Knowledge: gift or burden of innovation?

Siyuan Lyu

Department of Economics
Stony Brook University

January 11, 2025

Abstract

The accumulation of knowledge is crucial for the development of scientific innovation. However, in recent years, despite a significant increase in the volume of research in both academia and industry, there has been a noticeable decline in breakthrough innovations. Some studies suggest that the accumulation of knowledge leads to increased learning costs, prolonging the innovation process, promoting teamwork, and narrowing individual research fields. In contrast, other studies argue that the innovation process essentially involves the recombination of existing knowledge, akin to hybridization in biology, thereby suggesting that knowledge accumulation can accelerate innovation. This paper examines the role of knowledge accumulation in the innovation process by distinguishing between the accumulation of general field knowledge and the specific knowledge directly referenced in individual innovations. Through a theoretical model, following Arora et al., 2021, this paper discusses how knowledge accumulation can reduce innovation costs, while directly referencing existing knowledge increases the difficulty of innovation. Knowledge accumulation encourages citations, which exhibit a substitute effect with citations in the innovation cost equation, thereby mitigating the negative impact of citations on innovation. Additionally, by utilizing data compiled by Park et al., 2023 and NBER Patent Project database, this paper provides empirical evidence showing that field-level knowledge reserves promote both innovation and citations, whereas direct citations constrain innovation. Thus, the observed decline in innovation is more attributable to citation behavior rather than to the accumulation of knowledge.

1 Introduction

The accumulation of knowledge is a driving force behind scientific innovation and technological progress. In both academic and industrial sectors, the volume of research output has grown exponentially in recent years. Despite this expansion, the rate of breakthrough innovations appears to have declined, raising concerns about the efficiency of modern innovation processes. Several theories attempt to explain this paradox. One view suggests that the continuous accumulation of knowledge increases the complexity of learning, lengthens the innovation process, and encourages specialization in narrower research fields, often requiring larger collaborative teams.

On the other hand, an alternative perspective argues that innovation fundamentally involves the recombination of existing knowledge, much like biological hybridization. In this view, accumulated knowledge serves as a vast repository from which new, original ideas can emerge through novel combinations. This suggests that, contrary to concerns of stagnation, knowledge accumulation could actually accelerate the innovation process by providing more diverse elements for recombination.

This paper seeks to resolve these competing views by differentiating between two forms of knowledge: general field knowledge and the specific knowledge directly cited in innovations. We build on the theoretical framework proposed by Arora et al., 2021 to examine how the accumulation of broad, field-level knowledge can lower innovation costs, while citing specific, directly related knowledge may increase the difficulty of generating new ideas. By integrating data from the NBER Patent Project and recent research by Park et al., 2023, we provide empirical evidence that highlights the complex relationship between knowledge accumulation, citations, and innovation outcomes.

Our findings suggest that while field-level knowledge accumulation promotes both innovation and the citation of prior work, the act of directly referencing specific knowledge appears to constrain innovation. In particular, we argue that the observed decline in breakthrough innovations is driven more by changes in citation behavior than by the sheer accumulation of knowledge itself. This distinction offers a nuanced understanding of how modern knowledge ecosystems impact innovation and presents opportunities for optimizing the balance between knowledge accumulation and citation practices in fostering future breakthroughs.

The paper proposes a model that distinguishes between general field knowledge and citations in a research project (patent development). In this model:

- Researchers pursue citations by selecting optimal references and innovation.
- Knowledge reduces the marginal cost of innovation, while references increase it. Knowledge and references act as complements in influencing the marginal cost of innovation.
- Innovation decreases with an increase in references but increases with an increase in knowledge.
- References also increase with an increase in knowledge.
- In equilibrium, the negative effect from references outweighs the positive effect from knowledge, resulting in an overall decrease in innovation as knowledge accumulates.

The logic of the paper is explained in Figure 1.

[Place Figure 1 here.]

The paper also conducts several empirical regressions using the dataset. The results indicate that knowledge has a significantly positive effect on innovation, while references have a significantly negative effect on innovation. The empirical findings present three key results. First, the overall impact (or payoff) of a patent increases as both its level of innovation and the number of references it cites grow. This indicates that patents that are both innovative and deeply embedded in existing knowledge tend to have higher commercial or scientific value (Hall and Ziedonis, 2001; Lucas Jr, 2009; Akcigit et al., 2018; Akcigit and Kerr, 2018). Second, while innovation tends to decrease with the number of references cited, it increases with the accumulation of general public knowledge in the field. This suggests that patents benefit from a rich knowledge base in their field (Acemoglu et al., 2016) but excessive reliance on direct citations can constrain their innovative potential. Finally, we find that references themselves tend to increase as public knowledge accumulates, highlighting the complementary effect of growing knowledge reserves on the use of prior work in new inventions.

Taken together, these results suggest that the observed decline in breakthrough innovations is driven more by changes in citation behavior than by the sheer accumulation of knowledge itself. This distinction offers a nuanced understanding of how modern knowledge ecosystems impact innovation and presents opportunities for optimizing the balance between knowledge accumulation and citation practices in fostering future breakthroughs.

Related work. The evolution of innovation has been discussed in many papers as an important resource of economic development. As a significant input factor in production, innovation evolves as more research effort is invested, and can bring in multiple-period benefits consistently. Scientific research, however, is somewhere different from innovation, which is always measured as patents or other technology that can lead to a direct increase in economic profit. In a strand of theoretical models, innovation is regarded to be able to create a certain amount of value (i.e., innovation has a fair price) and the development of innovation is always improving itself, either by learning from others (Jovanovic and Nyarko, 1994) or by copying others' ideas (Jovanovic and Wang, 2020).

On the other hand, research cannot guarantee a positive output, and the direction of scientific development is heterogeneous. Also, the total value of the innovation is treasured the most in the production process rather than whether it is controversial and could bring revolutionary additions to the current technology. In fact, as long as it could bring incoming profits by affecting the production of the goods, the project has a value measured by the potential profits from it. For example, a minor change to a machine cannot be defined as a "super" invention, but it could increase the efficiency of the machine as well as future profits, which makes it quite valuable enough. As inUzzi et al., 2013, the highest-impact papers were not the ones that had the greatest novelty but had a combination of novelty and otherwise conventional combinations of prior work. While Wang et al., 2017 also shows that Novel research has a higher variance in its citations, reflecting its risky nature.

The paper also coordinates many theoretical and empirical papers that focus on the definition of innovation. Aside from Park et al., 2023, Uzzi et al., 2013 defines novelty as an unusual combination of pairs of cited papers; Wang et al., 2017 defines novelty as the recombination of pre-existing knowledge components in an unprecedented fashion; Kelly et al., 2021 uses the text similarity between patents to determine whether a patent is novel. In the theoretical model learning on the process of knowledge evolution, some economists also argue that the production of new ideas is made a function of the combination of all the old ideas from all the fields (Weitzman, 1998; Buera and Oberfield, 2020). In our paper, the addition to the current knowledge is determined by two parts, how much the project relies on the original idea (or the novelty level), and the value of the idea itself. Even if the researcher is quite confident about her own idea, it might not create so much value if the idea is not talented enough, which hinders the researcher with average intelligence from making full use of her plain idea.

The last strand of literature that this paper is trying to consonate is the key factors that affect the payoff of scientists or the impact of their scientific findings. As mentioned before, the impact of a finding not only relies on its innovation but also on its references. The highest-impact science is primarily grounded in exceptionally conventional combinations of prior work yet simultaneously features an intrusion of unusual combinations, and team size has a continually increasing relation with the likelihood of a high-impact paper (Verstak et al., 2014). Works that cite literature with a low mean age and high age variance are in a citation "hotspot", and Papers with authors in more locations and with longer reference lists get published in higher-impact journals and receive more citations(Uzzi et al., 2013; Lee et al., 2015; Mukherjee et al., 2017; Freeman and Huang, 2015). This paper considers that both the innovation investment and references add to the payoff of a scientist, and finds that as knowledge accumulates, more reliance on references "squeezes out" innovation investment, which is the main reason for declining disruptiveness.

The remainder of the paper is as follows: Section 2 provides an overlook of the development patterns in scientific research. Section 3 introduces the model and main results as well as their empirical implications; Section 4 presents the data and empirical strategies of the paper, whose results that support the findings in the theoretical model are shown in Section 5; Section 6 concludes the paper.

2 Science of Science

Becoming a more and more popular field, the Science of science has attracted the eyesights of physicists, sociologists, and of course, economists. It captures the development of each scientific field, using empirical and theoretical tools to track the growing pattern of research. It also discusses the possible approach to encourage new findings and to implement novel ideas into production in society. After taking tons of publication and patent data from different sources into account, there are several pieces of empirical evidence to support the following properties of scientific research.

Increase in workload and team size. The 20th century has witnessed an exponential increase in scientific research, represented by both the number of articles published in scientific journals and that of researchers around the world (Jones, 2011; Fortunato et al., 2018; Azoulay et al., 2018). As knowledge of each field keeps expanding, the overlapping of multiple fields makes it possible for interdisciplinary studies. Also, as the fruits of the lower branch have been thoroughly collected, the cost of making new progress is also increasing. Hence, it is quite natural that teamwork is becoming more and more frequent, with collaborations between researchers from different fields and identities (Jones et al., 2008). However, the unique ideas don't keep the same pace with the rapid development of scientific results, on the contrary, the novel findings that each researcher or team generates are decreasing. This will be discussed in detail in the following paragraph.

General decline in research novelty. Park et al., 2023 develop a new index called CD (consolidating/disruptive) index to evaluate the "disruptiveness" of a paper or a patent. The intuition is that if a paper or patent is disruptive, the subsequent work that cites it is less likely to also cite its predecessors. If a paper or patent is consolidating, subsequent work that cites it is also more likely to cite its predecessors. CD index of a paper or patent is calculated by the ratio between the numbers of its subsequent work that only cites it and that cites both itself and its citation, ranging from $[-1, 1]$, where -1 represents the paper or patent is completely consolidating and 1 denotes the work is an absolute radical one. After collecting 25 million papers in the Web of Science between 1945 and 2010, and 3.9 million patents in the United States between 1976 and 2010, they found that there is a consistent decline in the disruptiveness among different scientific fields (Life science, biomedicine, social science, physical science, and technology) in the past more than 60 years. Also, Shi et al., 2015 also find that the past three decades witness a flat increase in new things and methods in the field of biomedicine compared with the 1960s when there was a surge in scientific activities.

Reliance on "stars". One of the most famous sayings by scientists is Issac Newton's "If I have seen further it is by standing on the shoulders of Giants.", which reveals an important factor in scientific findings: reliance on previous work by other scientists. Especially, the productivity of scientists whose work relied on frontier research is significantly affected by the accessibility of these works (Iaria et al., 2018). Based on this, we can conclude that the scientists who are productive at frontier research, i.e., the "star" scientists, have a larger impact in their fields than the others (Jones et al., 2008). Azoulay et al., 2019 find that the death of star scientists usually leads to changes in their subfields, with more "outsiders" surge in and change the direction of the fields. This also supports that the choices of scientific ideas and methods are quite narrow, surrounding what the stars are thinking and doing. The clustering in research also indicates a possible reason that the ideas are getting more and more similar, that the entrant researchers are always following the dominant ones without daring to develop new ideas.

The role of existing knowledge in the development of innovation has been subject to two main discussions within the academic literature. First, the concept of the "Burden of knowledge" (Jones, 2009) suggests that the accumulation of knowledge stock may impede innovation. According to this view, the increasing knowledge stock extends the time required to obtain degrees, leading researchers to focus more narrowly on their fields of study and increasing collaboration. However, this intense focus and collaboration may inadvertently slow down the innovation process.

Conversely, the theory of "Recombinant growth" (Weitzman, 1998; Acemoglu et al., 2016) proposes that innovation often emerges from the combination of existing ideas. From this perspective, a greater knowledge stock facilitates innovation by providing a larger pool of ideas to draw from. This raises the question: what is the actual role of knowledge stock in research? Does it hinder or motivate innovation?

This research aims to explore the relationship between knowledge stock and innovation. By investigating the impact of knowledge accumulation on the pace and direction of innovation, this study seeks to provide insights into the mechanisms through which existing knowledge influences the innovation process.

