motorwagen represents a combination of three pre-existing ideas: bicycle, carriage, and internal combustion engine.

The theory that existing technologies are recombined to generate new inventions is confirmed by the analysis of US patents [80]. Each patent is classified by the US patent office (USPTO) using a unified scheme of technology codes (a class and a subclass). For example, one of the original patents for iPod, assigned to Apple Computer, Inc with Steve Jobs listed as one of the inventors [81], has a class-subclass pair 345/156, denoting class 345 (Computer Graphics Processing and Selective Visual Display Systems) and subclass 156 (Display Peripheral Interface Input Device). Examining all US patents dating from 1790 to 2010, researchers found that, during the 19th century, nearly half of all patents issued in the US were for single-code inventions—those that utilize a single technology, rather than combining multiple technology areas. Today, by contrast, 90% of inventions combine at least two codes, showing that invention is increasingly a combinatorial process.

This combinatorial view of innovation offers a way to quantify novelty in science. Indeed, scientific papers draw their references from multiple journals, signaling the domains from which they sourced their ideas [82-84]. Some of these combinations are anticipated, whereas others are novel, deviating from conventional wisdom.

If a paper cites a pair of journals that are rarely brought together, it may suggest that the paper introduces a novel combination of prior work. Take for instance a 2001 paper [85] in the *Journal of Biological Chemistry*, which pinpointed the protein with which a known antipsychotic drug interacts and used this insight to identify other biological effects. Its reference list is the first ever to cite both the journal *Gene Expression* and the *Journal of Clinical Psychiatry* [86], representing a novel combination of prior knowledge. On the other hand, other journals cited in the same paper, like the *Journal of Biological Chemistry* and *Biochemical Journal*, are frequently co-cited in the literature, an example of the conventional pairings that reflect more mainstream thinking in the field.

The Novelty Paradox

Evidence from a broad array of investigations consistently shows that rare combinations in scientific publications or patented inventions are associated with a higher likelihood that the publication or invention will achieve high impact. In other words, with novelty comes an increased chance of hitting a home run. This finding also validates the key premise of interdisciplinary research [87-89]—that many fruitful discoveries come from the cross-pollination of different fields and ways of thinking, combining previously disconnected ideas and resources [89-91].

Yet, while novel ideas often to lead to high-impact work, they also lead to higher degrees of uncertainty [82, 92, 93]. In fact, very novel ideas and combinations can just as well lead to failure as to a breakthrough. For example, an analysis of more than 17,000 patents found that the greater the divergence between the collaborators' fields of expertise, the higher the variance of the outcomes; highly divergent patents were more likely than average both to be a breakthrough, and to be a failure (Fig. 3.4.2) [93].

Similarly, papers that cite more novel combinations of journals are more likely to be in the top 1% cited papers in their field. Yet at the same time, they are also riskier, tending to take a longer time before they begin accumulating more citations [82]. The higher risk inherent in innovation may play a major role in determining what kind of innovation takes place (or doesn't) in academia. For instance, in the field of biochemistry, studying chemical relationships between unexplored compound pairs is much more novel than focusing on well-studied chemicals, and such strategies are indeed more likely to achieve high impact. But the risk of failure of exploring such previously unexplored combinations is so high that, as an analysis estimated, the additional reward may not justify the risk [92].

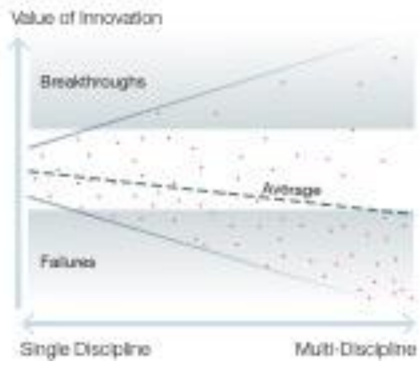


Figure 3.4.2 **Multi-disciplinary collaborations in patenting.** As collaborations among inventors become more multidisciplinary, the overall quality of their patents decreases. But multidisciplinary collaboration increases the variance of the outcome, meaning that both failures and breakthroughs are more likely. After Fleming [93].

The high variance in the impact of novel ideas may be rooted in the human bias against novelty. Studies of grant applications show that scientists tend to be biased against novel concepts before the work is realized. At a leading US medical school, researchers randomly assigned 142 world-class scientists to review 15 grant proposals. In parallel, the researchers measured the occurrences of rare combinations of keywords in each proposal [94]. For example, proposals combining the terms “Type 1 Diabetes” and “Insulin” were typical, whereas proposals with “Type 1 Diabetes” and “Zebrafish” presented a novel combination rarely seen in the literature. But would the more novel proposals be graded more or less favorably? The researchers found that proposals that scored high on novelty received systematically lower ratings than their less novel counterparts. Even nominally “interdisciplinary” grants are not immune to similar biases [95]. Analyzing all 18,476 proposals submitted to an Australian funding program, including both successful and unsuccessful applications, researchers measured how many different fields were represented in each proposal, which is weighted by how distant those fields were. The results indicated that the more interdisciplinary the proposed work, the lower the likelihood of being funded.

And, so we are left with a paradox. It is clear that novelty is essential in science—novel ideas are those that score big. Yet the novelty bias observed in grant applications suggests that an innovative scientist may have trouble getting the funding necessary to test these ideas at the first place. And, even if she does, novel ideas are more likely to fail than mediocre ones.

Is there anything we can do to ameliorate this paradox? Recent studies have offered one crucial insight: balance novelty with conventionality. Consider that Darwin devoted the first part of *On the Origin of*

Species to highly conventional, well-accepted knowledge about the selective breeding of dogs, cattle, and birds. In doing so, he exhibited an essential feature of many high-fitness ideas that do achieve great impact: They tend to be grounded in conventional combinations of prior work, while also merging hitherto uncombined, atypical knowledge. Analyzing 17.9 million papers spanning all scientific fields, researchers found that papers that introduced novel combinations, yet remained embedded in conventional work, were at least twice as likely to be hits than the average paper [84]. These results show that novelty can become especially influential when paired with familiar, conventional thought [84, 96].

3.4.2 Publicity (good or bad) amplifies citations

Does media coverage amplify scientific impact? Are we more likely to cite papers that have been publicized in the popular press? To answer these questions, let's turn to a major news outlet: *The New York Times*.

Since papers about human health are often of general interest, researchers in one study looked at whether articles published by *The New England Journal of Medicine* (*NEJM*) were covered by the *Times*. They compared the number of citations *NEJM* articles received when they had been written up in the *Times*, versus when they were not written up in the *Times*. [97]. They found that overall, articles covered by the *Times* received 72.8% more citations in the first year than the non-covered group.

But can we attribute this dramatic impact difference to the publicity that the *Times* offers? Or could it be that the *Times* simply covered outstanding papers, which would have gathered just as many citations without their coverage? A natural experiment allowed the researchers to find the definitive answer: The *Times* staff underwent a 12-week strike from August 10 to November 5, 1978. During this period, it continued to print a reduced "edition of record" but did not sell copies to the public. In other words, it continued to earmark articles it deemed worthy of coverage during the strike, but this information never reached its readership. During this period, researchers found, the citation advantage disappeared entirely—the articles that the *Times* selected for coverage did no better than those it didn't in terms of citations. Therefore, the citation advantage of attention-grabbing papers cannot be explained solely by their higher quality, novelty, or even mass appeal—it is also the result of media coverage itself.

It is not hard to see why publicity helps boost citations. Media coverage increases the reach of the audience, potentially allowing a wider group of researchers to learn about the findings. It may also act as a stamp of approval, bolstering the credibility of the paper in the eye of the scientific community. But perhaps the most basic reason is that media publicity is, more often than not, *good* publicity. Indeed, a TV station or newspaper does not pretend to be a check or balance on science. When media chooses to spend its limited air time or ink on a scientific study, it usually presents findings that are deemed genuine, interesting, and important—after all, if they weren't all of these things, then why bother wasting the audience's time?

Box 3.4.1 Media bias and science

The important role the media plays in disseminating science raises a critical question: does the press offer balanced coverage, or does it distort scientific findings as it tries to make them accessible and interesting to the public? Research on media coverage of medical studies found that journalists prefer to cover only initial findings, many of which are refuted by subsequent studies and meta-analyses. But journalists often fail to inform the public when the studies they covered are disproven [98, 99]. When a team of researchers examined 5,029 articles about risk factors for diseases and how those articles were covered in the media [99], they found that studies reporting positive associations about disease risk and protection (e.g., studies suggesting that a certain food may cause cancer, or that a certain behavior may help stave off heart disease) tend to be widely covered. In contrast, studies finding no significant association received basically zero media interest. Moreover, when follow-up analyses fail to replicate widely reported positive associations, these follow-up results are rarely mentioned in the media. This is troubling since the researchers found that, of the 156 studies reported by newspapers that initially described positive associations, only 48.7% were supported by subsequent studies.

This hints at the main tension between the media and the sciences: While the media tends to cover the latest advances, in science, it's the complete body of scientific work that matters. A single study can almost never definitively prove or disprove an effect, nor confirm or discredit an explanation [100, 101]. It takes many papers to obtain a nuanced and accurate conclusion. Indeed, the more novel a paper's initial findings, the more vulnerable they are to refutation [102].

The media's tendency to report simple, preliminary results can have severe consequences, as the media coverage of vaccinations reveals. Presently, many parents in the United States refuse to vaccinate their children against measles, mumps, and rubella (MMR), fearing that the vaccine could cause autism. Why do they believe that? This fear is rooted in a 1998 paper [103] published by Andrew Wakefield in *The Lancet*, which received world-wide media coverage. However, the original study was based on a sample of only 12 children, and follow-up studies unanimously failed to confirm the link. Furthermore, researchers later found that Wakefield had distorted his data, hence the paper was retracted with a clear statement identifying the deliberate falsification of the original research. He subsequently lost his license and was barred from practicing medicine. Yet, although Wakefield's findings have been widely refuted and discredited in the scientific community, the media's coverage of Wakefield's claim led to a decline in vaccination rates in the United States, the United Kingdom and Ireland. Consequently, measles and mumps have become more prevalent in these countries, resulting in serious preventable illness and deaths.

Media offers only good publicity for science, which may have important consequences for the public perception of science (see Box 3.4.1). But, the checks and balances used to ensure that scientific work is accurate and honest are maintained by scientists. Scientific critiques and rebuttals can come in many forms: some only offer an alternative interpretation of the original results, and others may refute only a part of a study. In most cases, however, rebuttals aim to highlight substantial flaws in published papers, acting as the first line of defense after scientific research has passed through the peer review system. Here, it seems, we finally have a form of *bad* publicity. But do these critiques and rebuttals diminish a paper's impact? And, if so, how much?

Comments, which tend to question the validity of a paper, are often seen as "negative citations", ostensibly making the original paper less trustworthy in the eye of the scientific community. Hence, one would expect commented papers to have less impact. Yet studies have revealed the opposite: Commented papers are not only cited more than non-commented papers—they are also significantly more likely to be among the most cited papers in a journal [50].

Similar results are uncovered by studies of negative citations—references that pinpoint limitations, inconsistencies, or flaws of prior studies [104]. Researchers used machine learning and natural language processing techniques to classify negative citations from a training set of 15,000 citations extracted from *Journal of Immunology*, categorizing the citations as "negative" or "objective" with the help of five immunology experts. They then used the tool to analyze 15,731 articles from the same journal. They found that papers pay only a slight long-term penalty in the total number of citations they receive after a negative one, and the criticized papers continue to garner citations over time, which paints the picture that it's better to receive negative attention than none at all.

Together, these results show that comments and negative citations seem to play a role that's the opposite of what is intended; they are early indicators of a paper's impact. Why does such bad publicity amplify citation impact?

The main culprit is a selection effect: Scientists are often reluctant to devote time to writing comments on weak or irrelevant results [105]. Hence, only papers perceived as potentially significant draw enough attention to be commented on in the first place. Moreover, while comments or negative citations are critical in tone, they often offer a more nuanced understanding of the results, advancing the argument presented in the paper rather than simply invalidating its key findings. In addition, comments also bring attention to the paper, further boosting its visibility. Even in science, it appears, there's no such thing as bad publicity.

Chapter 3.5

The Time Dimension of Science

The Library of Alexandria had a simple but ambitious goal: to collect all knowledge in existence at the time. Built in the Eastern Harbor of Alexandria in Egypt, where Queen Cleopatra first laid eyes on Julius Caesar, the library relied on a unique method to enrich its collection: Every ship that sailed into the busy Alexandria harbor was searched for books. If one was found, it was confiscated on the spot, and taken to the library where scribes would copy it word for word. The rightful owner was eventually returned the copy along with adequate compensation, while the library kept the original. Although historians continue to debate the precise number of books it managed to amass, the library at its peak was estimated to possess nearly half a million scrolls—that is, until Julius Caesar burned it all down in 48 BC.

Imagine you are at the library of Alexandria, watching the fire spread slowly towards the building. You are surrounded by the world's single greatest archive of knowledge and you know that it will soon be turned into ashes. You have time to step in and save a few scrolls. Which ones should you rescue? Should you collect the oldest scrolls, containing the ideas that have stood the test of time? Or should you run for the most recent documents, as they synthesize the best of past knowledge? Or perhaps you should pick some documents at random. The fire is approaching, and time is running out—what would you do?

Tomorrow's biggest discoveries necessarily build on past knowledge. But when choosing which discoveries to build upon, how far back should we go? And, relatedly, how long will others continue to cite our own work before the fires of time render it irrelevant? These are the questions we will explore in this chapter. We will pinpoint the unique combinations of old and relatively new knowledge that are most likely to produce new breakthroughs. In doing so, we will see that the way we build on past knowledge follows clear patterns—and we will explore how these patterns shape future scientific discourse. Let's begin with a simple question: How far back into the past do scientists typically anchor their inquiries?

3.5.1 Myopic or not?

The famous Newton quote, "If I have seen further than others, it is by standing upon the shoulders of giants," suggests that older work, tested by time, is the foundation for new discoveries. However, Francis Bacon disagreed [8], arguing that discoveries occur only when their time has come. If Bacon is right, then it is the most *recent* work that drives breakthroughs. So, if you don't want to miss out on a discovery, you'd better stay at the bleeding edge of knowledge.

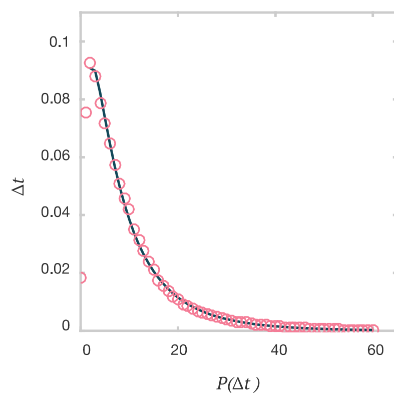


Figure 3.5.1 **The Age of Citations.** To understand where new knowledge comes from, we can look how old are the references of a paper. The age distribution of citations for papers published in 2010 shows a peak around two years and decays quickly afterwards. Much research has been devoted to reasoning about how best to characterize the distribution of citation ages. A classic meta-analysis of over 20 studies indicates that the age of references is best approximated by an exponential distribution [65]. A more recent analysis suggests that a lognormal distribution with an exponential cutoff offers a better fit [17], shown in the figure as a solid line. After Yin and Wang [17].

Who is right, Newton or Bacon? One way to test the two theories is to compile the age distribution of the references in research papers, measuring the time gap between a work's publication year and the publication years of its references. This approach has a long history: librarians used to look at the ages of references to determine which older journal volumes could be discarded to free up shelf space [106].

Figure 3.5.1 shows the probability that a paper cites a paper published t years earlier. The distribution vividly documents the myopic nature of science: most references point to works published just two to three years earlier. One may argue that this is a consequence of the fact that there were many more papers published in recent years than in years further back [65, 107]. But even when we account for the exponential growth of literature, the distribution nonetheless decays rapidly with time. While scholars do regularly reach back to “vintage” knowledge more than 20 years old, the likelihood of citing research older than that decays rather quickly.

The distribution in Fig. 3.5.1 captures the collective foraging patterns of scientists, showing that we tend to rely heavily on recent information, yet balance it occasionally with vintage, or canonical knowledge. Individual scientists may differ, however, in how they search for past knowledge, raising the question: could the varied patterns of information foraging help determine the impact of their work? Is there a search strategy that is particularly well-suited to prompting tomorrow’s breakthroughs?

3.5.2 The Hotspot of Discovery

Imagine four papers that differ in the way they cite previous literature (Fig. 3.5.2). Which one will have the highest impact? We can make a case for each of them. The deep-reach paper (Fig. 3.5.2A) primarily cites old papers, building on well-established, well-tested classics. By contrast, the paper in Fig. 3.5.2B draws only upon hot, recent topics that are absorbing the community’s attention at the moment. The paper in Fig. 3.5.2C combines the best of both worlds, featuring a judicious mix of new and vintage knowledge, with an emphasis on more recent discoveries, while the one in Fig. 3.5.2D draws upon knowledge somewhat uniformly across time.

Which strategy leads to the greatest impact? To answer this question, a team of researchers analyzed more than 28 million papers in the Web of Science, capturing two parameters for each paper [108]: the mean age of its references ($< T >$) and the coefficient of variation (σ_T) for the age distribution of its references, which captures how far those references’ publication dates are spread out. Then the researchers measured the number of citations each paper had accumulated eight years after its publication. They considered a work of high-impact (i.e. a hit paper,) if it lay within the top 5th percentile of citation counts in its subfield.