Utilizing data from NBER Patent Project database and following the methodology outlined by Park et al., 2023, a dataset was constructed comprising patents published between 1976 and 2006. This dataset was used to create three key variables:

- **Innovation Index (I)**: This index measures the disruptiveness of innovation. It is defined as the fraction of patents citing the patent in question but not citing any of its references, divided by the total number of patents citing the patent. A value closer to 1 indicates a higher level of disruptiveness.
- **Knowledge (K)**: For each year and field, knowledge is defined as the total number of patents published in that field up to and including the specified year. All patents published within the same field in the same year are assigned the same knowledge value.

- **References (R)**: This variable represents the total number of patents cited by each patent.
- The annual mean values of Innovation, Knowledge, and References were calculated for each subfield. To ensure consistency in the scale, the value of Innovation was scaled up by a factor of 5.

These variables allow for the examination of the relationship between knowledge stock and innovation, as well as the impact of references on innovation within the dataset of patents published in the United States between 1976 and 2006.

[Place Figure 2 here.]

Figure 2 demonstrates that as knowledge accumulates, the number of references increases, while disruptiveness declines. Is this decline caused by both types of existing ideas, or by only one?

Figure 3 illustrates the distribution of patent innovativeness relative to the number of citations across six fields as categorized by NBER in the year of 1978. Subfigures (a) through (f) represent fields Chemical, Computers & Communications, Drugs & Medical, Electrical & Electronic, Mechanical, and Others, respectively. Within the same year and across each field, all patents are subjected to the same knowledge reserves within that field. In each subfigure, the vertical axis represents the average level of innovativeness, while the horizontal axis shows groups with an increasing number of citations from left to right. It is evident that as the number of citations increases, the average level of innovativeness steadily decreases. This trend is consistent across all six fields, indicating that, after controlling for the effect of knowledge accumulation, patents with a higher number of citations tend to be less innovative.

[Place Figure 3 here.]

Figure 4 shows all the patents that are divided into three groups based on their amount of references relative to the average. In each group, the subfigure shows both the group mean value of knowledge and innovation among the patents with similar reference amounts. It is consistent in each group that the fields with more knowledge stock are inclined to develop more innovative patents. By comparing innovativeness between groups, it decreases from Figure 4(a), 4(b) to 4(c), which adds to the negative relationship between references and innovation.

[Place Figure 4 here.]

The paper also investigates whether knowledge could accelerate patents' referring patterns in Figure 5, which shows both absolute and relative amounts of references with respect to knowledge over time.

[Place Figure 5 here.]

3 Conceptual Framework

The following will present a simple model of an innovation game where two researchers[1] are competing for citations/impacts by choosing the novelty scale. The main characteristics of the model are as follows:

- The payoff is related to citations or impacts of the research instead of its disruptiveness. As mentioned in Uzzi et al., 2013, the most impactful research is not the one that overturns the past findings, but the one deeply rooted in literature. There is no need to stress the importance of citations for the scientists, yet it also leads to higher values for the innovators in the firms (Poege et al., 2019). Hence the model assumes that the researcher cares both reliance on past knowledge and the new ideas.

- The role of knowledge. The model captures the value of knowledge on two channels: (i) Reliance on past knowledge (citations) could add to the payoff directly; (ii) Following Arora et al., 2021, knowledge could also help to cut the unit cost of innovation. The "burden" of knowledge, i.e., the cost of learning it is also included.

- The value and uncertainty brought by new ideas. The new ideas will add to the value of the whole project, while it is also risky to succeed (Foster et al., 2015 ;citewang2017bias;Carson et al., 2023;Agrawal et al., 2024). This requires that the payoff function is concave in innovation to capture its uncertainty. There also exists competition as well as a spillover effect between firms' innovation.

The model maintains the previous characteristics of a research game: (i) Innovation is risky, which is captured by a concave payoff function with respect to innovation; (ii) Value of knowledge, which cuts the cost of innovation. However, the model applies some new properties related to academia and industry: (i) The knowledge is separated into two types: private knowledge, which adds to the value of a research project and incurs an education cost (if private knowledge is obtained via collaboration, the cost contains personnel administration); and public knowledge, which helps to decrease the unit innovation cost; (ii) Role of the rival becomes more complicated, both add to the value of the research depending on the similarity of the two research and attracts attention depending on the diversity of the two researchers.

3.1 Settings

The setting of the model is as follows. There are two researchers developing their research strategy independently. Each research strategy is a bundle of (I, α) where I_i denotes researcher i's own novel idea and α_i is her reliance on private knowledge (the past knowledge the researcher learns from training). The private knowledge could not surpass the current public knowledge $\alpha_i \in [0, k]$. Each researcher's revenue depends on both the choice of innovation and private knowledge. There is a

[1] The "researcher" here not only refers to the individual scientist in the academia but also includes research teams in all kinds of research institutions.

technical spillover between the two researchers' innovations, which implies imitation from the other.

The game has two stages. In the first stage, both agents choose (α_i, α_j) as how many references they are going to learn; in the second stage, agents choose (I_i, I_j) as how much innovation they will input into the project, and then the payoffs realize.

In many cases, the role of innovation in firms' revenue is described as to cut the unit production cost (d'Aspremont and Jacquemin, 1988; López and Vives, 2019; Antón et al., 2024) and henceforth contributes to increasing the profits. However, this setting lies an over-simplified assumptions on innovation, which describes the unit cost as a linear function of innovation. This assumption ignores the risk of innovation and assumes that as long as there is R&D investment, the profit will benefit from it proportionally. However, more innovation inputs in a project could lead to a large uncertainty, as it takes longer time to finish a project, or making the project less likely acceptable for others (Hill and Stein, 2021; Carson et al., 2023). Besides, the previous models tend to focus on the competition of quantity, whose marginal cost is determined by innovation, and it is different to the focus of this paper. To capture the risk and the value of innovation, our model assumes a revenue function concave in innovation, which is similar to Arora et al., 2021. The difference is our model also indicates that references not only cuts the innovation cost, but also contributes to the revenue function directly. The revenue of the project is symmetric and determined by both agents' choice:

$$\Pi_(I_i, I_j, \alpha_i) = I_i - \frac{c_1}{2} I_i^2 + c_2 I_i I_j + b\alpha_i$$

where $c_1 > 0$, and $b > 0$.

Here both innovation and references add to the impact of the research, and there exists an incomplete substitution between the two items, which is measured by $b > 0$. The revenue is concave in its first argument, that the revenue will increase in the innovativeness of the project, while the marginal return of innovation is changing in a diminishing rate, collaborating with the fact that innovation in a project is risky regarding of its impact. c_2 captures the nature of the strategic interactions, including the scale of the spillover effect as well as the competition between the innovations. Here we assume c_2 is symmetric between two researchers. If c_2 is positive, innovations are strategic complements and the researcher's revenue increases in both own innovation and the rival's innovation; and if c_2 is negative, innovations are strategic substitutes and rival's innovation would dampen the researcher's revenue.

The researcher i pays two sorts of costs: (i) Innovation cost. This is the cost of implementing and developing her novel idea, and the marginal cost of innovation is $\phi_i(k; \alpha_i) I_i$ with $\frac{\partial \phi_i}{\partial k} < 0$, $\frac{\partial \phi_i}{\partial \alpha_i} > 0$, and $\frac{\partial^2 \phi_i}{\partial \alpha_i \partial k} < 0$. Similar to Arora et al., 2021, this means that reliance on past knowledge could help to cut the innovation cost, while references add the innovation cost. The last inequality also requires that public knowledge act as an mitigation of the marginal effect of references on the innovation cost. (ii) Education cost. This is the cost of learning the knowledge and is represented by $C(\alpha_i) = \gamma\alpha_i$, an increasing convex function. We assume there exists a stable Nash Equilibrium. This requires that $D = c_1^2 - c_2^2 > 0$, or $|c_1| > |c_2|$.

3.2 Equilibrium

Given the above settings, agent i is maximizing the following payoff:

$$v_i = I_i - \frac{c_1}{2} I_i^2 + c_2 I_i I_j + b\alpha_i - \phi_i(\alpha_i; k) I_i - \gamma\alpha_i$$

The relationship between two researchers' innovation depends on the sign of c_2, the strategic interactions. If $c_2 > 0$, i.e., innovations are strategic complements, then both researchers would benefit from the spillover of rivals' innovation, and if one researcher increases its innovation input, the other will also innovate more. If $c_2 < 0$, i.e., innovations are strategic substitutes, then the force of competition for attention would surpass that of technical spillover, and an increase in one researcher's innovation would decrease that of its rival.

The choice of reliance on past knowledge is related to its own contribution, education cost, and the scale of innovation. If ϕ_i is convex in α_i, then an increase in references would decrease innovation. On the one hand, it increases the cost of innovation in an increasing rate, makes it less profitable to innovate; on the other hand, it contributes more to the payoff of the project, which also "squeezes out" innovation.

In the second stage, the agents' optimal choice of innovation is given as

$$I_i = \frac{c_1(1 - \phi_i(\alpha_i; k)) + c_2(1 - \phi_j(\alpha_j; k))}{D}$$

$$I_j = \frac{c_1(1 - \phi_j(\alpha_j; k)) + c_2(1 - \phi_i(\alpha_i; k))}{D}$$

To assure the existence of nonnegative results, it requires $\phi_i, \phi_j \leq 1$. Both innovations rely on both references and public knowledge.

3.3 Empirical Implications

3.3.1 What influences innovation?

Here we provide the main results and the intuition. The returns to innovation depend on its scale, and the return to references depends on both its value and its negative effect on innovation. In equilibrium, if both agents' innovations are strategic substitutes, and an increase in innovation by one firm reduces innovation by the other.

From the equilibrium there is

$$\frac{\partial I_i^*}{\partial \alpha_i} = -\frac{c_1}{D}\frac{\partial \phi_i}{\partial \alpha_i} < 0$$

That at the equilibrium, innovation decreases in own references, since the reliance on knowledge could hinder the progress of developing new ideas and add to innovation cost.

The relationship between innovation and the rival's references is depending on

$$\frac{\partial I_j^*}{\partial \alpha_i} = -\frac{c_2}{D}\frac{\partial \phi_i}{\partial \alpha_i}$$

via the innovation cost of the rival. Note that if $c_2 > 0$, $\frac{\partial I_j}{\partial \alpha_i} < 0$, i.e., researcher j also decreases its innovation in response to the decrease in research by researcher i. This is because that innovations are strategic complements.

The response of innovation input to public science is

$$\frac{\partial I_i}{\partial k} = -\frac{1}{D}(c_1 \frac{\partial \phi_i}{\partial k} + c_2 \frac{\partial \phi_j}{\partial k})$$

$$\frac{\partial I_j}{\partial k} = -\frac{1}{D}(c_1 \frac{\partial \phi_j}{\partial k} + c_2 \frac{\partial \phi_i}{\partial k})$$

If there exists strategic complementarity between innovations, i.e., $c_1 > 0$, both firms innovate more in response to an increase in public science. However, if the innovations are strategic substitutes, then one (but not both) firm may reduce innovation. In particular, if the innovation costs of a firm are not very responsive to public science, the effect of a rival increasing its innovation may cause the firm to reduce its innovation. However, note that

$$\frac{\partial I_i}{\partial k} + \frac{\partial I_j}{\partial k} = -\frac{1}{D}(c_1 + c_2)(\frac{\partial \phi_i}{\partial k} + \frac{\partial \phi_j}{\partial k}) > 0$$

This implies that innovation on average increases with public science.
The above discussions yield the following findings on the role of public knowledge and references on innovation.

Proposition 1. *In the equilibrium, the optimal choice for innovation is determined by the firm's innovation cost and its competition with the rival. And there is* $\frac{\partial I_i^*}{\partial \alpha_i^*} < 0, \frac{\partial I_j^*}{\partial \alpha_i^*} < 0, \frac{\partial I_i^*}{\partial k} > 0$ *and* $\frac{\partial I_j^*}{\partial k} > 0$ *under the assumptions on innovation cost functions and strategic complementarity, i.e., innovation decreases in both own and rival's references and increases in public knowledge.*

Both of the propositions could be derived and proved from the first-order conditions above.

The returns to reliance on past knowledge and public knowledge depend on the scale of innovation because the former increases the unit cost of innovation while the latter reduces the cost. In equilibrium, researcher i refers to more existing knowledge, and thus will have a higher marginal return from references yet a lower marginal return from innovation. If innovations of two researchers are strategic substitutes, an increase in references by one researcher reduces innovation by the other. On the other hand, the shared public knowledge could always encourage the generation of new innovations.

3.3.2 What influences references?

At an interior maximum, the direction of the effect of public science on references is given by $\frac{\partial^2 v_i}{\partial \alpha_i \partial k}$. When $\frac{\partial^2 \phi_i}{\partial \alpha_i \partial k} < 0$ and $|\frac{\partial^2 I_i}{\partial \alpha_i \partial k}|$ is large enough,

$$\frac{\partial^2 v_i}{\partial \alpha_i \partial k} = -\frac{\partial I_i}{\partial k}\frac{\partial \phi_i}{\partial \alpha_i} - I_i \frac{\partial^2 \phi_i}{\partial \alpha_i \partial k} > 0$$

That is to say, the marginal return to references decrease in public knowledge. This is illustrated in the above equation. The first term indicates that public knowledge increases innovation, which imperfectly substitute references; the second term's sign depends on the sign of $\frac{\partial^2 \phi_i}{\partial \alpha_i \partial k}$, and if it is negative, then public knowledge could mitigate the negative effect from references on innovation, which adds to the marginal contribution of references to payoff.

In the first stage, Firm i's optimal choice for references satisfies

$$\frac{\partial \Pi_i}{\partial I_j}\frac{\partial I_j^*}{\partial \alpha_i} - \frac{\partial \phi_i}{\partial \alpha_i} I_i^* + b - \gamma = 0$$

Take the derivative with respect to k and then reorganize:

$$\frac{d\alpha_i}{dk} = -\frac{\frac{\partial^2 v_i}{\partial \alpha_i \partial k}}{\frac{\partial^2 v_i}{\partial \alpha_i^2}} > 0$$

When $\frac{\partial^2 \phi_i}{\partial \alpha_i \partial k} < 0$ and ϕ_i is convex in α_i, as k increases, α_i also increases; i.e., references increase in knowledge when references and knowledge are substitutes. This leads to the second result of the model.

Proposition 2. *In the equilibrium, the optimal choice for references is determined by the firm's optimal innovation, education cost, and the firm's own innovation cost function. And there is* $\frac{\partial \alpha_i^*}{\partial k} > 0$ *under the assumptions on innovation cost functions, i.e., references increase in public knowledge.*

Empirically, as k increases, the researcher's marginal returns to references depends on the supply of public science: the supply of public science will enhance the effect of references on innovation if it decreases the marginal cost of innovation. Conversely, if public science and references are substitutes, then public science could mitigate the effect of references on innovation. Under the assumption of substitutability, knowledge makes the marginal contribution of references on innovation cost smaller and more profitable to increase references, henceforth decrease innovation as well as increase optimal references. Therefore, the result in 5 indicates that references are increasing in public knowledge.

3.3.3 Role of public knowledge

The substitutability between references and public knowledge also leads to

$$\frac{\partial^2 I_i}{\partial \alpha_i \partial k} = -\frac{c_1}{D}\frac{\partial^2 \phi_i}{\partial \alpha_i \partial k} < 0$$

that accumulation of knowledge accelerates the negative effect of references on innovation.
And if $|\frac{\partial^2 I_i}{\partial \alpha_i \partial k}|$ is large enough, there is

$$\frac{\partial v_i^2}{\partial \alpha_i \partial k} = -\frac{\partial^2 \phi_i}{\partial \alpha_i \partial k} I_i - \frac{\partial \phi_i}{\partial \alpha_i}\frac{dI_i}{dk} > 0$$

that the accumulation of knowledge encourages references.

Proposition 3. *Knowledge amplifies the negative impact of references on innovation, and if* $|\frac{\partial^2 I_i}{\partial \alpha_i \partial k}|$ *is large enough, then there must be*

$$\frac{dI_i}{dk} = [\frac{\partial I_i}{\partial k}(>0) + \frac{\partial I_i}{\partial \alpha_i}(<0)\frac{d\alpha_i}{dk}(>0)] < 0$$

That the overall effect of knowledge on innovation is negative.

The model presented in this study offers a nuanced narrative that explores how researchers engage with two distinct types of existing ideas: references and public, general knowledge. In the process of innovation, these two types of knowledge play divergent roles. On the one hand, the accumulation of general knowledge serves as a powerful catalyst for innovation by reducing the costs associated with both direct innovation and innovation that builds upon references. General knowledge, with its broad applicability and foundational nature, can provide researchers with the tools and insights necessary to push the boundaries of their fields.

On the other hand, references, which represent specific, cited works directly related to the research at hand, can act as a double-edged sword. While they provide a necessary framework and context for ongoing research, they also introduce complexity and can create barriers to truly disruptive innovation. The reliance on references may make it more challenging for researchers to generate novel ideas, as they may be constrained by the existing body of work they must navigate and acknowledge.

The relationship between general knowledge and references is complex and not entirely straightforward. While the accumulation of general knowledge can facilitate both the process of innovation and the act of referencing existing work, the overall impact of this accumulation on innovation is, paradoxically, negative. This counterintuitive outcome arises because the negative influence of references—specifically, the way they can stifle the emergence of groundbreaking ideas—tends to outweigh the positive effects of general knowledge. As a result, despite the advantages that accumulated knowledge might offer, the dominant effect of increased references leads to a decline in the rate and impact of innovation.

This model thus underscores the delicate balance that exists between drawing on past knowledge and pushing the frontier of innovation, highlighting the potential pitfalls of over-reliance on existing ideas.

4 Data and Empirical empirical strategies

4.1 Data

The paper constructs an unbalanced panel of patents combines the measure of innovation from Park et al., 2023, from United States Patent Office (USPTO) data, and the counts of their impacts and references, as well as their originality and generality from NBER Patent Data Project (PDP) respectively. The combined dataset covers the period 1977-2006. We begin with 308,796 utility patents (excluding Non-utility patents such as Design, Plants, and other Non-utility patents, as classified in File, 2001) and are successfully filed between 1977 and 2006.

In this section we investigate the empirical relationship between public knowledge within each NBER classified field and subfield, amount of references, and innovation. Specifically, we are interested in how innovation and references inputs in a patent depend on the extent to which the knowledge is accumulated within in its corresponding field and on the extent to which the innovation spills over to neighboring patents in the same technology field. As in our theoretical framework we study the general implications of general knowledge and do not restrict ourselves to the study of any particular single industry.

4.2 Variables

4.2.1 Measure of Innovation

Novelty in research is calculated by a measurement method introduced by Park et al., 2023, that treats the novelty of a patent as to what extent it is creating new things that no previous patents have done and later patents would learn from. This measure of novelty is based on the premise that impactful and groundbreaking work often arises from the synthesis of previously unrelated ideas or knowledge domains. By identifying these novel combinations, the authors offer a systematic approach to assess how much a particular work departs from conventional paths.

The measure is calculated using citation data, which reflect the intellectual heritage of a paper or patent by showing which prior works are being referenced. More specifically, novelty is defined by looking at how citations in a particular paper or patent combine knowledge from different scientific or technological areas that have not been frequently linked before. In this way, the measure of novelty does not merely consider the content of the work itself but examines its relational structure to existing knowledge. The idea is that if the work draws from areas that are rarely cited together, it demonstrates a higher degree of novelty.

To quantify novelty, the authors leverage citation networks, where nodes represent papers or patents, and edges represent citation links between them. Novelty is calculated by analyzing how often certain citation combinations have appeared historically. The more unique the combination of cited references, the higher the novelty score for that paper or patent. In essence, novelty is inversely related to how common a particular combination of citations is. A paper citing two works from entirely different fields that are seldom combined would score higher in novelty compared to a paper that cites works frequently paired together.

The recombination concept is crucial because groundbreaking innovations often result from connecting distant areas of knowledge. For instance, combining advances in biology with techniques from artificial intelligence might result in novel insights that wouldn't emerge if a researcher simply built on well-established work within just one domain. This broader scope of knowledge integration is at the heart of the measure of novelty.

The measure of novelty is valuable in various contexts, particularly for understanding how breakthroughs in research and technology come about. By providing a quantitative measure of novelty, this approach allows researchers, policymakers, and funding agencies to identify highly innovative work, which might otherwise be overlooked. It could also be used to inform research funding decisions or to design programs that incentivize interdisciplinary collaborations, which are often associated with high novelty.

However, since novelty as defined before relies heavily on citation patterns, which may not capture all forms of innovation. For example, some groundbreaking work may not immediately be recognized as novel in citation networks but may still have significant long-term impacts. Additionally, the reliance on citation data assumes that citations accurately reflect intellectual connections, which is not always the case, particularly when strategic or social factors influence citation behavior.

In conclusion, the measure of novelty offers a robust way to understand how papers and patents introduce new ideas by combining previously unconnected fields. It allows for the identification of works that may push the boundaries of existing knowledge, making it a valuable tool for assessing the landscape of innovation.

4.2.2 Measure of knowledge and references

Knowledge. The variable knowledge ($Knowledge_{st}$) is designed to capture the cumulative stock of knowledge within a specific technological field over time. For each year and field, knowledge is defined as the total number of patents published in that field up to and including the specified year. This measure accounts for the accumulation of intellectual property and the growing pool of innovations that subsequent patents in the field may draw upon.

The calculation of $Knowledge_{st}$ begins by identifying all patents published in a given field up to the target year. Each patent is assigned to its respective technological field based on predefined categories, and then, for each year, the total number of patents in that field is summed. The cumulative total for each field represents the body of knowledge available at that point in time. For example, if a particular field had 100 patents published up until 1999, and an additional 20 patents were published in the year 2000, the knowledge value for that field in 2000 would be 120.

Importantly, all patents published in the same field and within the same year share the same knowledge value. This approach standardizes the level of accumulated knowledge for all inventions within a given field at a specific point in time. This allows for a comparative analysis of patents within the same field and year, as they are subject to the same level of cumulative prior knowledge.

The knowledge variable is crucial for understanding the innovation landscape because it provides a measure of the existing intellectual foundation on which new inventions build. In fields with a higher knowledge value, inventors have more prior work to reference, potentially leading to more incremental advances. Conversely, fields with a lower knowledge value may offer more opportunities for novel or disruptive innovations, as there is less existing knowledge to build upon or constrain new ideas.

This variable can be used to examine how the accumulation of knowledge in a field influences patenting behavior and the types of innovations that emerge. For instance, fields with rapidly growing knowledge values might see more incremental innovations, while fields with lower knowledge values could witness more breakthroughs as inventors explore less-charted territory. Additionally, knowledge could serve as a control variable in econometric analyses to account for differences in innovation potential across fields and over time.

The use of knowledge as a cumulative measure allows researchers to explore how the quantity of prior patents in a field affects new patenting activity. This variable can help determine whether the growth in the stock of prior patents stimulates or inhibits further innovation. For example, it may be used to test whether fields with a larger existing body of knowledge experience a higher rate of patenting or whether the saturation of knowledge in a field stifles the potential for disruptive innovations.

References. The variable references ($References_{ist}$) represents the total number of prior patents cited by a given patent. This variable is designed to quantify how much existing knowledge a new invention draws upon. Each time a patent is filed, it includes citations to previous patents that are relevant to the invention, either as foundational knowledge or to demonstrate the novelty and scope of the new invention. The references variable captures the extent to which each patent connects to and builds upon prior innovations.

For each patent in the dataset, the number of cited patents is counted, and this value is assigned as the references variable. The number of references can vary significantly across patents, depending on the complexity of the technology, the field of invention, and the patent's relationship to prior work. Patents that are highly incremental in nature may cite many earlier patents, reflecting their reliance on established technologies. In contrast, patents that are more groundbreaking may cite fewer prior patents if they introduce fundamentally new concepts or approaches.

The references variable is constructed at the level of individual patents, meaning that each patent in the dataset is assigned its own unique references value based on the total number of citations it includes. This variable serves as an indicator of the breadth of prior knowledge a patent draws from, and it can be used to assess the degree of connectivity between the new patent and the existing body of intellectual property.