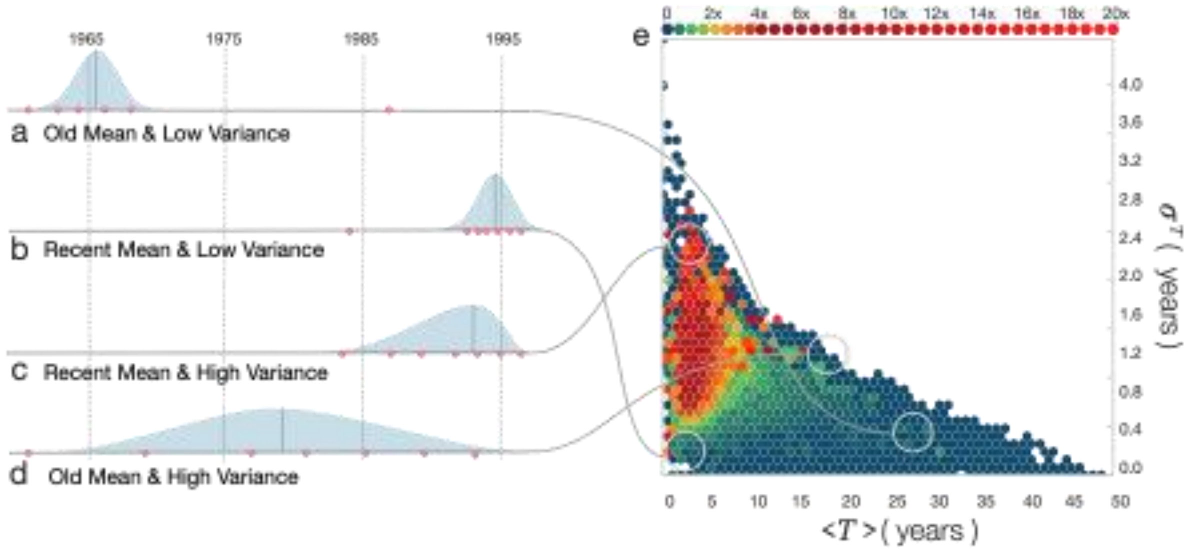


Figure 3.5.2 **Knowledge Hotspots Predict High-Impact.** (a—d) Potential information search patterns characterizing individual papers. (E) The hotspot, shown in red, captures "hit" papers that cite prior knowledge with a low mean age, $< T >$, and a high age coefficient of variation, σ_T (data shown for the year 1995, capturing N = 546,912 publications). The background rate is the likelihood that a paper chosen at random is in the top 5% of citations for papers in its field. Papers in the hotspot are on average more than two times more likely than the background rate to be hits. Notably, 75% of papers are outside the hotspot, and their likelihood of being a hit is no greater than expected by chance. After Mukherjee *et al.* [108].

The relationship between $< T >$, σ_T, and impact is captured by the heat plot shown in Fig. 3.5.2E. Each point represents the $< T >$ and σ_T values of papers published in 1995 and the color corresponds to the probability of finding hit papers at that particular combination of T_μ and σ_T , defining the "hotspots" of discovery. Three main conclusions emerge from Fig. 3.5.2E:

- Papers with low $< T >$ and high σ_T are in the middle of the hotspot, offering the highest impact. In other words, papers whose references are centered just behind the bleeding edge of knowledge but which also combine an unusually broad mix of old and new (as in Fig. 3.5.2C) tend to garner the most citations. Papers of this sort are 2.2 times more likely to become homeruns in their field than a randomly chosen publication.
- Papers that center their references exclusively on new knowledge (those with low $< T >$ and low σ_T, Fig. 1.2B) have a surprisingly low impact, rarely exceeding what would be expected by chance. This means, the conventional bias towards heavily referencing recent work is misguided; more recent references are only valuable when supplemented with older references.

- Papers whose references are centered in “vintage” knowledge (papers with high $< T >$ and low σ_T, Fig. 3.5.2A) have a particularly low impact. Their likelihood of becoming homeruns in their field is about half of what is expected by chance. This is quite remarkable, given that some 27% of all papers fall in this category.

Figure 3.5.2E shows that being at the cutting edge is key for future impact, but only when paired with an appreciation for older work. Building only on recent knowledge amounts to throwing a dart to determine your paper’s impact—you are leaving it entirely to the whims of chance. Furthermore, the lowest impact works are those that are stuck in the past, building exclusively on vintage knowledge while oblivious of recent developments. These are not merely sweeping generalizations: Disaggregating the data by field reveals the same pattern of hotspots in nearly every branch of science. What’s more, the effect of referencing patterns has become more universal in recent decades. While at the beginning of the postwar era about 60% of fields displayed a clear link between hotspots and hits, by the 2000s, this link was detectable in almost 90% of research fields.

This means, despite important differences among scientific fields in methods, culture, and the use of data versus theory, the way a scientist chooses to build upon earlier work consistently influences whether or not her work will make a splash, no matter her field.

3.5.3 The Growing Impact of Older Discoveries

Access to scholarly knowledge has fundamentally changed over the years. Today’s students may never set foot in a library. Instead, they hunt for the relevant literature from their browsers, using powerful search engines like Google Scholar. And while they may be impressed by the heavy bookshelves and neat piles of printed journals in a professor’s office, it may be hard for them to imagine how he once relied on them to keep up with the latest developments in his field. Many scientists today first learn about their colleagues’ new research from social media and they feel satisfied when someone retweets their latest preprints.

But how do the changes in information access alter the way we build on past knowledge? On one hand, new tools and technologies allow us to reach deeper into the past [109-111]. Today every journal offers online access and a digital archive of its older articles, making all research accessible 24/7 for anyone with internet. These changes make older knowledge more accessible [111], ostensibly increasing the chances that scientists will build upon it, especially since search engines often allow users to see the most relevant results first, not just the most recent ones.

Yet there are also reasons to believe that these changes may in fact narrow the ideas scientists build upon [112]. Following hyperlinks is likely to steer researchers towards the current prevailing opinion, rather than something deep within the archives. Indeed, trips to the library might have been inefficient, but they also offered an opportunity for serendipity; flipping through dusty pages, scholars were forced to interact with findings from the distant past, noticing papers and knowledge they were not necessarily seeking. Therefore, digital publishing has the potential to make us more myopic, its convenience shifting us away from the classics and towards more recent work. Moreover, the digital tools we rely on also push us closer than ever to the frontier of knowledge. Today's scientist receives an email alert the moment a new paper in her field has been published or is shared on social media. She does not even need to wait for a journal publication to learn about the latest advances, as pre-print sharing (wherein a paper is circulated publicly before being submitted to a peer-reviewed journal) is now the norm in several disciplines. Taken together, these changes may further reduce the time gap between what scientists produce and what they build upon.

To see the true impact of digitization, researchers studied papers published between 1990 and 2013, counting how many cited papers that were at least 10 years old [111]. They found that the fraction of older citations grew steadily between 1990 and 2013, even accelerating after 2000. The shift was most pronounced between 2002 and 2013, when sophisticated search engines emerged. By 2013, four of the nine broad research areas had at least 40% citations to older articles, the "humanities, literature and arts" category leading the pack with 51%.

This trend is not limited to the last two decades. Plotting the average age of references over time, we find that since the 1960s, scientists have been systematically reaching deeper into the literature, citing older and older papers (Fig. 3.5.3). Why did such deep referencing start in the 1960s? While the answer is not entirely clear, some speculate that this change is rooted in the advent of peer review [90]. Before the 1960s, papers were accepted mainly at the editors' discretion [113]. By the mid-20th century the increasing specialization of science had intensified the need for expert opinions—yet disseminating a paper to geographically dispersed readers was difficult until photocopying became available in 1959 [114]. *Nature* officially introduced peer review in 1967, and the practice became mainstream throughout scholarly publishing shortly thereafter. It is likely that these reviewers could point out relevant older work for the authors to consider, not only boosting references to canonical literature, but also inducing a behavioral change, prompting authors to more carefully credit earlier work.

Together, these results indicate that technological shifts have made it easier to find the most relevant articles, irrespective of their age. Happily, important advances are no longer getting lost on shelves, but are able to continue to influence research for decades longer. But just how long can a researcher expect their corpus of work to remain relevant?

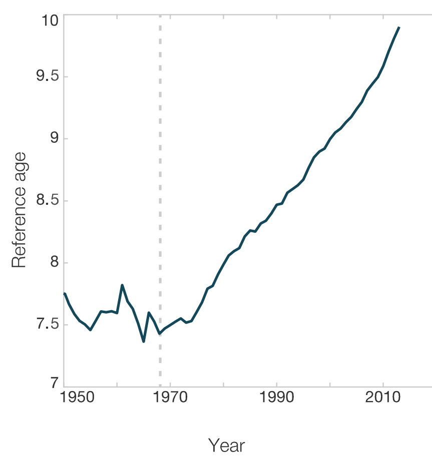


Figure 3.5.3 **Building on the Past**. The average age of the references of papers stayed around 7.5 years from 1950 to 1970, after which we observe a remarkable increasing trend, indicating that scientists are systematically reaching deeper into the literature. The dotted line marks the year 1967, when *Nature* officially introduced peer review. The fact that the start of peer review coincided with changes in referencing patterns has led researchers to hypothesize a link between the two [90].

Box 3.5.1 The age distribution of citations: Two distinct approaches.

Historically, two rather distinct approaches have been used to measure the age distribution of citations. The *retrospective (citation from)* approach considers papers cited by a publication during a particular year and analyzes the age distribution of these citations [39], looking back in time [115-117]. By contrast, the *prospective (citations to*) approach [39, 115-118] studies the distribution of citations gained over time by papers published in a given year.

Research shows that the two approaches are connected through precise mathematical relationships, allowing us to derive and even predict one approach from the other [17]. Yet, while both approaches measure the age of citations, they capture different processes. The retrospective approach, which we have used so far in this chapter, measures how far back in time the authors of a paper look as they cite earlier references, characterizing the citing "memory" of a paper. The prospective approach, on the other hand, represents a collective measure, quantifying how a particular paper is remembered over time. Hence, the prospective approach (which we will use next) is more useful for understanding how impact changes with time. Note, however, that studies prior to 2000 mostly relied on the retrospective approach. The reason is mainly computational: the retrospective approach is easier to implement, since the *citing* paper is fixed.

3.5.4 Your expiration date

As we saw in Fig. 3.5.1, our chance of referencing an older paper decays quickly with the paper's age. For the scientist, this prompts a critical, if uncomfortable, question: Do papers have an expiration date, after which they lose their relevance to the scientific community?

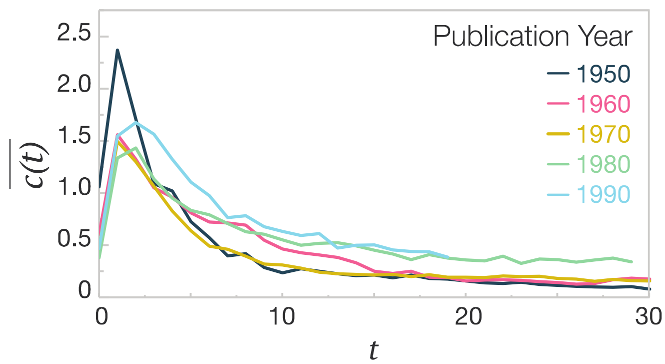


Figure 3.5.4 **The Jump-decay Pattern of Citations.** The average number of citations as a function of the time elapsed since publication. Each line represents a group of papers, the color denoting their publication year. As the curves indicate, most citations come within the first five years after publication; after that, the chance of a paper being cited again drops dramatically.

To answer this question, we can consider the average number of citations $\overline{c(t)}$ a paper acquires each year after publication. As Fig. 3.5.4 shows, citations follow a *jump-decay pattern* [66, 119]: a paper's citation rate rises quickly after publication, reaching a peak around year two or three, after which it starts to drop. In other words, a typical paper achieves its greatest impact within the first few years of its publication. After that, its impact diminishes quickly. As Box 3.5.2 explains, this jump-decay pattern is the rationale behind the two-year window used to measure the "impact factor," a metric used to evaluate journals.

The jump-decay citation pattern suggests that, if we wish to fairly compare the impacts of two papers, we must first factor in their age. A paper published in 2000 may have a higher citation count than a 2010 paper simply because it has been around longer (Fig. 3.5.5a). Yet, as Fig. 3.5.5B indicates, a paper's cumulative citations do not increase indefinitely, but saturate after a few years. For example, by 2006 the citation distributions of papers published in 1991 or 1993 were statistically indistinguishable (Fig. 3.5.5b). In other words, the citation distributions converge to a steady-state form after a certain time period (with that period varying from journal to journal) [40]. From that point on, the number of new citations a paper garners will be negligible—those papers have reached their expiration date.

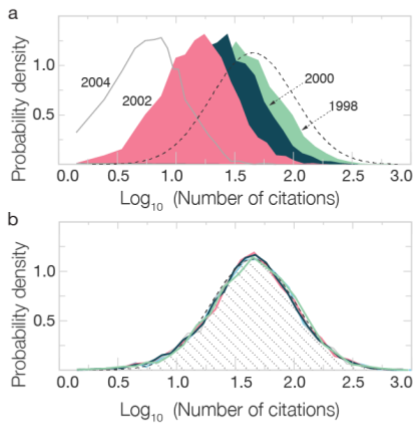


Figure 3.5.5 **Time evolution of citation distributions.** Taking papers published by *Journal of Biological Chemistry*, Panel a shows the distribution of citations accrued by the end of 2006 [40]. Papers published in 2004 constitute the far-left curve, because they had only two years to collect citations. Papers published in earlier years, which have had more time to collect citations, are further to the right. However, as papers get older, this temporal shift wanes. In Panel b, we consider the citations to papers published between 1991 and 1993. Since those published in 1991 have mostly stopped acquiring new citations by 2006, their citation distribution is identical to that of papers published in 1993. After Stringer *et al.* [40].

Box 3.5.2 Impact factor. The impact factor (IF), often used to quantify the importance of a journal, captures the average number of citations that papers published in that journal collect within two years of publication. *Nature*, for example, has an impact factor of 38.1 for the year 2015, while *Physical Review Letters* has an impact factor of 7.6, *American Sociological Review* 4.3, and *Journal of Personality and Social Psychology* 4.7. Lately, however, the impact factor has started to take on a new, concerning function. When a paper is published in a certain journal, readers increasingly use that journal's impact factor to evaluate the potential importance of the paper. This is as misguided as trying to judge a book by its cover. Indeed, as Fig. B3.5.6 shows, papers published within the same journal can have vastly different impacts, telling us that the impact factor cannot predict an individual paper's impact.

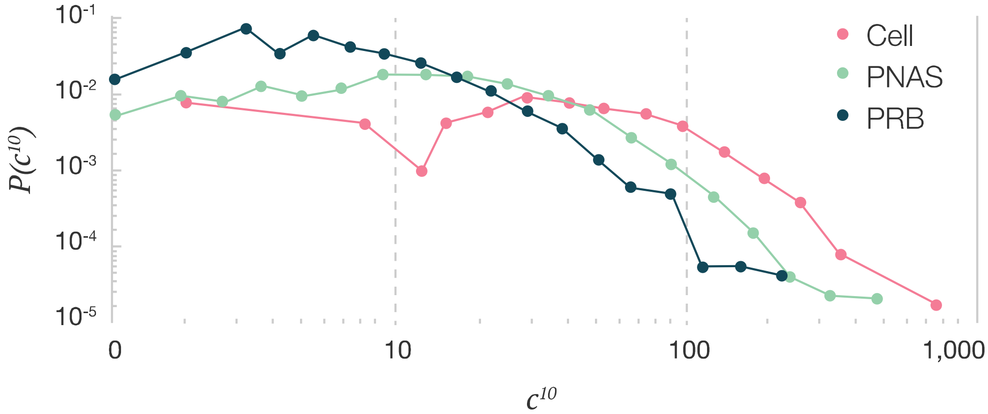


Figure B3.5.6: **Impact factor does not predict the impact of individual papers.** (A) Distribution of the cumulative citations ten years after publication (c^{10}) for all papers published in *Cell* (IF=33.62 in 1992), *PNAS* (IF=10.48), and *Physical Review B* (*PRB*) (IF = 3.26). Papers published by the same journal have vastly different impacts, despite sharing the same impact factor.

Summary

At the beginning of this chapter, we posed a difficult question: If you were at the Library of Alexandria watching the flames grow nearer, which research would you save—the old, or the new? The issues discussed in this chapter not only help us to answer this hypothetical question, but may also help scientists more realistically understand the lifespan of their work.

As we have discussed, scientists follow reproducible patterns when drawing upon prior knowledge. The cutting edge of new research is integral to the discovery process, given that the vast majority of references are to papers published within the last two to three years. Yet citing work from the frontier does not by itself guarantee impact: Papers with a judicious mix of new and canonical knowledge are twice as likely to be homeruns than typical papers in the field. Therefore, while building upon cutting-edge work is key, a paper's impact is lifted when it considers a wide horizon of research. Put simply, if the person at the Library Alexandria wanted to ensure that great scientific discoveries would still be possible going forward, she should try to stack up on newly-published scrolls, but also grab a few older ones.

In general, a paper garners most of its citations within the first couple years of publication. However, old papers are never entirely forgotten—hence, focusing exclusively on the impact garnered within the first few years is likely to miss the long-term impact of a discovery. That means it is only fair to compare the impact of two papers if they have had a similar amount of time to be cited, or if both papers have been around long enough that they have already collected the bulk of their citations. Often, we may not know a paper’s true impact for more than a decade.

For scientists, there is another takeaway: Each of our papers have an expiration date. Whether we like it or not, there will come a point when every paper we have written stops being relevant to the community. However, this dark cloud has a silver lining: Thanks to digitization, older articles are getting more and more attention, indicating that the collective horizon of science is expanding.

But if every paper will sooner or later stop acquiring citations, what is the “ultimate” impact for each paper? And can we estimate a paper’s total impact in advance?

Chapter 3.6

Ultimate impact

The jump-decay citation patterns described in the last chapter calls to mind the notion of "15 minutes of fame." Are we left to conclude that the papers we publish—works we spend countless days and nights researching, writing, and agonizing over—will only be read for at most a few years after publication? Figure 3.6.1 shows the yearly citation count for 200 randomly selected papers published between 1960 and 1970 in *Physical Review*, conveying a clear message: citation patterns for individual papers are complicated. While, in aggregate citations may follow a uniform jump-decay pattern, citations of individual papers do not appear to be dictated by any apparent temporal pattern. Instead, they are remarkably variable. While most work is forgotten shortly after publication, a few papers seem to have an especially long lifespans. How can we make sense of this diverse set of trajectories?

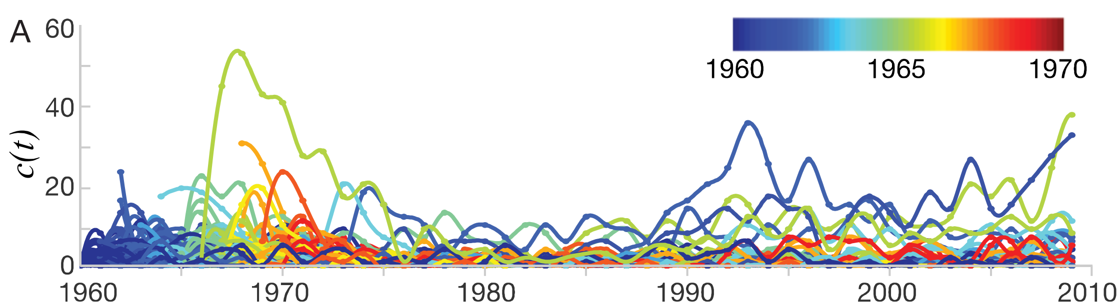


Figure 3.6.1 **Citation histories of individual papers**. We randomly selected 200 papers published between 1960 and 1970 in *Physical Review*, showing how many citations each acquired yearly following its publication. The color of the lines indicates the publication year. Blue papers were published around 1960, whereas the red were published closer to 1970.