The references variable is key to understanding the relationship between a patent and the existing technological landscape. It provides insight into how much a new invention relies on established ideas, allowing researchers to assess the degree of originality or novelty present in the patent. Patents with a high number of references may suggest that the invention builds upon a large body of prior work, whereas patents with fewer references might indicate that the invention is more novel or independent of existing technologies.

4.2.3 Other variables

Originality. Originality measures the breadth of different technological fields cited by a patent. A higher Originality score suggests that the patent draws from a diverse range of sources, integrating knowledge from multiple areas to create something new. This variable is essential for assessing how innovative or interdisciplinary a patent is, as broader citations imply a greater potential for novel insights and cross-field breakthroughs.

Age. The Age variable represents the mean age of the patents or prior work cited by the current patent. This measure provides insight into whether the new patent builds on recent advancements or relies on older foundational work. A lower Age value may suggest that the patent is highly current and responsive to the latest developments, while a higher Age value might indicate that the invention leans on more established, older knowledge.

Age Variance. Age Variance captures the dispersion in the age of work cited by a patent, reflecting the diversity of time periods from which the cited patents originate. A high Age Variance means that the patent draws upon a broad range of knowledge across different time periods, while a low variance suggests that the patent is more narrowly focused on work from a specific time window. This variable provides insight into the temporal breadth of the knowledge base that the patent is building upon.

Team Production. Team Production measures the mean number of prior works produced by the team members involved in a patent's creation, expressed in logarithmic form. This variable assesses the experience and productivity of the inventors, where a higher value indicates that the team has a strong track record of prior contributions. It reflects the collective expertise brought to bear on the new invention and may correlate with the quality or impact of the resulting patent.

Diversity. The Diversity variable represents the diversity of the work cited by the patent, typically measured by the variety of different subfields from which the patent draws its references. A high Diversity score suggests that the patent is leveraging knowledge from a wide range of technological areas, indicating a potentially interdisciplinary or novel approach. This measure is useful for understanding how broad or specialized the patent's knowledge base is.

Competition. Competition measures the average innovation in the same year and subfield, indicating the level of competitive pressure within that specific technological space. A higher value means that the subfield is highly active, with numerous innovations occurring simultaneously. This variable helps to contextualize the innovation environment for each patent, as intense competition can either spur rapid advancements or make it more challenging for any single patent to stand out.

Table 1 provides an overview of the descriptions and resources of key variables.

[Place Table 1 here.]

Table 2 presents the summary statistics of key variables in the dataset. The CD-5 index, which measures the disruptiveness of patents, has a mean value of 0.13 with a standard deviation of 0.30, ranging from -1 to 1. This indicates that in general, the patents are creating breakthroughs slightly. On average, there are 10.62 (in logarithm) patents published before in the same subfield, while the logarithm number of patents published in the same year and subfield has a mean value of 8.29. On average, the patents cite 1.61 previous ones with a mean age of 3.08 years. Originality, measured by the variety in different fields of researcher's references, has a mean value of 0.52. The dispersion in the age of works cited has a mean value of 3.08 years. The mean number of prior works produced by team members is 1.35, while the diversity of works cited has a mean value of 0.98. Lastly, the average innovation in the same year and subfield has a mean value of 0.14.

[Place Table 2 here.]

4.3 Empirical strategies

To investigate the effects of general subfield knowledge and citations on innovation, the paper adopts a linear model as follows:

$$Innovation_{ist} = \beta_0 + \beta_1 Knowledge_{st} + \beta_2 References_{ist} + \mathbf{X}_{ist}\sigma + \epsilon_{st}$$

where $Innovation_{ist}$ indicates patent i published in subfield s by the year of t. This is measured mainly by CD-5 index, and CD-10 index is also used as a robustness check. $Knowledge_{st}$ and $References_{st}$ are the numbers of existing patents published before t within the subfield s, and of references cited by i respectively. Other variables include competition (patents published in the same year and same subfield), total backward citations from the patent, originality, mean age of work cited, dispersion in age of work cited, mean number of prior works produced by team members, diversity of work cited, and average innovation in the same year and subfield. Time and subfield fixed effects are controlled.

5 Empirical evidences

Using the specialized dataset generated in Section 4, the paper conducts some basic empirical regressions to support the discussions in Section 3.

5.1 What affects innovations?

Table 4 presents the core findings derived from the model discussed earlier. The dependent variables in this analysis are the CD-5 index and the CD-10 index, which serve as indicators of disruptive innovation in patents. A higher value in these indices signifies a greater degree of disruptiveness, indicating that the patent has made a significant impact by introducing innovations that deviate from the existing knowledge base and potentially pave the way for new directions in research and development.

One of the most notable results from the table is the effect of backward citations on innovation. The analysis reveals a significantly negative relationship between the number of backward cites from the patent and its disruptiveness. This suggests that when a patent heavily relies on existing knowledge, as evidenced by a higher number of backward citations, it is less likely to produce a truly disruptive innovation. This finding highlights the constraining effect of existing knowledge frameworks on the ability to generate novel and groundbreaking ideas.

Conversely, the knowledge accumulated within the same subfield emerges as a crucial factor that significantly enhances the patent's innovation potential. This positive contribution of knowledge is evident both within a 5-year and a 10-year timeframe, underscoring the importance of a well-established knowledge base in fostering innovative activities. The findings suggest that when researchers have access to a rich repository of accumulated knowledge, they are better equipped to push the boundaries of innovation, leading to more disruptive patents.

In addition, the table highlights the complex dynamics of competition in the innovation process. Specifically, the analysis reveals that competition based on quantity—where the focus is on producing more patents—tends to discourage innovation. This is likely because quantity-driven competition may lead to incremental innovations that prioritize volume over quality or novelty. On the other hand, competition that centers on the novelty of inventions appears to have a motivational effect, encouraging patent developers to innovate more aggressively and to pursue more disruptive ideas. This distinction between quantity and novelty in competitive environments underscores the importance of fostering a culture of innovation that values breakthrough ideas over mere output.

[Place Table 4 here.]

The paper extends its analysis by running the primary regression model with an additional focus on field-specific knowledge and competitive dynamics. The results of this analysis are presented in Table 5, which reflects a trend similar to that observed in Table 4. This consistency across different models lends robust support to the theoretical discussions outlined in Section 2. Specifically, the findings reinforce the notion that references, when over-relied upon, can stifle innovation by anchoring new research too closely to existing ideas, thereby reducing the likelihood of generating novel and disruptive innovations.

In contrast, the knowledge accumulated within the same subfield and broader field appears to be a significant asset in fostering novelty. This suggests that while references may exert a constraining effect, the broader base of knowledge serves as a foundation that can spur innovation, providing the necessary tools and insights to push the boundaries of current understanding. This dual role of knowledge—as both a facilitator of innovation and, when filtered through excessive referencing, a potential hindrance—highlights the complex interplay between different types of existing ideas in the innovation process.

The findings in Table 5, therefore, further substantiate the argument that knowledge within a specific field is more likely to be a "gift" to innovation, enhancing the potential for novelty and breakthrough discoveries, while the act of referencing too heavily can "squeeze out" the space needed for truly innovative ideas to emerge. This nuanced understanding of the relationship between knowledge and innovation is crucial for guiding both academic research and policy-making aimed at fostering a more innovative and dynamic research environment.

[Place Table 5 here.]

5.2 Competition between innovation and references

The paper then deeply investigates the possible mechanism that explains the positive effect of knowledge and the negative effect of references. The paper adopts a linear model as follows:

$$Citation_{ist} = \beta_0 + \beta_1 Innovation_{ist} + \beta_2 References_{ist} + \mathbf{X}_{ist}\sigma + \epsilon_{st}$$

where $Citation_{ist}$ counts the total forward cites to the patent and is used as a proxy for the impacts of the patent, which is the main payoff of its developers. $Innovation_{ist}$ and $References_{st}$ are the same variable as in the main regression. Other variables include competition (patents published in the same year and same subfield), total backward citations from the patent, originality, mean age of work cited, dispersion in age of work cited, mean number of prior works produced by team members, diversity of work cited, and average innovation in the same year and subfield. Time and subfield fixed effects are controlled.

Table 6 presents an analysis of the relationship between a patent's disruptiveness, its backward citations, and the overall impact, as measured by the citations it receives. The results suggest that both disruptiveness and backward citations significantly contribute to a patent's impact, but they function as substitutes rather than complements. This implies that while a patent's ability to disrupt existing technology plays a crucial role in its influence, the number of backward citations—indicating how much it builds on existing work—can offset or substitute the need for disruptiveness.

Additionally, the table highlights other factors that may influence the number of citations a patent receives. Among these, competition and the diversity of citations emerge as significant negative determinants. This suggests that in more competitive environments, or when citations are drawn from a more diverse array of prior work, the overall impact of a patent may be diminished. These findings align with the discussion in Proposition 1, reinforcing the notion of a potential negative correlation between innovation and references. Specifically, as the reliance on references increases, it may constrain the degree of innovation, thereby reducing the disruptiveness and ultimate impact of the patent. This nuanced relationship underscores the complex dynamics at play in the process of innovation, where the balance between building on existing knowledge and breaking new ground must be carefully managed to maximize a patent's influence.

[Place Table 6 here.]

5.3 What affects references?

Lastly, the paper aims to provide empirical evidence for the positive correlation between knowledge and references. To test this relationship, the paper uses $References_{st}$ as the dependent variable and $Knowledge_{st}$ as well as $Innovation_{ist}$ as the primary independent variable. This analysis is conducted while controlling for innovation and other characteristics of the patent to ensure that the observed effects are not confounded by these factors. The results, presented in Table 7, demonstrate that the relationship between knowledge and references remains robust and statistically significant. Specifically, the data show that as accumulated knowledge increases, the number of references a patent receives also rises. This finding supports Proposition 2, which posits that an increase in the general body of knowledge within a field leads to a greater number of citations for patents. Thus, the empirical evidence reinforces the notion that accumulated knowledge not only enhances the overall innovation landscape but also increases the extent to which new patents build upon existing research.

[Place Table 7 here.]

Integrating the empirical findings discussed earlier, the paper draws a compelling conclusion regarding the dual role of innovation and references in determining the overall impact of a patent, which is quantified by the total citations it garners. The analysis reveals that both innovation and references are significant contributors to a patent's influence within its field, highlighting a complex interplay that suggests a substitution effect between these two critical components of a research project. Essentially, while innovation drives the creation of novel ideas, references anchor these innovations within the existing body of knowledge, both playing distinct yet interconnected roles in shaping a patent's significance.

However, the study also identifies a crucial factor: the accumulation of knowledge within the same field or subfield has a profound effect on both innovation and references. This growing body of knowledge serves as a double-edged sword. On one hand, as knowledge accumulates, it acts as a catalyst for innovation, directly enhancing the capacity of researchers to generate novel and impactful ideas. This positive relationship underscores the importance of a robust knowledge base in fueling the creative processes that lead to scientific and technological advancements.

On the other hand, the paper uncovers a more nuanced and somewhat paradoxical dynamic. The accumulation of knowledge also exerts a negative influence on innovation through its impact on references. As knowledge grows, it inevitably leads to an increase in the number of references that researchers must navigate and incorporate into their work. This increased reliance on references introduces greater complexity and potentially stifles the creative aspects of innovation, as researchers become more constrained by existing frameworks and less able to break free from established paradigms. The empirical results indicate that this negative impact from references is not only significant but substantial enough to outweigh the positive effects derived from the accumulation of knowledge.

As a result, there is an observable and concerning trend: innovation in patents tends to decline as the overall body of knowledge in a field continues to expand. This finding is particularly striking, as it suggests that while knowledge is indispensable for innovation, its interaction with references can lead to diminishing returns, particularly in terms of the novelty and disruptiveness of new inventions. These insights are in alignment with the theoretical discussions outlined in Section 3 of the paper, reinforcing the complex and sometimes counterintuitive relationships between knowledge, references, and innovation within the broader context of scientific and technological progress.

6 Conclusion

Is knowledge a gift that fuels the novelty of scientific discovery, or is it a burden that impedes innovation? This paper seeks to unravel this complex question by introducing a novel approach that differentiates between general field knowledge and the specific references that researchers rely on in their projects. Through the construction and analysis of a theoretical model, the paper argues that knowledge, in its broad and general sense, serves as a significant asset. It acts as a gift to the innovation process, reducing the costs associated with both direct innovation and innovations that are derivative of existing references. In essence, general knowledge provides the foundational insights and tools necessary for researchers to advance their fields and push the boundaries of what is known.