If all papers were to follow the same jump-decay pattern, then we could easily determine the future impact of any paper after just its first few years. However, since individual discoveries take wildly different paths toward their ultimate impact, estimating the future impact of a paper appears a hopeless task. Making the matter worse is the fact that many influential discoveries are notoriously under-appreciated at first—research suggests that the more a discovery deviates from the current paradigm, the longer it will take to be appreciated by the community [120]. Indeed, given the myriad factors that affect how a new discovery is received—from the work's intrinsic value to its timing, publishing venue, and sheer chance—finding regularities in the citations of individual papers remains elusive.

Yet, as we will show in this chapter, beneath the randomness and the apparent lack of order lie some striking patterns. And these patterns make the citation dynamics of individual papers quite predictable.

3.6.1 Citation Dynamics of Individual Papers

Let's recap the mechanisms that we have already shown to affect the impact of a paper [66]:

- First is the *exponential growth* of science (Ch. **3.1**). In order for papers to gain new citations, new papers must be published; hence, the rate at which these new papers are published affects how existing papers will accumulate citations.
- Second, *preferential attachment* captures the fact that highly cited papers are more visible and thus more likely than their less-cited counterparts to be cited again (Ch. **3.3**).
- Third, *fitness* captures the inherent differences between papers, accounting for the perceived novelty and importance of a discovery (Ch. **3.3**).
- And lastly, *Aging* captures how new ideas are integrated into subsequent work: Every paper's propensity for citations eventually fades (Ch. **3.5**), in a fashion best described by a log-normal survival probability.

What do these factors tell us about citation patterns? It turns out that we can combine these features in a mathematical model, and then solve the model analytically (see Appendix 1.5 for more detail), arriving to a formula describing the cumulative number of citations acquired by paper *i* at time *t* after publication:

$$c_i^t = m\left(\mathrm{e}^{\lambda_i \phi\left(\frac{\ln t - \mu_i}{\sigma_i}\right)} - 1\right) \qquad (3.6.1)$$

Here, $\phi(x) \equiv \frac{2}{\sqrt{\pi}} \int_{-\infty}^{x} e^{-y^2/2} dy$ is the cumulative normal distribution, and m corresponds to the average number of references each new paper cites. As Fig. 3.6.2 illustrates, for different parameters (3.6.1) describes a wide range of citation trajectories, helping to account for the different citation patterns observed in Fig. 3.6.1.

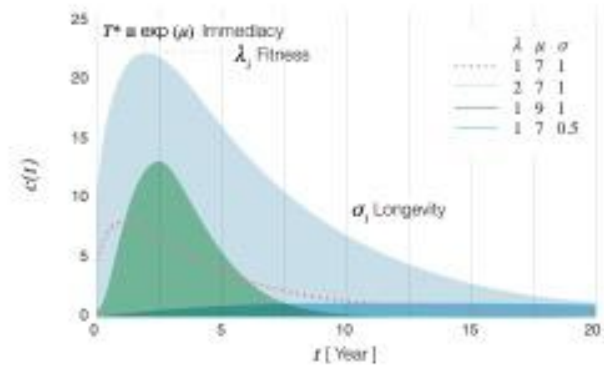


Figure 3.6.2 **Illustrations of Eq. (3.6.1)**. The citation history of paper *i* is characterized by three parameters: (1) the *relative fitness* λ_i capturing a paper's ability to attract citations relative to other papers; (2) the *immediacy* μ_i captures how quickly a paper will get attention from the community, governing the time required for a paper to reach its citation peak, and (3) the *longevity* σ_i captures how quickly the attention decays over time.

Equation (3.6.1) makes a remarkable prediction: as noisy and unpredictable as citation patterns of individual papers may seem (Fig. 3.6.1), they are all governed by the same universal equation. The differences between papers can be reduced to differences in three fundamental parameters, λ, μ, and σ. Indeed, Eq. (3.6.1) predicts that if we know the $(\lambda_i, \mu_i, \sigma_i)$ parameters for each paper and rescale the formula accordingly, using $\tilde{t} \equiv (\ln t - \mu_i) / \sigma_i$ and $\tilde{c} \equiv \ln\left(1 + \frac{c_i^t}{m}\right)/\lambda_i$, then each paper's citation history must follow the same universal curve:

$$\tilde{c} = \Phi(\tilde{t}), \quad (3.6.2)$$

In other words, all citation histories, as different as they seem, appear to be governed by a single formula. Next, we will test whether the empirical evidence bears out this unexpected prediction.

3.6.2 Citation Dynamics: Remarkably Universal

Let us select four papers whose citation histories are clearly dissimilar (Fig. 3.6.3A). Yet, Eq. (1.1) tells us that, if we obtain the set of $(\lambda_i, \mu_i, \sigma_i)$ parameters that best describe each paper's citation history and rescale them following (3.6..2), all these different curves must collapse into a single one. As Fig. 3.6.3B illustrates, that is precisely what happens. In other words, even though we selected these papers specifically for their diverse citation histories, none of them appears to defy this universal pattern. In fact, we could choose any number of papers published in different decades by different journals in different disciplines, as we did with some 8,000 physics papers in Fig. 3.6.3c, and according to Eq. (1.2) they should all collapse on the same universal curve. As our sample of papers grows, there are inevitably some that have had bizarre and irregular citation histories (see also Box 3.6.1). Yet, remarkably, once we determine the parameters for each paper and rescale their trajectories accordingly, their citation histories largely follow the same universal curve (1.2).

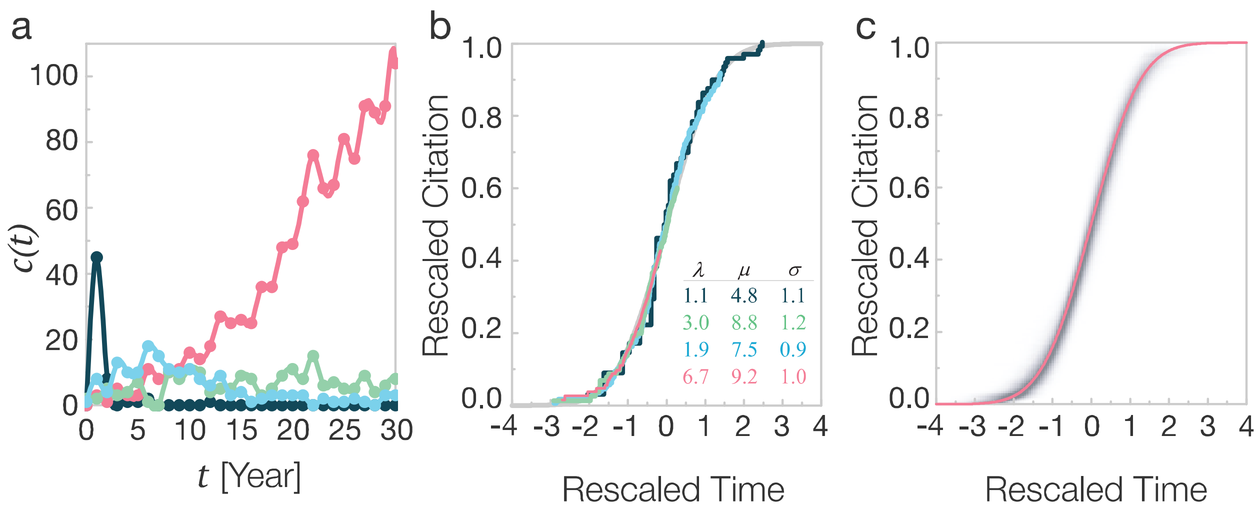


Figure 3.6.3 **Universality in Citation dynamics.** (a) Citation histories of four papers published in Physical Review in 1964, selected for their distinct patterns: a "jump-decay" pattern (blue), a delayed peak (magenta), a constant number of citations over time (green), and an increasing number of citations each year (red). (b) Citations of an individual paper are determined by three parameters: fitness λ_i, immediacy μ_i, and longevity σ_i. Upon rescaling the citation history of each paper in (a) by the appropriate $(\lambda_i, \mu_i, \sigma_i)$ parameters, the four papers collapse into a single universal function. [66]. (c) We rescaled all the papers published between 1950 and 1980 in *Physical Review* that garnered more than 30 citations in 30 years (~8,000 papers). After Wang [66].

As such, all observed differences in citation histories can be attributed to three measurable parameters: fitness, immediacy and longevity. This reveals a surprising degree of regularity in a system that seems noisy, unpredictable, and driven by countless other factors. This regularity is rooted in the fact that citations are a collective measure, reflecting the collective opinion of the scientific community on a paper's importance. Hence individual actions do not alter much a paper's path to success or oblivion—it is the collective action of countless scientists that shape their impact. As such, recognition from the community as a collective follows highly reproducible patterns and is also dominated by a small number of detectable mechanisms.

But why does this universality matter?

Box 3.6.1 Sleeping beauties and second acts

There are some papers whose citations deviate from the typical rise-and-fall trajectories of Eq. (1.2). These come in two categories. First, there are "sleeping beauties," papers whose importance is not recognized until years after publication [121-123] The classic paper by Garfield [124] that introduced citation indices offers a fine example (Fig. B3.6.4). Published in 1955, this sleeping beauty only awoke after 2000, thanks to an increased interest in citation networks. Other sleeping beauties only awaken when they are independently discovered by a new scientific community. One example is the 1959 mathematics paper by Paul Erdős and Alfréd Rényi on random networks [125], whose impact exploded in the 21st century following the emergence of network science (Fig. B3.6.4).

The "Theory of Superconductivity" paper by Bardeen, Cooper, and Schrieffer (BCS) [76] offers an example of a "second act," [39] another form of atypical trajectories (Fig. B3.6.4). The paper took off quickly after its publication in 1957, and even won its authors the Nobel Prize in physics in 1972. Yet, soon the enthusiasm dwindled, bottoming out in 1985. However, the 1986 discovery of high-temperature superconductivity made the BCS paper relevant again, helping it experience a "second act".

These atypical trajectories are not particularly rare [122]. Indeed, in multi-disciplinary journals more than 7% of papers can be classified as sleeping beauties. In specialized disciplines such as biology, chemistry, statistics, and physics, the fraction is around 2-3%. We can extend the citation model discussed in the previous section to include a second peak, helping to predict about 90% of the atypical cases [123].

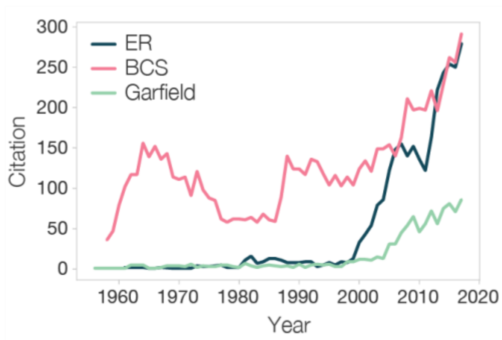


Figure B3.6.4 **Atypical Citation Histories**. Garfield's classic paper on citations [124] is an exemplary case of a sleeping beauty (green curve). Similarly, the Erdős-Rényi paper (blue curve) was highly regarded within mathematics but had only limited impact outside the field. However, the emergence of network science in 1999 drew new, multidisciplinary attention to the paper, fueling its explosive citation count. The superconductivity paper by Bardeen, Cooper, and Schrieffer (orange curve) experienced a second act when high-temperature superconductivity was discovered.

3.6.3 Ultimate Impact

While the ups and downs of a discovery's impact presents a fascinating puzzle, most scientists are more interested in the net result of all these fluctuations: They want to know what a paper's cumulative impact will be when all is said and done. Equation (1.1) offers an elegant way to calculate this cumulative impact. Indeed, if we set the time period to infinity, then (1.1) represents the total number of citations a paper will ever acquire during its lifetime, or its *ultimate impact* (c^{∞}),

$$c_i^{\infty} = m\left(\mathrm{e}^{\lambda_i} - 1\right). \tag{1.3}$$

Recall that we needed three parameters to describe the time-dependent citation count of a paper. Yet, when it comes to ultimate impact, according to (1.3), only one parameter is relevant: the relative fitness λ. It does not matter how soon the paper starts to garner attention (immediacy, μ) or how fast its appeal decays over time (longevity, σ). *Its ultimate impact is determined by its relative fitness only*, capturing the paper's importance relative to its peers.

As such, we may evaluate the long-term impact of a paper without considering the journal in which it was published. To illustrate this, we can pick papers with a comparable fitness $\lambda \approx 1$, but published in three very different journals and follow their citation histories over their first 20 years. Despite being published in journals with widely different readership and impact, the total citation counts acquired by these papers are remarkably similar (Fig. 3.6.5). This is precisely what (1.3) tells us: as they had similar fitness λ, they must ultimately have the same impact, $c^{\infty} = 51.5$. The takeaway: Just as we don't judge a book by its cover, we can't judge a paper by its journal. *Cell* may appear to be a more coveted venue than *PNAS*—but as long as the papers have the same fitness, ultimately they will acquire the same number of citations.

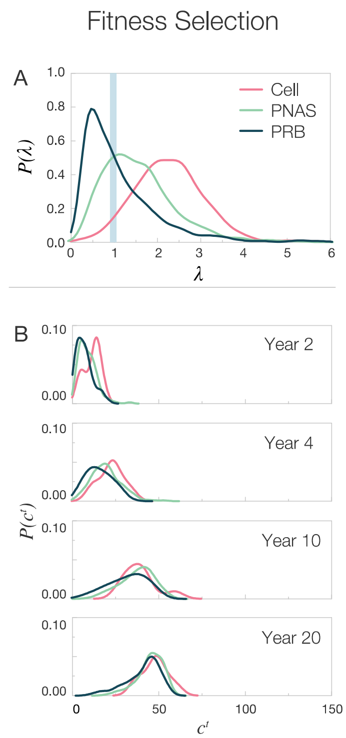


Figure 3.6.5 **Ultimate Impact**. We select three journals with different readership and impact: 1) *Physical Review B* (*PRB*), with an impact factor (IF) around 3.26 in 1992, is the largest journal within the *Physical Review* family, covering a specialized domain of physics; 2) *PNAS* (IF=10.48) is a high-impact multi-disciplinary journal covering all areas of science; and 3) *Cell* (IF=33.62) is an elite biology journal. For each paper published in these journals, we measure the fitness λ, obtaining their distinct journal-specific fitness distribution (A). We then select all papers with comparable fitness $\lambda \approx 1$ published in each journal and follow their citation histories. As expected, the paper's particular path depends on the journal's prestige: In early years, *Cell* papers run slightly ahead and *PRB* papers stay behind, resulting in distinct citation distributions from year 2 to year 4 ($T = 2 \div 4$). Yet, by year 20 the cumulative number of citations acquired by these papers converges in a remarkable way (B). This is precisely what (1.3) tells us: given their similar fitness λ, eventually these papers should have the same ultimate impact $c^{\infty} = 51.5$. After Wang *et al.* [66].

3.6.4 Future Impact

Figure 3.6.6 shows a typical weather forecast, the kind that often airs on TV news. It warns of a hurricane brewing in the Caribbean and predicts where and when it will hit Tennessee. This kind of detailed prediction can help residents and emergency personnel make better plans, to stock food and seek shelter.

Figure 3.6.6 **Hurricane Forecast.** Although the storm is thousands of miles away from Tennessee, residents know well in advance when and where it is going to hit their state and its likely intensity upon arrival. Can we develop similar predictive tools for science?

How did scientists acquire such a predictive capability? First, massive amounts of data on past hurricane trajectories allowed them to study in minute detail the patterns characterizing these trajectories and the physical factors that drive them. Then, using that information, they built predictive models that tell us where hurricanes tend to go, given where they have already been. Can we use the same approach to predict the impact of a publication?

It turns out, we can adapt the model introduced in this chapter to predict future citations of each paper. For this we use a paper's citation history up to year T_{Train} after publication to train the model, allowing us to estimate the parameters λ_i, μ_i, σ_i for the paper. We illustrate this in Fig. 3.6.7A, where we use a five-year training period to predict the subsequent trajectories of three papers. The figure shows the predicted most likely citation path (red line) with an uncertainty envelope (shaded area) for each paper. Comparing our predictions to the actual citation histories, we find that two of the three papers indeed fell within the envelope. For the third paper, the model overestimated future citations. But using a longer training period ameliorates this problem. Indeed, if we increase the training period from 5 years to 10 years, the cone of predictions shrinks (Fig. 3.6.7B) and the model's predictive accuracy improves: The third paper now falls well within the prediction envelope, while the expected paths of the other two papers now align even more closely with their real-life citation trajectories.

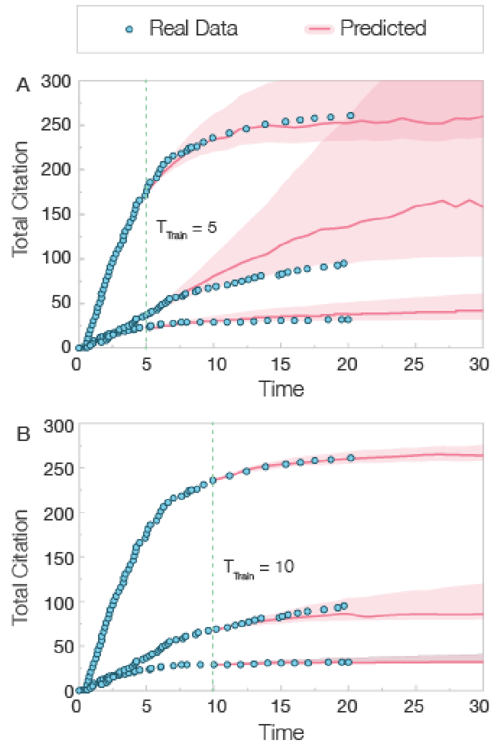


Figure 3.6.7 **Predicting future citations**. We can adapt the citation model to predict future citations of any given paper, by learning the model parameters from its existing citation histories. The uncertainty in estimating these parameters then translate into a prediction envelop with its most likely trajectory, similar to what we saw in the hurricane's example. As we observe more of its citation records, the envelop shrinks, and more accurately encapsulates a paper's citation history. After Wang *et al.* [66].

In summary, each paper follows wildly different paths on its way to achieving its ultimate impact. Yet, as we showed in this chapter, this apparent variability in citation dynamics hides a remarkable degree of regularity. Indeed, citation histories can be accurately captured by a simple model with just three parameters: fitness, immediacy, and longevity. Once we discern the three parameters characterizing a paper's citation history, we can determine how and when the paper will accumulate its citations.

Importantly, while three parameters are required to determine the particular pattern of citation growth, when it comes to predicting the paper's ultimate impact, representing the total number of citations accumulated over a paper's lifetime, only one parameter, the fitness, matters. Indeed, as we show, papers that have the same fitness acquire the same number of citations in the long run, regardless of where they are published. In contrast to impact factor and short-term citation count, which lack predictive power, ultimate impact offers a journal-independent measure of a paper's long-term impact.