However, the model also reveals that references, while necessary for situating new research within the existing body of work, can transform from a supportive resource into a constraining burden. As researchers increasingly rely on references, the process of innovation becomes more challenging.

The need to build upon and acknowledge previous work can create a form of intellectual inertia, making it harder to achieve breakthroughs and generate truly novel ideas. This duality in the role of knowledge—its capacity to both facilitate and hinder innovation—leads to an ambiguous relationship between general knowledge and references.

While the accumulation of knowledge can make both innovation and referencing more accessible, the overall impact of knowledge accumulation on innovation is found to be negative. This paradoxical outcome occurs because the negative effects of relying on references, which can stifle creativity and originality, tend to outweigh the positive effects of general knowledge. Consequently, as knowledge continues to accumulate, the innovation process is increasingly burdened by the weight of past references, leading to a decline in the novelty and impact of new scientific findings.

In addition to the theoretical insights, the paper also presents empirical evidence that supports the model's main conclusions. By analyzing data and trends within specific research fields, the study demonstrates that the observed patterns align with the theoretical predictions, reinforcing the idea that while knowledge is undoubtedly valuable, its interaction with references can ultimately inhibit the innovation process. The findings underscore the importance of carefully managing the balance between building on existing knowledge and fostering the conditions for novel, groundbreaking research.

References

[1] Daron Acemoglu, Ufuk Akcigit, and William R Kerr. Innovation network. *Proceedings of the National Academy of Sciences*, 113(41):11483–11488, 2016.

[2] Ajay Agrawal, John McHale, and Alexander Oettl. Artificial intelligence and scientific discovery: A model of prioritized search. *Research Policy*, 53(5):104989, 2024.

[3] Ufuk Akcigit and William R Kerr. Growth through heterogeneous innovations. *Journal of Political Economy*, 126(4):1374–1443, 2018.

[4] Ufuk Akcigit, Santiago Caicedo, Ernest Miguelez, Stefanie Stantcheva, and Valerio Sterzi. Dancing with the stars: Innovation through interactions. Technical report, National Bureau of Economic Research, 2018.

[5] Miguel Antón, Florian Ederer, Mireia Giné, and Martin Schmalz. Innovation: the bright side of common ownership? *Management Science*, 2024.

[6] Ashish Arora, Sharon Belenzon, Konstantin Kosenko, Jungkyu Suh, and Yishay Yafeh. The rise of scientific research in corporate america. Technical report, National Bureau of Economic Research, 2021.

[7] Ashish Arora, Sharon Belenzon, and Lia Sheer. Knowledge spillovers and corporate investment in scientific research. *American Economic Review*, 111(3):871–898, 2021.

[8] Pierre Azoulay, Joshua Graff-Zivin, Brian Uzzi, Dashun Wang, Heidi Williams, James A Evans, Ginger Zhe Jin, Susan Feng Lu, Benjamin F Jones, Katy Börner, et al. Toward a more scientific science. *Science*, 361(6408):1194–1197, 2018.

[9] Pierre Azoulay, Christian Fons-Rosen, and Joshua S Graff Zivin. Does science advance one funeral at a time? *American Economic Review*, 109(8):2889–2920, 2019.

[10] Francisco J Buera and Ezra Oberfield. The global diffusion of ideas. *Econometrica*, 88(1): 83–114, 2020.

[11] Richard T Carson, Joshua S Graff Zivin, and Jeffrey G Shrader. Choose your moments: Peer review and scientific risk taking. Technical report, National Bureau of Economic Research, 2023.

[12] Claude d'Aspremont and Alexis Jacquemin. Cooperative and noncooperative r & d in duopoly with spillovers. *The American Economic Review*, 78(5):1133–1137, 1988.

[13] Data File. Lessons, insights and methodological tools,". *NBER Working Paper*, 8498:40, 2001.

[14] Santo Fortunato, Carl T Bergstrom, Katy Börner, James A Evans, Dirk Helbing, Staša Milojević, Alexander M Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, et al. Science of science. *Science*, 359(6379):eaao0185, 2018.

[15] Jacob G Foster, Andrey Rzhetsky, and James A Evans. Tradition and innovation in scientists' research strategies. *American sociological review*, 80(5):875–908, 2015.

[16] Richard B Freeman and Wei Huang. Collaborating with people like me: Ethnic coauthorship within the united states. *Journal of Labor Economics*, 33(S1):S289–S318, 2015.

[17] Bronwyn H Hall and Rosemarie Ham Ziedonis. The patent paradox revisited: an empirical study of patenting in the us semiconductor industry, 1979-1995. *rand Journal of Economics*, pages 101–128, 2001.

[18] Ryan Hill and Carolyn Stein. Race to the bottom: Competition and quality in science. *Northwestern University and UC Berkeley*, 2021.

[19] Alessandro Iaria, Carlo Schwarz, and Fabian Waldinger. Frontier knowledge and scientific production: evidence from the collapse of international science. *The Quarterly Journal of Economics*, 133(2):927–991, 2018.

[20] Benjamin F Jones. The burden of knowledge and the "death of the renaissance man": Is innovation getting harder? *The Review of Economic Studies*, 76(1):283–317, 2009.

[21] Benjamin F Jones. As science evolves, how can science policy? *Innovation policy and the economy*, 11(1):103–131, 2011.

[22] Benjamin F Jones, Stefan Wuchty, and Brian Uzzi. Multi-university research teams: Shifting impact, geography, and stratification in science. *science*, 322(5905):1259–1262, 2008.

[23] Boyan Jovanovic and Yaw Nyarko. Learning by doing and the choice of technology., 1994.

[24] Boyan Jovanovic and Zhu Wang. Idea diffusion and property rights. Technical report, National Bureau of Economic Research, 2020.

[25] Bryan Kelly, Dimitris Papanikolaou, Amit Seru, and Matt Taddy. Measuring technological innovation over the long run. *American Economic Review: Insights*, 3(3):303–320, 2021.

[26] You-Na Lee, John P Walsh, and Jian Wang. Creativity in scientific teams: Unpacking novelty and impact. *Research policy*, 44(3):684–697, 2015.

[27] Ángel L López and Xavier Vives. Overlapping ownership, r&d spillovers, and antitrust policy. *Journal of Political Economy*, 127(5):2394–2437, 2019.

[28] Robert E Lucas Jr. Ideas and growth. *Economica*, 76(301):1–19, 2009.

[29] Satyam Mukherjee, Daniel M Romero, Ben Jones, and Brian Uzzi. The nearly universal link between the age of past knowledge and tomorrow's breakthroughs in science and technology: The hotspot. *Science advances*, 3(4):e1601315, 2017.

[30] Michael Park, Erin Leahey, and Russell J Funk. Papers and patents are becoming less disruptive over time. *Nature*, 613(7942):138–144, 2023.

[31] Felix Poege, Dietmar Harhoff, Fabian Gaessler, and Stefano Baruffaldi. Science quality and the value of inventions. *Science advances*, 5(12):eaay7323, 2019.

[32] Feng Shi, Jacob G Foster, and James A Evans. Weaving the fabric of science: Dynamic network models of science's unfolding structure. *Social Networks*, 43:73–85, 2015.

[33] Brian Uzzi, Satyam Mukherjee, Michael Stringer, and Ben Jones. Atypical combinations and scientific impact. *Science*, 342(6157):468–472, 2013.

[34] Alex Verstak, Anurag Acharya, Helder Suzuki, Sean Henderson, Mikhail Iakhiaev, Cliff Chiung Yu Lin, and Namit Shetty. On the shoulders of giants: The growing impact of older articles. *arXiv preprint arXiv:1411.0275*, 2014.

[35] Jian Wang, Reinhilde Veugelers, and Paula Stephan. Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy*, 46(8):1416–1436, 2017.

[36] Martin L Weitzman. Recombinant growth. *The Quarterly Journal of Economics*, 113(2): 331–360, 1998.

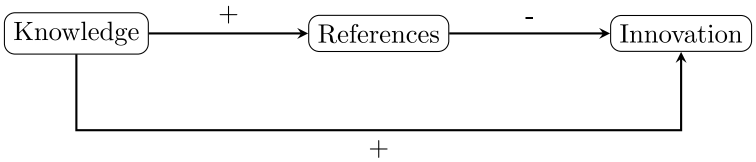


Figure 1: The impacts from knowledge on innovation

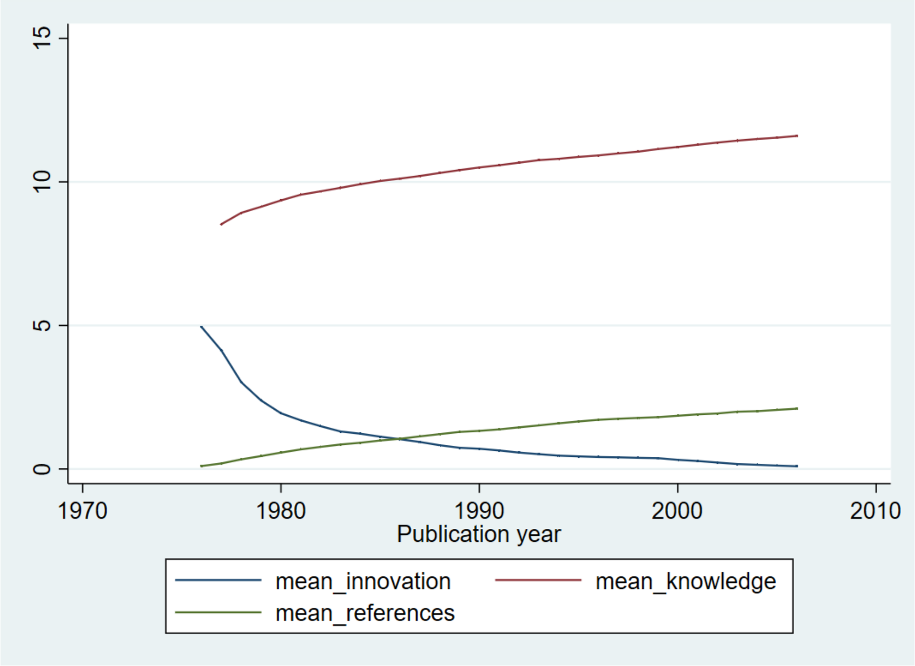
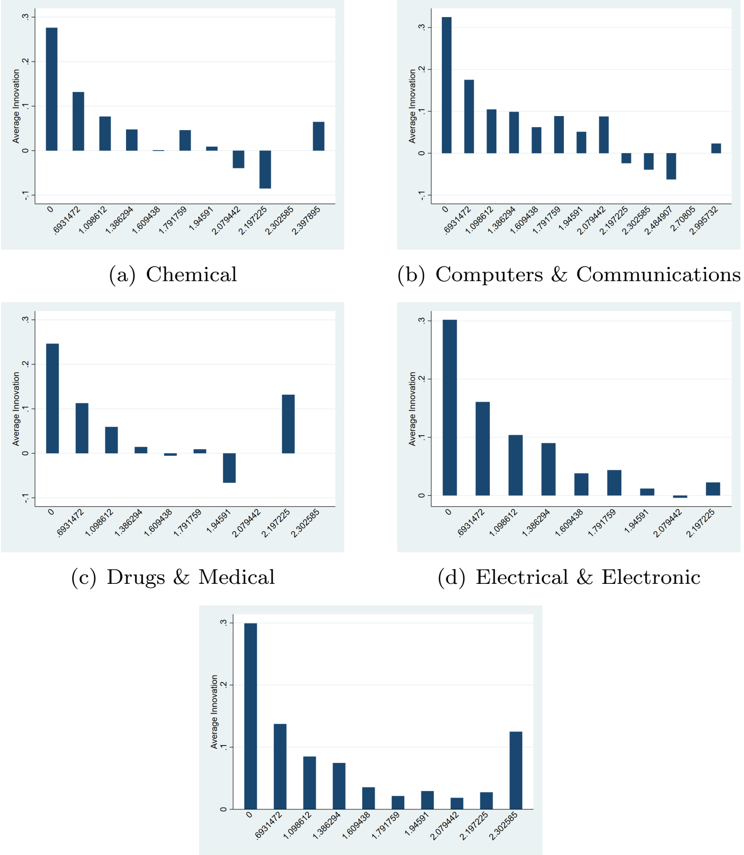


Figure 2: The changes in patents' knowledge, references and disruptiveness



(a) Chemical

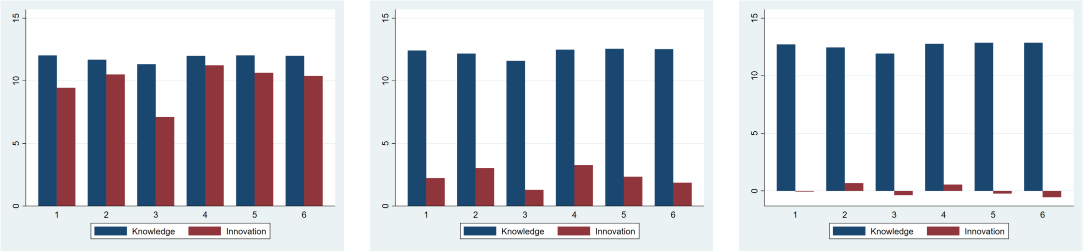
(b) Computers & Communications

(c) Drugs & Medical

(d) Electrical & Electronic

(e) Mechanical

Figure 3: Distribution of innovativeness on references in 1978
Note: The category of patents is defined by NBER patent classifications.



(a) Few References (b) Average References (c) Abundant References

Figure 4: Average knowledge and innovation among 6 NBER categories

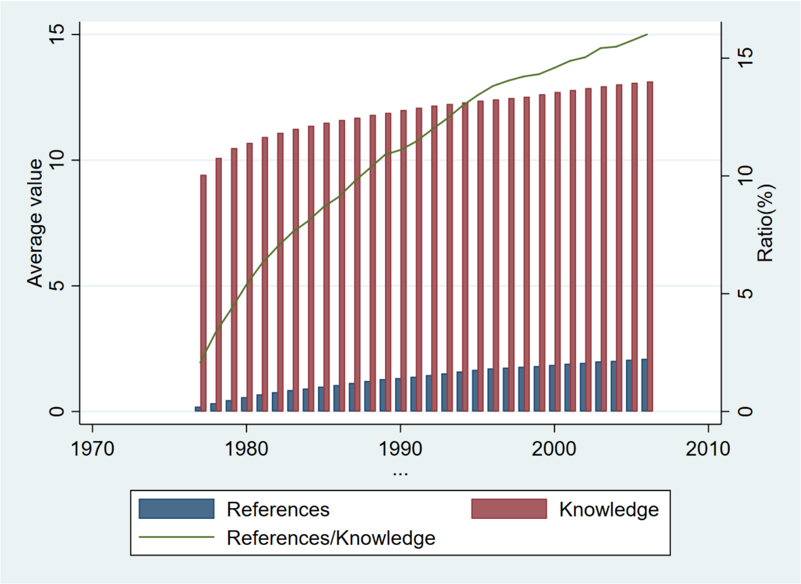


Figure 5: Absolute and relative amounts of references with respect to knowledge

Table 1: Variable Definition

Variable Name	Variable Measurement	Source
Innovation	CD-5 index	Park et al., 2023
Field Knowledge	Patents published before in the same subfield (in Logarithm)	
Subfield Knowledge	Patents published in the same year and same subfield (in Logarithm)	
References	Total backward cites from the patent	NBER Patent Project
Originality	Fields of patents cited from the patent	NBER Patent Project
Age	Mean age of work cited	Park et al., 2023
Age Variance	Dispersion in age of work cited	
Team Production	Mean number of prior works produced by team members (in Logarithm)	
Diversity	Diversity of work cited	
Competition	Average innovation in the same year and subfield	

Table 2: Summary statistics

	Mean	SD	Min	Max	N
CD-5 index	0.13	0.30	-1.00	1.00	2.9e+06
Patents published before in the same subfield (in Logarithm)	10.62	1.08	0.69	12.50	3.1e+06
Patents published in the same year and same subfield (in Logarithm)	8.29	0.79	0.69	9.88	3.1e+06
Total backward cites from the patent	1.61	0.97	0.00	6.67	2.8e+06
Originality	0.52	0.35	0.00	1.00	2.5e+06
Mean age of work cited	7.10	3.81	-36.00	30.00	2.8e+06
Dispersion in age of work cited	3.08	2.35	0.00	19.00	2.8e+06
Mean number of prior works produced by team members (in Logarithm)	1.35	1.17	0.00	7.27	3.1e+06
Diversity of work cited	0.98	0.01	0.89	1.00	3.1e+06
Average innovation in the same year and subfield	0.14	0.16	-0.11	1.00	3.1e+06

Note: This table displays pairwise correlations for the main explanatory variables relating to knowledge and references.

Table 3: Correlations between explanatory variables

Variables	(1)	2	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1)Patents published before in the same field (in Logarithm)	1.000								
(2)Patents published in the same year and field (in Logarithm)	0.801	1.000							
(3)Total backward cites from the patent	0.374	0.296	1.000						
(4)Originality	0.061	0.018	0.077	1.000					
(5Mean age of work cited	0.431	0.219	0.261	0.112	1.000				
(6)Dispersion in age of work cited	0.481	0.312	0.569	0.119	0.588	1.000			
(7)Mean number of prior works produced by team members (in Logarithm)	0.244	0.203	0.172	-0.033	0.028	0.068	1.000		
(8)Diversity of work cited	-0.213	-0.315	-0.246	0.068	-0.042	-0.120	-0.195	1.000	
(9)Average innovation in the same year and field	-0.870	-0.542	-0.423	-0.041	-0.423	-0.469	-0.249	0.307	1.000

Table 4: Regression within subfields

	(1) CD-5 index	(2) CD-10 index
Patents published before in the same subfield (in Logarithm)	0.0120***	0.0120***
	(0.000)	(0.000)
Patents published in the same year and same subfield (in Logarithm)	-0.0149***	-0.0177***
	(0.000)	(0.000)
Total backward cites from the patent	-0.0451***	-0.0529***
	(0.000)	(0.000)
Originality	0.0131***	0.0145***
	(0.000)	(0.000)
Mean age of work cited	0.00488***	0.00532***
	(0.000)	(0.000)
Dispersion in age of work cited	-0.00497***	-0.00542***
	(0.000)	(0.000)
Mean number of prior works produced by team members (in Logarithm)	0.00252***	0.00256***
	(0.000)	(0.000)
Diversity of work cited	-0.345***	-0.378***
	(0.000)	(0.000)
Average innovation in the same year and subfield	0.158***	0.171***
	(0.000)	(0.000)
Constant	0.382***	0.455***
	(0.000)	(0.000)
Observations	2454595	2465161
R^2	0.065	0.096

p-values in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 5: Regression within fields

	(1) CD-5 index	(2) CD-10 index
Patents published before in the same field (in Logarithm)	0.0330***	0.0309***
	(0.000)	(0.000)
Patents published in the same year and field (in Logarithm)	-0.0272***	-0.0290***
	(0.000)	(0.000)
Total backward cites from the patent	-0.0448***	-0.0528***
	(0.000)	(0.000)
Originality	0.0125***	0.0144***
	(0.000)	(0.000)
Mean age of work cited	0.00469***	0.00514***
	(0.000)	(0.000)
Dispersion in age of work cited	-0.00502***	-0.00555***
	(0.000)	(0.000)
Mean number of prior works produced by team members (in Logarithm)	0.00287***	0.00308***
	(0.000)	(0.000)
Diversity of work cited	-0.304***	-0.262***
	(0.000)	(0.000)
Average innovation in the same year and field	0.167***	0.188***
	(0.000)	(0.000)
Constant	0.221***	0.244***
	(0.000)	(0.000)
Observations	2454595	2465161
R^2	0.062	0.092

p-values in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 6: What makes a highly cited patent?

	(1) Total forward cites to the patent
CD-5 index	0.600***
	(0.000)
Average innovation in the same year and subfield	-0.579***
	(0.000)
Patents published before in the same subfield (in Logarithm)	-0.617***
	(0.000)
Patents published in the same year and same subfield (in Logarithm)	0.151***
	(0.000)
Total backward cites from the patent	0.300***
	(0.000)
Average references in the same year and subfield	-0.429***
	(0.000)
Originality	0.0399***
	(0.000)
Mean age of work cited	-0.0449***
	(0.000)
Dispersion in age of work cited	-0.00488***
	(0.000)
Mean number of prior works produced by team members (in Logarithm)	0.00255***
	(0.000)
Diversity of work cited	-0.816***
	(0.000)
Average innovation in the same year and subfield	0
	(.)
Constant	7.134***
	(0.000)
Observations	1884944
R^2	0.127

p-values in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 7: What affects references?

	(1)
	Total backward cites from the patent
Patents published before in the same subfield (in Logarithm)	0.186***
	(0.000)
Average references in the same year and subfield	0.657***
	(0.000)
Patents published in the same year and same subfield (in Logarithm)	-0.0452***
	(0.000)
CD-5 index	-0.955***
	(0.000)
Originality	0.136***
	(0.000)
Mean age of work cited	-0.00544***
	(0.000)
Dispersion in age of work cited	0.110***
	(0.000)
Mean number of prior works produced by team members (in Logarithm)	0.0700***
	(0.000)
Diversity of work cited	-1.707***
	(0.000)
Average innovation in the same year and subfield	0.669***
	(0.000)
Constant	0.357**
	(0.004)
Observations	2454595
R^2	0.157

p-values in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 8: Heterogenous analysis

	(1) Chemical	(2) omputers & Communications	(3) Drugs & Medical	(4) Electrical & Electronic	(5) Mechanical
Patents published before in the same subfield (in Logarithm)	-0.0586***	0.00335*	0.0125***	-0.0144***	-0.00271
	(0.000)	(0.049)	(0.000)	(0.000)	(0.252)
Patents published in the same year and same subfield (in Logarithm)	0.0593***	-0.00403*	-0.00902**	0.0165***	-0.0125***
	(0.000)	(0.043)	(0.001)	(0.000)	(0.000)
Total backward cites from the patent	-0.0458***	-0.0427***	-0.0247***	-0.0536***	-0.0512***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Originality	0.0128***	0.00191***	0.00876***	0.00440***	0.0206***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Mean age of work cited	0.00495***	0.00448***	0.00479***	0.00513***	0.00445***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Dispersion in age of work cited	-0.00493***	-0.00470***	-0.00441***	-0.00476***	-0.00537***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Mean number of prior works produced by team members (in Logarithm)	0.00150***	0.00210***	0.000264	0.00304***	0.00495***
	(0.000)	(0.000)	(0.333)	(0.000)	(0.000)
Diversity of work cited	0.123	0.442***	-0.493***	-0.180**	-0.705***
	(0.209)	(0.000)	(0.000)	(0.004)	(0.000)
Constant	0.121	-0.317***	0.463***	0.322***	0.929***
	(0.217)	(0.000)	(0.000)	(0.000)	(0.000)
Observations	376991	483002	215308	505684	436789
R^2	0.048	0.096	0.035	0.079	0.051

p-values in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

A Appendix I Theory

A.1 Single Researcher under a dynamic setting

In each period t, the firm brings up a research project (I, α) and the knowledge evolves:

$$\max_{I_t,\alpha_t} v_t = I_t - \frac{c}{2}I_t^2 + b\alpha_t - \phi(\alpha_t; k_t)I_t - \gamma\alpha_t$$

$$k_{t+1} = k_t + I_t$$

First-order conditions:

$$1 - cI_t - \phi(\alpha_t; k_t) = 0$$

$$b - \frac{\partial\phi(\alpha_t; k_t)}{\partial\alpha_t}I_t - \gamma = 0$$

The following Fig.6 shows the dynamic evolution of innovation, references, and knowledge. As time goes by, the optimal innovation would decrease over time, while references would increase, and the accumulation of knowledge would slow down.