The predictive accuracy of the model raises an interesting question: Could the act of predicting an article's success alter that success—or the scientific discourse at large? [126] There is reason to think so. Widespread, consistent prediction of an idea's future impact will necessarily speed up its acceptance and may undermine ideas that are predicted to fare worse, by suggesting that they do not deserve our attention or resources. As such, predicting an article's reception could act as a self-fulfilling prophecy, leading to the premature abortion of valuable ideas. On the plus side, however, the ability to better predict an article's future impact could accelerate a paper's life cycle—the time between a paper's publication and its acceptance in the community.

Bibliography

1. Maury, J.-P., *Newton: understanding the cosmos*. 1992: Thames & Hudson.
2. de Solla Price, D., *Science since babylon*. 1961, Yale University Press.
3. de Solla Price, D., *Little science, big science*. 1963: Columbia University Press. 119.
4. Gilbert, G.N. and S. Woolgar, *The quantitative study of science: An examination of the literature.* Science Studies, 1974: p. 279-294.
5. Sinatra, R., et al., *Quantifying the evolution of individual scientific impact.* Science, 2016. **354**(6312): p. aaf5239.
6. Khabsa, M. and C.L. Giles, *The number of scholarly documents on the public web.* PLOS one, 2014. **9**(5): p. e93949.
7. Sinha, A., et al. *An overview of microsoft academic service (mas) and applications*. in *Proceedings of the 24th international conference on world wide web*. 2015. ACM.
8. Bacon, F., *On innovations.* Bacon's essays... in two volumes, 1881. **1**: p. 84-85.
9. Baldwin, M., *'Keeping in the race': physics, publication speed and national publishing strategies in Nature, 1895–1939.* The British Journal for the History of Science, 2014. **47**(2): p. 257-279.
10. Editorial, *Form follows need.* Nature Physics, 2016. **12**.
11. Csiszar, A., *The scientific journal: Authorship and the politics of knowledge in the nineteenth century*. 2018: University of Chicago Press.
12. Cyranoski, D., et al., *Education: the PhD factory.* Nature news, 2011. **472**(7343): p. 276-279.
13. Wendler, C., et al., *The Path Forward: The Future of Graduate Education in the United States.* Educational Testing Service, 2010.
14. Schools, C.o.G., *Ph. D. completion and attrition: Policy, numbers, leadership, and next steps*. 2004, Author Washington, DC.
15. Schillebeeckx, M., B. Maricque, and C. Lewis, *The missing piece to changing the university culture.* Nature biotechnology, 2013. **31**(10): p. 938-941.
16. Merton, R.K., *Singletons and multiples in scientific discovery: A chapter in the sociology of science.* Proceedings of the American Philosophical Society, 1961: p. 470-486.
17. Yin, Y. and D. Wang, *The time dimension of science: Connecting the past to the future.* Journal of Informetrics, 2017. **11**(2): p. 608-621.
18. Vale, R.D., *Accelerating scientific publication in biology.* Proceedings of the National Academy of Sciences, 2015. **112**(44): p. 13439-13446.
19. Woolston, C., *Graduate survey: a love-hurt relationship.* Nature, 2017. **550**(7677): p. 549-552.
20. Powell, K., *The future of the postdoc.* Nature, 2015. **520**(7546): p. 144.
21. Zolas, N., et al., *Wrapping it up in a person: Examining employment and earnings outcomes for Ph. D. recipients.* Science, 2015. **350**(6266): p. 1367-1371.
22. Editorial, *Make the most of PhDs.* Nature, 2015. **528**.
23. Bloom, N., et al., *Are ideas getting harder to find?* 2017, National Bureau of Economic Research.
24. Milojevic, S., *Quantifying the cognitive extent of science.* Journal of Informetrics, 2015. **9**(4): p. 962-973.
25. Dong, Y., et al. *A century of science: Globalization of scientific collaborations, citations, and innovations*. in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2017. ACM.
26. Van Noorden, R., et al., *The top 100 papers.* Nature, 2014. **514**(7524): p. 550-553.
27. de Solla Price, D.J., *Networks of Scientific Papers.* Science, 1965. **149**(3683): p. 510-515.
28. Garfield, E. and I.H. Sher, *New factors in the evaluation of scientific literature through citation indexing.* American Documentation, 1963. **14**(3): p. 195-201.
29. Pareto, V., *Cours d'économie politique*. 1964, Librairie Droz.
30. Barabási, A.-L., *Network science*. 2015: Cambridge University.
31. Vázquez, A., *Statistics of citation networks.* arXiv preprint cond-mat/0105031, 2001.
32. Lehmann, S., B. Lautrup, and A.D. Jackson, *Citation networks in high energy physics.* Physical Review E, 2003. **68**(2): p. 026113.

33. Radicchi, F., S. Fortunato, and C. Castellano, *Universality of citation distributions: Toward an objective measure of scientific impact.* Proceedings of the National Academy of Sciences, 2008. **105**(45): p. 17268-17272.
34. Seglen, P.O., *The skewness of science.* Journal of the American Society for Information Science, 1992. **43**(9): p. 628-638.
35. Bommarito, I.I., J. Michael, and D.M. Katz, *Properties of the United States code citation network.* arXiv preprint arXiv:0911.1751, 2009.
36. Eom, Y.-H. and S. Fortunato, *Characterizing and modeling citation dynamics.* PloS one, 2011. **6**(9): p. e24926.
37. Menczer, F., *Evolution of document networks.* Proceedings of the National Academy of Sciences, 2004. **101**(suppl 1): p. 5261-5265.
38. Radicchi, F. and C. Castellano, *Rescaling citations of publications in physics.* Physical Review E, 2011. **83**(4): p. 046116.
39. Redner, S., *Citation Statistics from 110 Years of Physical Review.* Physics Today, 2005. **58**: p. 49.
40. Stringer, M.J., M. Sales-Pardo, and L.A.N. Amaral, *Effectiveness of journal ranking schemes as a tool for locating information.* Plos one, 2008. **3**(2): p. e1683.
41. Castellano, C. and F. Radicchi, *On the fairness of using relative indicators for comparing citation performance in different disciplines.* Archivum immunologiae et therapiae experimentalis, 2009. **57**(2): p. 85-90.
42. Stringer, M.J., M. Sales-Pardo, and L.A.N. Amaral, *Statistical validation of a global model for the distribution of the ultimate number of citations accrued by papers published in a scientific journal.* Journal of the American Society for Information Science and Technology, 2010. **61**(7): p. 1377-1385.
43. Wallace, M.L., V. Larivière, and Y. Gingras, *Modeling a century of citation distributions.* Journal of Informetrics, 2009. **3**(4): p. 296-303.
44. Anastasiadis, A.D., et al., *Tsallis q-exponential describes the distribution of scientific citations—a new characterization of the impact.* Scientometrics, 2010. **83**(1): p. 205-218.
45. van Raan, A.F., *Two-step competition process leads to quasi power-law income distributions: Application to scientific publication and citation distributions.* Physica A: Statistical Mechanics and its Applications, 2001. **298**(3): p. 530-536.
46. Van Raan, A.F., *Competition amongst scientists for publication status: Toward a model of scientific publication and citation distributions.* Scientometrics, 2001. **51**(1): p. 347-357.
47. Kryssanov, V.V., et al., *We cite as we communicate: A communication model for the citation process.* arXiv preprint cs/0703115, 2007.
48. Barabási, A.-L., C. Song, and D. Wang, *Publishing: Handful of papers dominates citation.* Nature, 2012. **491**(7422): p. 40-40.
49. Aksnes, D.W., *Citation rates and perceptions of scientific contribution.* Journal of the American Society for Information Science and Technology, 2006. **57**(2): p. 169-185.
50. Radicchi, F., *In science "there is no bad publicity": Papers criticized in comments have high scientific impact.* Scientific reports, 2012. **2**.
51. Moravcsik, M.J. and P. Murugesan, *Some results on the function and quality of citations.* Social studies of science, 1975. **5**(1): p. 86-92.
52. Cole, J.R. and S. Cole, *Social stratification in science.* 1973.
53. Cronin, B., *RESEARCH BRIEF RATES OF RETURN TO CITATION.* Journal of documentation, 1996. **52**(2): p. 188-197.
54. Lawani, S.M. and A.E. Bayer, *Validity of citation criteria for assessing the influence of scientific publications: New evidence with peer assessment.* Journal of the American Society for Information Science, 1983. **34**(1): p. 59-66.
55. Luukkonen, T., *Citation indicators and peer review: their time-scales, criteria of evaluation, and biases.* Research Evaluation, 1991. **1**(1): p. 21-30.
56. Oppenheim, C. and S.P. Renn, *Highly cited old papers and the reasons why they continue to be cited.* Journal of the American Society for Information Science, 1978. **29**(5): p. 225-231.
57. Rinia, E.J., et al., *Comparative analysis of a set of bibliometric indicators and central peer review criteria: Evaluation of condensed matter physics in the Netherlands.* Research policy, 1998. **27**(1): p. 95-107.
58. Radicchi, F., A. Weissman, and J. Bollen, *Quantifying perceived impact of scientific publications.* Journal of Informetrics, 2017. **11**(3): p. 704-712.
59. Barabási, A.-L., *The Formula: The Universal Laws of Success*. 2018: Hachette UK.