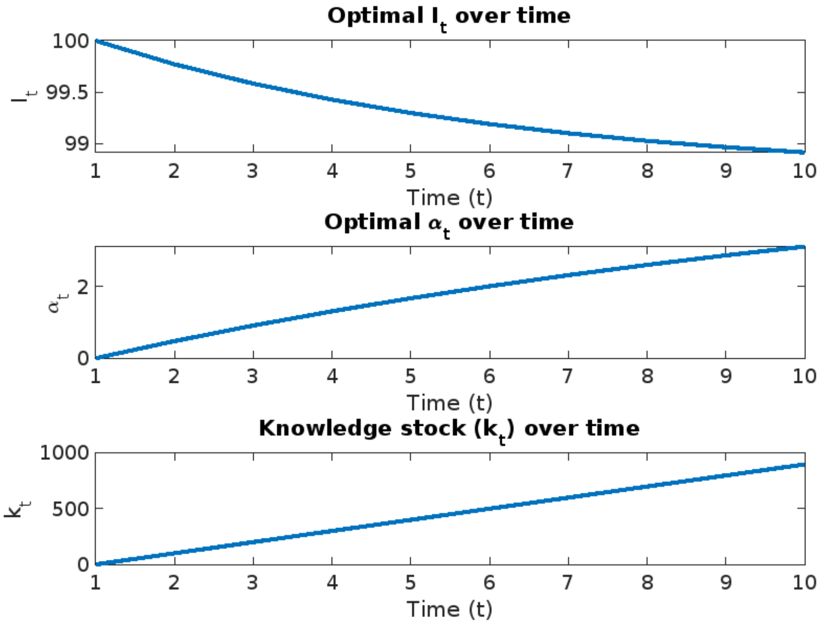


Figure 6: Single Researcher case

Note: in the simulation, $c = 0.01; b = 1; \gamma = 0.1; \phi(\alpha, k) = \frac{\alpha^2}{k+1}$.

A.2 Setting

The setting of the model is as follows. There are two researchers developing their research strategy independently. Each research strategy is a bundle of (I, α) where I_i denotes researcher i's own novel idea and α_i is her reliance on private knowledge (the past knowledge the researcher learns from training). The private knowledge could not surpass the current public knowledge $\alpha_i \in [0, k]$. Each researcher's revenue depends on both the choice of innovation and private knowledge. **There is a technical interaction between the two researchers' innovations, which implies either competition or spillover from each other.**

The game has two stages. In the first stage, both agents choose (α_i, α_j) as how many references they are going to learn; in the second stage, agents choose (I_i, I_j) as how much innovation they will input into the project, and then the payoffs realize.

The revenue of the project is symmetric and determined by both agents' choice:

$$\Pi_i(I_i, I_j, \alpha_i) = I_i - \frac{c_1}{2} I_i^2 + c_2 I_i I_j + b\alpha_i$$

where $c_1 > 0$, and $b > 0$.

Here both innovation and references add to the impact of the research, and there exists an incomplete substitution between the two items, which is measured by $b > 0$. The revenue is concave in its first argument ($\frac{\partial^2 \Pi_i}{\partial I_i^2} = -c_1 < 0$), that the revenue will increase in the innovativeness of the project, while the marginal return of innovation is changing in a diminishing rate, collaborating with the fact that innovation in a project is risky regarding of its impact. c_2 captures the nature of the strategic interactions, including the scale of the spillover effect as well as the competition between the innovations. Here we assume c_2 is symmetric between two researchers. If c_2 is positive, innovations are strategic complements and the researcher's revenue increases in both own innovation and the rival's innovation ($\frac{\partial \Pi_i}{\partial I_j} = c_2 I_i > 0$); and if c_2 is negative, innovations are strategic substitutes and rival's innovation would dampen the researcher's revenue.

The researcher i pays two sorts of costs: (i) Innovation cost. This is the cost of implementing and developing her novel idea, and the marginal cost of innovation is $\phi_i(k; \alpha_i) I_i$ with $\frac{\partial \phi}{\partial k} < 0$ and $\frac{\partial \phi}{\partial \alpha_i} > 0$, and $\frac{\partial^2 \phi}{\partial \alpha_i \partial k} < 0$. Similar to Arora et al., 2021, this means that reliance on past knowledge could help to cut the innovation cost, while references add the innovation cost. The last inequality also requires that public knowledge act as an mitigation of the marginal effect of references on the innovation cost.) (ii) Education cost. This is the cost of learning the knowledge and is represented by $C(\alpha_i) = \gamma\alpha_i$, an increasing function with $b > \gamma$. We assume there exists a stable Nash Equilibrium. This requires that $D = c_1^2 - c_2^2 > 0$, or $|c_1| > |c_2|$.

Given the above settings, agent i is maximizing the following payoff:

$$v_i = I_i - \frac{c_1}{2} I_i^2 + c_2 I_i I_j + b\alpha_i - \phi_i(\alpha_i; k) I_i - \gamma\alpha_i$$

A.3 Innovation

In the second stage, FOCs of Innovation are given as

$$1 - c_1 I_i + c_2 I_j - \phi(\alpha_i; k) = 0$$

$$1 - c_1 I_j + c_2 I_i - \phi(\alpha_j; k) = 0$$

and SOCs for both arguments are $-c_1 < 0$.
Write $\phi(\alpha_i; k)$ as ϕ_i, and $\phi(\alpha_j; k)$ as ϕ_j, then the agents' optimal choice of innovation is given as

$$I_i^* = \frac{c_1(1-\phi_i) + c_2(1-\phi_j)}{D}$$

$$I_j^* = \frac{c_1(1-\phi_j) + c_2(1-\phi_i)}{D}$$

To assure the existence of nonnegative results, it requires $\phi_i, \phi_j \leq 1$.

From the equilibrium there is

$$\frac{\partial I_i^*}{\partial \alpha_i} = -\frac{c_1}{D}\frac{\partial \phi_i}{\partial \alpha_i} < 0$$

The relationship between innovation and the rival's references is depending on

$$\frac{\partial I_j^*}{\partial \alpha_i} = -\frac{c_2}{D}\frac{\partial \phi_i}{\partial \alpha_i}$$

via the innovation cost of the rival. Note that if $c_2 > 0$, $\frac{\partial I_j^*}{\partial \alpha_i} < 0$, i.e., researcher j also decreases its innovation in response to the decrease in research by researcher i.

The response of innovation input to public science is

$$\frac{\partial I_i^*}{\partial k} = -\frac{1}{D}(c_1\frac{\partial \phi_i}{\partial k} + c_2\frac{\partial \phi_j}{\partial k}) > 0$$

$$\frac{\partial I_j^*}{\partial k} = -\frac{1}{D}(c_1\frac{\partial \phi_j}{\partial k} + c_2\frac{\partial \phi_i}{\partial k})$$

If there exists strategic complementarity between innovations, i.e., $c_2 > 0$, both firms innovate more in response to an increase in public science. Otherwise, if the innovations are strategic substitutes, then one (but not both) firm may reduce innovation. However, note that

$$\frac{\partial I_i^*}{\partial k} + \frac{\partial I_j^*}{\partial k} = -\frac{1}{D}(c_1 + c_2)(\frac{\partial \phi_i}{\partial k} + \frac{\partial \phi_j}{\partial k}) > 0$$

always holds given that $|c_1| > |c_2|$. This implies that innovation on average increases with public science.

A.4 References

In the first stage, taking into account how it impact the equilibrium choices of I_i^* and I_j^*, firm i' chooses α_i to maximize

$$v_i = \Pi_i(I_i^*, I_j^*, \alpha_i) - \phi_i(\alpha_i; k)I_i^* - \gamma\alpha_i$$

where

$$\Pi_i(I_i^*, I_j^*, \alpha_i) = I_i^* - \frac{c_1}{2}I_i^{*2} + c_2 I_i^* I_j^* + b\alpha_i$$

since both I_i^* and I_j^* are functions of α_i, the FOC w.r.t. α_i is written as

$$\begin{aligned}\frac{\partial v_i}{\alpha_i} &= \frac{\Pi_i}{\partial I_i}\frac{\partial I_i^*}{\partial \alpha_i} + \frac{\Pi_i}{\partial I_j}\frac{\partial I_j^*}{\partial \alpha_i} + b - \frac{\partial \phi_i}{\partial \alpha_i}I_i^* + \phi_i\frac{\partial I_i^*}{\partial \alpha_i} - \gamma \\ &= (\frac{\Pi_i}{\partial I_i} - \phi_i)\frac{\partial I_i^*}{\partial \alpha_i} + \frac{\partial \Pi_i}{\partial I_j}\frac{\partial I_j^*}{\partial \alpha_i} - \frac{\partial \phi_i}{\partial \alpha_i}I_i^* + b - \gamma = 0\end{aligned}$$

In the second stage, I_i^* satisfies

$$\frac{\partial v_i}{\partial I_i} = \frac{\Pi_i}{\partial I_i} - \phi_i = 0$$

by envelope theorem, the optimal α_i satisfies that

$$\frac{\partial \Pi_i}{\partial I_j}\frac{\partial I_j^*}{\partial \alpha_i} - \frac{\partial \phi_i}{\partial \alpha_i}I_i^* + b - \gamma = 0$$

The first term stands for the indirect impact on innovation from references, that it would cut the rival's innovation which has a spillover effect on firm i's payoff. The direct impact on innovation is reflected in the second term, that references hinder innovation by adding to its unit cost. The term $b - \gamma$ captures the marginal return of references on payoff. Plugging $\frac{\partial \Pi_i}{\partial I_j}$ into the above equation:

$$c_2 I_i^*\frac{\partial I_j^*}{\partial \alpha_i} - \frac{\partial \phi_i}{\partial \alpha_i}I_i^* + b - \gamma = -\frac{c_2^2}{D}\frac{\partial \phi_i}{\partial \alpha_i} - \frac{\partial \phi_i}{\partial \alpha_i}I_i^* + b - \gamma = 0 \Rightarrow \frac{\partial \phi_i}{\partial \alpha_i}I_i^* = \frac{(b-\gamma)D}{c_1^2}$$

Second-order derivative:

$$-\frac{c_1^2}{D}[\frac{\partial^2 \phi_i}{\partial \alpha_i^2}I_i^* - \frac{c_1}{D}(\frac{\partial \phi_i}{\partial \alpha_i})^2]$$

the unique existence of interior maximum requires that ϕ_i is convex in α_i and

$$\frac{\partial^2 \phi_i}{\partial \alpha_i^2}I_i^* - \frac{c_1}{D}(\frac{\partial \phi_i}{\partial \alpha_i})^2 > 0$$

At an interior maximum, the direction of the effect of public science on references is given by $\frac{\partial^2 v_i}{\partial \alpha_i \partial k}$, this is because taking the total derivative of F.O.C. w.r.t. α_i:

$$\frac{\partial v_i}{\alpha_i} = 0$$

$$\frac{\partial^2 v_i}{\partial \alpha_i^2}d\alpha_i + \frac{\partial^2 v_i}{\partial \alpha_i \partial k}dk = 0$$

$$\frac{d\alpha_i}{dk} = -\frac{\frac{\partial^2 v_i}{\partial \alpha_i \partial k}}{\frac{\partial^2 v_i}{\partial \alpha_i^2}}$$

Given that ϕ_i is convex in α_i and

$$\frac{\partial^2 v_i}{\partial \alpha_i^2} = -\frac{c_1^2}{D}[\frac{\partial^2 \phi_i}{\partial \alpha_i^2} I_i^* - \frac{c_1}{D}(\frac{\partial \phi_i}{\partial \alpha_i})^2]$$

Now take a look at the sign of $\frac{\partial^2 v_i}{\partial \alpha_i \partial k}$:

$$\frac{\partial v_i}{\partial \alpha_i} = -\frac{c_1^2}{D}\frac{\partial \phi_i}{\partial \alpha_i} I_i^* + b - \gamma$$

$$\frac{\partial^2 v_i}{\partial \alpha_i^2} = -\frac{c_1^2}{D}[\frac{\partial^2 \phi_i}{\partial \alpha_i^2} I_i^* - \frac{c_1}{D}(\frac{\partial \phi_i}{\partial \alpha_i})^2]$$

$$\frac{\partial^2 v_i}{\partial \alpha_i \partial k} = -\frac{c_1^2}{D}[\frac{\partial I_i^*}{\partial k}\frac{\partial \phi_i}{\partial \alpha_i} + I_i^* \frac{\partial^2 \phi_i}{\partial \alpha_i \partial k}]$$

or the sign of $\frac{d\alpha_i}{dk}$ is the opposite of

$$\frac{\partial I_i^*}{\partial k}\frac{\partial \phi_i}{\partial \alpha_i} + I_i^* \frac{\partial^2 \phi_i}{\partial \alpha_i \partial k}$$

The first term in the above equation is positive. The second is positive if public science and references are complements in changing the unit cost of innovation and negative otherwise. Combined together, the first term reflects the trade-off between references and innovation: public knowledge increases innovation, yet references make it more costly and cut the payoff; the second term represents the interaction between public science and research in changing innovation costs. If $\frac{\partial^2 \phi_i}{\partial \alpha_i \partial k} > 0$, that the accumulation of knowledge would amplify the negative impact of references on innovation, then both effects would result in that knowledge decreases the marginal return of references. However, if $\frac{\partial^2 \phi_i}{\partial \alpha_i \partial k} < 0$ and $|\frac{\partial^2 \phi_i}{\partial \alpha_i \partial k}|$ is large enough, $\frac{\partial^2 v_i}{\partial \alpha_i \partial k} > 0$ would hold, and knowledge would encourage references since it mitigates its negative impact on innovation and henceforth increases its marginal return.