60. Jaffe, A.B., *Patents, patent citations, and the dynamics of technological change.* NBER Reporter, 1998(Summer): p. 8-11.
61. Jaffe, A.B., M.S. Fogarty, and B.A. Banks, *Evidence from patents and patent citations on the impact of NASA and other federal labs on commercial innovation.* The Journal of Industrial Economics, 1998. **46**(2): p. 183-205.
62. Trajtenberg, M., *A penny for your quotes: patent citations and the value of innovations.* The Rand Journal of Economics, 1990: p. 172-187.
63. Hall, B.H., A.B. Jaffe, and M. Trajtenberg, *Market value and patent citations: A first look*. 2000, national bureau of economic research.
64. Harhoff, D., et al., *Citation frequency and the value of patented inventions.* Review of Economics and statistics, 1999. **81**(3): p. 511-515.
65. Price, D.d.S., *A general theory of bibliometric and other cumulative advantage processes.* Journal of the American society for Information science, 1976. **27**(5): p. 292-306.
66. Wang, D., C. Song, and A.-L. Barabási, *Quantifying long-term scientific impact.* Science, 2013. **342**(6154): p. 127-132.
67. Eggenberger, F. and G. Pólya, *Über die statistik verketteter vorgänge.* ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik, 1923. **3**(4): p. 279-289.
68. Yule, G.U., *A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FRS.* Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character, 1925: p. 21-87.
69. Gibrat, R., *Les inégalités économiques*. 1931: Recueil Sirey.
70. Zipf, G.K., *Human behavior and the principle of least effort*. 1949: addison-wesley press.
71. Simon, H.A., *On a class of skew distribution functions.* Biometrika, 1955: p. 425-440.
72. Merton, R.K., *The Matthew effect in science.* Science, 1968. **159**(3810): p. 56-63.
73. Barabási, A.-L. and R. Albert, *Emergence of scaling in random networks.* science, 1999. **286**(5439): p. 509-512.
74. Newman, M., *Networks: an introduction*. 2010: Oxford University Press.
75. Newman, M.E.J., *The first-mover advantage in scientific publication.* EPL (Europhysics Letters), 2009. **86**(6): p. 68001.
76. Bardeen, J., L.N. Cooper, and J.R. Schrieffer, *Theory of superconductivity.* Physical Review, 1957. **108**(5): p. 1175.
77. Bianconi, G. and A.-L. Barabási, *Competition and multiscaling in evolving networks.* EPL (Europhysics Letters), 2001. **54**(4): p. 436.
78. Bianconi, G. and A.-L. Barabási, *Bose-Einstein condensation in complex networks.* Physical Review Letters, 2001. **86**(24): p. 5632.
79. Fleming, L., S. Mingo, and D. Chen, *Collaborative brokerage, generative creativity, and creative success.* Administrative science quarterly, 2007. **52**(3): p. 443-475.
80. Youn, H., et al., *Invention as a combinatorial process: evidence from US patents.* Journal of The Royal Society Interface, 2015. **12**(106): p. 20150272.
81. Robbin, J., S. Jobs, and P. Schiller, *Method and apparatus for use of rotational user inputs*. 2003, Google Patents.
82. Wang, J., R. Veugelers, and P. Stephan, *Bias against novelty in science: A cautionary tale for users of bibliometric indicators.* Research Policy, 2017. **46**(8): p. 1416-1436.
83. Lee, Y.-N., J.P. Walsh, and J. Wang, *Creativity in scientific teams: Unpacking novelty and impact.* Research policy, 2015. **44**(3): p. 684-697.
84. Uzzi, B., et al., *Atypical combinations and scientific impact.* Science, 2013. **342**(6157): p. 468-472.
85. Phiel, C.J., et al., *Histone deacetylase is a direct target of valproic acid, a potent anticonvulsant, mood stabilizer, and teratogen.* Journal of Biological Chemistry, 2001. **276**(39): p. 36734-36741.
86. Stephan, P., R. Veugelers, and J. Wang, *Reviewers are blinkered by bibliometrics.* Nature News, 2017. **544**(7651): p. 411.
87. Van Noorden, R., *Interdisciplinary research by the numbers.* Nature, 2015. **525**(7569): p. 306-307.
88. Wagner, C.S., et al., *Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature.* Journal of informetrics, 2011. **5**(1): p. 14-26.
89. Larivière, V., S. Haustein, and K. Börner, *Long-distance interdisciplinarity leads to higher scientific impact.* Plos one, 2015. **10**(3): p. e0122565.
90. Sinatra, R., et al., *A century of physics.* Nature Physics, 2015. **11**(10): p. 791-796.

91. Leahey, E. and J. Moody, *Sociological innovation through subfield integration.* Social Currents, 2014. **1**(3): p. 228-256.
92. Foster, J.G., A. Rzhetsky, and J.A. Evans, *Tradition and innovation in scientists' research strategies.* American Sociological Review, 2015. **80**(5): p. 875-908.
93. Fleming, L., *Breakthroughs and the" long tail" of innovation.* MIT Sloan Management Review, 2007. **49**(1): p. 69.
94. Boudreau, K.J., et al., *Looking across and looking beyond the knowledge frontier: Intellectual distance, novelty, and resource allocation in science.* Management Science, 2016. **62**(10): p. 2765-2783.
95. Bromham, L., R. Dinnage, and X. Hua, *Interdisciplinary research has consistently lower funding success.* Nature, 2016. **534**(7609): p. 684.
96. Kim, J., C.-Y. Lee, and Y. Cho, *Technological diversification, core-technology competence, and firm growth.* Research Policy, 2016. **45**(1): p. 113-124.
97. Phillips, D.P., et al., *Importance of the lay press in the transmission of medical knowledge to the scientific community.* The New England journal of medicine, 1991. **325**(16): p. 1180.
98. Gonon, F., et al., *Why most biomedical findings echoed by newspapers turn out to be false: the case of attention deficit hyperactivity disorder.* PloS one, 2012. **7**(9): p. e44275.
99. Dumas-Mallet, E., et al., *Poor replication validity of biomedical association studies reported by newspapers.* PloS one, 2017. **12**(2): p. e0172650.
100. Peng, R.D., *Reproducible research in computational science.* Science, 2011. **334**(6060): p. 1226-1227.
101. Open Science Collaboration, *Estimating the reproducibility of psychological science.* Science, 2015. **349**(6251): p. aac4716.
102. Ioannidis, J.P., *Why most published research findings are false.* PLoS medicine, 2005. **2**(8): p. e124.
103. Wakefield, A.J., et al., *RETRACTED: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children*. 1998, Elsevier.
104. Catalini, C., N. Lacetera, and A. Oettl, *The incidence and role of negative citations in science.* Proceedings of the National Academy of Sciences, 2015. **112**(45): p. 13823-13826.
105. Fang, F.C. and A. Casadevall, *Retracted science and the retraction index.* Infection and immunity, 2011. **79**(10): p. 3855-3859.
106. Sandison, A., *Densities of use, and absence of obsolescence, in physics journals at MIT.* Journal of the American Society for Information Science, 1974. **25**(3): p. 172-182.
107. Candia, C., et al., *The universal decay of collective memory and attention.* Nature Human Behaviour, 2019. **3**(1): p. 82.
108. Mukherjee, S., et al., *The Age of Past Knowledge and Tomorrow's Scientific and Technological Breakthroughs.* Science Advances, 2017.
109. Odlyzko, A., *The rapid evolution of scholarly communication.* Learned Publishing, 2002. **15**(1): p. 7-19.
110. Larivière, V., É. Archambault, and Y. Gingras, *Long-term variations in the aging of scientific literature: From exponential growth to steady-state science (1900–2004).* Journal of the American Society for Information Science and technology, 2008. **59**(2): p. 288-296.
111. Verstak, A., et al., *On the Shoulders of Giants: The Growing Impact of Older Articles.* arXiv preprint arXiv:1411.0275, 2014.
112. Evans, J.A., *Electronic publication and the narrowing of science and scholarship.* science, 2008. **321**(5887): p. 395-399.
113. Burnham, J.C., *The evolution of editorial peer review.* Jama, 1990. **263**(10): p. 1323-1329.
114. Spier, R., *The history of the peer-review process.* TRENDS in Biotechnology, 2002. **20**(8): p. 357-358.
115. Burrell, Q.L., *Modelling citation age data: Simple graphical methods from reliability theory.* Scientometrics, 2002. **55**(2): p. 273-285.
116. Glänzel, W., *Towards a model for diachronous and synchronous citation analyses.* Scientometrics, 2004. **60**(3): p. 511-522.
117. Nakamoto, H., *Synchronous and diachronous citation distribution.* 1988.
118. Pan, R.K., et al., *The memory of science: Inflation, myopia, and the knowledge network.* arXiv preprint arXiv:1607.05606, 2016.
119. Parolo, P.D.B., et al., *Attention decay in science.* Journal of Informetrics, 2015. **9**(4): p. 734-745.
120. Kuhn, T.S., *The structure of scientific revolutions*. 1962: University of Chicago press.
121. Van Raan, A.F., *Sleeping beauties in science.* Scientometrics, 2004. **59**(3): p. 467-472.
122. Ke, Q., et al., *Defining and identifying Sleeping Beauties in science.* Proceedings of the National Academy of Sciences, 2015: p. 201424329.

123. He, Z., Z. Lei, and D. Wang, *Modeling citation dynamics of "atypical" articles.* Journal of the Association for Information Science and Technology, 2018. **69**(9): p. 1148-1160.
124. Garfield, E., *Citation Indexes for Science.* Science, 1955. **122**(3159): p. 108-111.
125. Erdős, P. and A. Rényi, *On the evolution of random graphs.* Publ. Math. Inst. Hung. Acad. Sci, 1960. **5**(1): p. 17-60.
126. Evans, J.A., *Future Science.* Science, 2013. **1245218**(44): p. 342.