Take the derivative with respect to k and then reorganize:

$$\frac{d\alpha_i}{dk} = -\frac{\frac{\partial^2 v_i}{\partial \alpha_i \partial k}}{\frac{\partial^2 v_i}{\partial \alpha_i^2}} > 0$$

When $\frac{\partial^2 \phi_i}{\partial \alpha_i \partial k} < 0$ and ϕ_i is convex in α_i, as k increases, α_i also increases; i.e., references increase in knowledge when references and knowledge are substitutes. This leads to the second result of the model.

A.5 Knowledge

At the equilibrium,

$$\frac{\partial v_i}{\partial k} = \frac{\partial \Pi_i}{\partial I_j}\frac{\partial I_j^*}{\partial k} - \frac{\partial \phi_i}{\partial k} I_i^* > 0$$

that the payoffs increase in the accumulation of knowledge when $c_2 > 0$.

Empirically, as k increases, the researcher's marginal returns to references depends on the supply of public science: the supply of public science will enhance the effect of references on innovation if it decreases the marginal cost of innovation. Conversely, if public science and references are substitutes, then public science could mitigate the effect of references on innovation. Under the assumption of substitutability, knowledge makes the marginal contribution of references on innovation cost smaller and more profitable to increase references, henceforth decrease innovation as well as increase optimal references. Therefore, the result in 5 indicates that references are increasing in public knowledge.

The substitutability between references and public knowledge also leads to

$$\frac{\partial^2 I_i}{\partial \alpha_i \partial k} = -\frac{c_1}{D}\frac{\partial^2 \phi_i}{\partial \alpha_i \partial k} > 0$$

that accumulation of knowledge mitigates the negative effect of references on innovation.

An example of $\phi(\alpha; k)$

Take a specific expression of $\phi(\alpha; k) = \exp(\alpha - k) \leq 1$, that

$$\frac{\partial \phi}{\partial \alpha} = e^{\alpha - k} > 0, \frac{\partial^2 \phi}{\partial \alpha^2} = e^{\alpha - k} > 0, \frac{\partial \phi}{\partial k} = -e^{\alpha - k} < 0, \frac{\partial^2 \phi}{\partial \alpha \partial k} = -e^{\alpha - k} < 0$$

then

$$I_i^* = \frac{c_1(1 - e^{\alpha_i - k}) + c_2(1 - e^{\alpha_j - k})}{D}, I_j^* = \frac{c_2(1 - e^{\alpha_i - k}) + c_1(1 - e^{\alpha_j - k})}{D}$$

and

$$\frac{\partial I_i^*}{\partial \alpha_i} = -\frac{c_1 e^{\alpha_i - k}}{D} < 0, \frac{\partial I_j^*}{\partial \alpha_i} = -\frac{c_2 e^{\alpha_i - k}}{D} < 0, \frac{\partial I_i^*}{\partial k} = \frac{c_1 e^{\alpha_i - k} + c_2 e^{\alpha_j - k}}{D} > 0$$

The the optimal α_i^* satisfies

$$\frac{e^{\alpha_i - k}}{D}[c_1(1 - e^{\alpha_i - k}) + c_2(1 - e^{\alpha_j - k})] = \frac{(b - \gamma)D}{c_1^3}$$

Combined with the optimal condition for α_j^*:

$$\frac{e^{\alpha_j - k}}{D}[c_1(1 - e^{\alpha_j - k}) + c_2(1 - e^{\alpha_i - k})] = \frac{(b - \gamma)D}{c_1^3}$$

To solve the system of equation, first substitute $u = e^{\alpha_i - k}$ and $v = e^{\alpha_j - k}$ for simplicity:

$$\begin{cases} u[c_1(1 - u) + c_2(1 - v)] = A \\ v[c_1(1 - v) + c_2(1 - u)] = A \end{cases}$$

where $A = \frac{(b-\gamma)D^2}{c_1^3}$. Combining the two equations:

$$(u-v)[(c_1+c_2)-(u+v)c_1]=0$$

There exists a symmetric equilibrium where

$$e^{\alpha_i^*-k}=e^{\alpha_j^*-k}=\frac{1\pm\sqrt{1-\frac{4A}{c_1+c_2}}}{2}<1$$

and

$$\alpha_i^*=\alpha_j^*=k+\ln\frac{1\pm\sqrt{1-\frac{4A}{c_1+c_2}}}{2}$$

when $c_1^3 \geq 4(b-\gamma)(c_1-c_2)$. Second-order condition:

$$e^{\alpha_i-k}[c_1(1-2e^{\alpha_i-k})+c_2(1-e^{\alpha_j-k})]$$

When $\alpha_i^*=\alpha_j^*=u$, SOC becomes

$$u[c_1(1-2u)+c_2(1-u)]$$

When $u=\frac{1-\sqrt{1-\frac{4A}{c_1+c_2}}}{2}$, or $2u=1-\sqrt{1-\frac{4A}{c_1+c_2}}<1$, SOC must be positive. Only when $u=\frac{1+\sqrt{1-\frac{4A}{c_1+c_2}}}{2}$, $2u=1+\sqrt{1-\frac{4A}{c_1+c_2}}>1$, SOC becomes nonpositive when

$$2u=1+\sqrt{1-\frac{4A}{c_1+c_2}}>\frac{2(c_1+c_2)}{2c_1+c_2}=1+\frac{c_2}{c_1+c_2}$$

or

$$\sqrt{1-\frac{4A}{c_1+c_2}}>\frac{c_2}{c_1+c_2}\Rightarrow 1-\frac{4A}{c_1+c_2}>\frac{c_2^2}{(c_1+c_2)^2}$$

$$\Rightarrow\frac{4A}{c_1+c_2}<\frac{c_1^2+2c_1c_2}{(c_1+c_2)^2}$$

$$\Rightarrow 4A<\frac{c_1^2+2c_1c_2}{c_1+c_2}$$

i.e., when $0\leq 4A=\frac{4(b-\gamma)D^2}{c_1^3}<\frac{c_1^2+2c_1c_2}{c_1+c_2}$, there is a unique symmetric equilibrium that

$$\alpha_i^*=\alpha_j^*=k+\ln\frac{1+\sqrt{1-\frac{4A}{c_1+c_2}}}{2}\equiv k+\ln u$$

with

$$\frac{d\alpha_i^*}{dk}=\frac{d\alpha_j^*}{dk}=1>0$$

And

$$\frac{dI_i}{dk}=\frac{\partial I_i}{\partial k}+\frac{\partial I_i}{\partial\alpha_i}\frac{d\alpha_i}{dk}=\frac{c_1(1-u)+c_2(1-u)}{D}-\frac{c_1u}{D}=\frac{c_1(1-2u)+c_2(1-u)}{D}<0$$

where the last inequality holds because of SOC.

B Appendix II

Table 9: NBER Patent Classification

NBER patent category code	NBER patent category name	NBER subcategory code (2-digit)	NBER subcategory name
1	Chemical	11	Agriculture, Food, Textiles
		12	Coating
		13	Gas
		14	Organic Compounds
		15	Resins
		19	Miscellaneous-chemical
2	Computers & Communications	21	Communications
		22	Computer Hardware & Software
		23	Computer Peripherals
		24	Information Storage
		25	Electronic Business Methods & Software
3	Drugs & Medical	31	Drugs
		32	Surgery & Medical Instruments
		33	Biotechnology
		39	Miscellaneous–Drugs & Medical
4	Electrical & Electronic	41	Electrical Devices
		42	Electrical Lighting
		43	Measuring & Testing
		44	Nuclear & X-rays
		45	Power Systems
		46	Semiconductor Devices
		49	Miscellaneous Electric
5	Mechanical	51	Materials Processing & Handling
		52	Metal working
		53	Motors, Engines, Parts
		54	Optics
		55	Transportation
		59	Miscellaneous-Mechanical

B.0.1 Heterogenous analysis

For most fields, except for Computers & Communications and Drugs & Medical, the coefficient of this variable is negative and significant. This suggests that in fields like Chemical, Electrical & Electronic, and Mechanical, the accumulation of prior field-level knowledge reduces innovation, likely due to increased complexity or saturation of knowledge. In contrast, the positive coefficient for Computers & Communications and Drugs & Medical implies that accumulated knowledge in these fields promotes innovation, perhaps because these areas are more dynamic or adaptable to knowledge recombination.

In fields like Chemical and Electrical & Electronic, the coefficient is positive, indicating that the concentration of patents within a particular year and subfield promotes innovation. This could suggest that increased activity within a subfield stimulates competition and encourages innovation. However, in other fields, such as Computers & Communications, the effect is negative, implying that a dense knowledge space may lead to crowding and limit opportunities for truly innovative breakthroughs.

The coefficient for this variable is consistently negative across all fields, which suggests that the more patents a given patent cites, the more challenging it becomes to innovate. This aligns with the idea that an over-reliance on past work can stifle creativity, as inventors may feel constrained by existing frameworks or technological paradigms. The negative effect of citations on innovation is evident across all fields, implying that directly building on prior work can limit the originality or novelty of new patents.

B.0.2 Considering the impact of reference variety

Across all models, the coefficient for total backward citations remains negative and highly significant, further reinforcing the earlier finding that citing more prior work tends to reduce the CD-5 index, a measure of innovation or patent impact. This suggests that reliance on extensive prior work may constrain the originality or disruptiveness of new inventions. In Model (2), where the squared term is included, the non-linear relationship indicates that the negative impact of citations diminishes at higher citation counts, though it remains substantial.

The positive and significant coefficient for the squared term in Model 2 suggests a non-linear relationship between backward citations and the CD-5 index. Initially, as patents cite more prior work, innovation decreases (as shown by the negative coefficient on the linear term), but beyond a certain point, the rate of decline slows. This could imply that at high levels of citation, additional references are less detrimental, possibly due to diminishing marginal returns of knowledge saturation.

The coefficient for originality is positive and significant in both Models 3 and 4. This confirms the earlier finding that patents citing a wider range of fields tend to be more innovative or impactful. Drawing from diverse knowledge sources allows for novel recombination, leading to more original and potentially disruptive inventions.

Table 10: Impact from variety of references

	(1)	(2)	(3)	(4)
	CD-5 index	CD-5 index	CD-5 index	CD-5 index
Total backward cites from the patent	-0.0616***	-0.161***		-0.130***
	(0.000)	(0.000)		(0.000)
Square of total backward cites		0.0260***		0.0196***
		(0.000)		(0.000)
Originality			0.00472***	0.00893***
			(0.000)	(0.000)
Diversity of work cited			0.217***	0.0647*
			(0.000)	(0.016)
Mean age of work cited	0.00747***	0.00534***	0.00538***	0.00409***
	(0.000)	(0.000)	(0.000)	(0.000)
Dispersion in age of work cited	-0.0102***	-0.00388***	-0.0105***	-0.00311***
	(0.000)	(0.000)	(0.000)	(0.000)
Mean number of prior works produced by team members (in Logarithm)	0.00289***	0.00188***	-0.000611***	0.00179***
	(0.000)	(0.000)	(0.000)	(0.000)
Constant	0.118***	0.182***	-0.187***	0.0920***
	(0.000)	(0.000)	(0.000)	(0.001)
Observations	2264247	2264247	2017774	2017774
R^2	0.104	0.131	0.022	0.082

p-values in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$