Combinatorial Innovation and Research Strategies: Theoretical Framework and Empirical Evidence from Two Centuries of Patent Data

Matthew Clancy

Working Paper No. 15002
January 2015
I develop a knowledge production function where new ideas are built from combinations of pre-existing elements. Parameters governing the connections between these elements stochastically determine whether a new combination yields a useful idea. Researchers use Bayesian reasoning to update their beliefs about the value of these parameters and thereby improve their selection of viable research projects. The optimal research strategy is a mix of harvesting the ideas that look best, given what researchers currently believe, and performing exploratory research in order to obtain better information about the unknown parameters. Moreover, this model predicts research productivity in any one field declines over time if new elements for combination or new information about underlying parameters are not discovered. I investigate some of these properties using a large dataset, consisting of all US utility patents granted from 1836 to 2012. I use fine-grained technological classifications to show that optimal research in my model is consistent with actual innovation outcomes, and that the model can be used to improve the forecasting of patent activity in different technology classes.
0 - Introduction

This paper develops a new model of knowledge production, based on two premises.

First, all knowledge is composed of pre-existing parts. There is no *creatio ex nihilo*, wherein new ideas or technologies spring into existence fully formed from out of a void. Look deep enough and even the most creative ideas reveal themselves to be complex structures and arrangements of parts that were already there. These “parts” may be methods, techniques, concepts, mental models, designs, relationships, conventions, symbols, materials, facts, and so forth. In the empirical exercise, I will use technology subclasses from the US Patent Classification System as proxies for these parts, where a subclass delineates “processes, structural features, and functional features” or a technology.¹ What makes an idea new and creative is not what it is built from, but what combination of parts it is built from. It is the new connections between parts that matters, not the parts in and of themselves.

Second, more often than not, ideas with similar structures and parts behave in similar ways. An incremental improvement on an idea or technology usually maintains most of the parts that make up the idea or technology, while also preserving the features and behavior we care about. For example, to make a better smartphone, it is reasonable to maintain many of the parts that make it work (touchscreens, software, batteries, cellular receiver, etc.). Conversely, we would be surprised if a radical change in all of its parts preserved the original features and behavior. To make a better jet plane, it is unreasonable to start with the design of a smartphone. The two technologies have so little in common, one would do better to go back to the drawing board.

This microfounded model of knowledge production, coupled with a simple model of an innovating agent, is consistent with several stylized facts about the knowledge discovery process. For example;

- Ideas can build on each other, but also be exhausted.
- There is a natural mechanism for knowledge to accumulate and spillover from one application to another.
- Technological paradigms can be locked in and dominate subsequent research.
- The productivity of research is variable over time.
- Incremental innovation will tend to peter out in the absence of radical innovations.

Moreover, this model is testable. Using data on 8.3 million patents, I show my model is consistent with actual researcher behavior. Furthermore, it suggests novel metrics of technological opportunity, which I show to be significant predictors of research activity.

The plan for this paper is as follows. I will proceed from two special cases that can be solved analytically, to more general models that can only be solved numerically, before testing the model empirically. After a review of some related literature (Section 1), I will first lay out the knowledge production function used in this paper, and embed this function in a simple model of a solitary agent conducting research (Section 2). I then discuss the optimal research strategy when there is no

uncertainty about model parameters (Section 3). In general, however, parameter uncertainty is a key feature of this model, and the next section discusses the nature of researcher beliefs about model parameters (Section 4). I then discuss the optimal research strategy in the special case where the model takes the form of a multi-armed bandit problem (Section 5), and then numerically solve some simulations where all model parameters are uncertain (Section 6). Since closed form solutions to the general problem are impossible, I statistically characterize features of optimal research strategies in the general setting, and discuss the relationship of these features to the existing literature on innovation (Section 7). I then turn to the testing of the model with data on US patents (Section 8). I show the propensity to patent using particular elements of knowledge is consistent with the predictions of the model (Section 9), and that this model improves the ability to predict changes in aggregate patenting behavior, relative to benchmark (Section 10). Finally, I discuss the empirical results (Section 11) before offering some thoughts on directions for further research (Section 12).

1 - Background

I now consider the motivating premises in more detail. The combinatorial nature of knowledge is clearest for physical objects like technology. Consider a laptop. Besides being an object with features and behaviors I value – for example, being able to run programs I buy or write – it is also a configuration of tightly integrated components: a screen, trackpad, keyboard, hard drive, battery, etc. All of these components existed, in some form or another, before the invention of the laptop. In an important sense, the invention of the laptop consisted in finding a configuration of suitable components that could operate in concert. A similar exercise can be performed on other technologies. For example, a car is an integrated system of wheels, guidance systems, engines, structural supports, etc.

Non-physical creations can also be understood as combinations. Works of fiction draw on a common set of themes, styles, character archetypes, and other tropes; musical compositions rely on combinations of instruments, playing styles, and other conventions; and paintings deploy common techniques, symbols, and conventions. Indeed, even abstract ideas can be understood this way. In an essay on mathematical creation, Henri Poincaré, noted, “[Mathematical creation] consists precisely in not making useless combinations and in making those which are useful and which are only a small minority.”

Weitzman (1998) is the first to incorporate this feature of knowledge creation into the knowledge production function. In Weitzman’s model, innovation consists of pairing “idea-cultivars” to see if they yield a fruitful innovation (a new idea-cultivar), where the probability an idea-pair will bear fruit is an increasing function of research effort. If successful, the new idea-cultivar is included in the set of possible idea-cultivars that can be paired in the next period. Weitzman’s main contribution is to show that combinatorial processes eventually grow at a rate faster than exponential growth processes, so that, absent some extreme assumptions about the cost of research, in the limit growth

---

3 So-called because the hybridization of ideas in the model is analogous to the hybridization of plant cultivars.
eventually becomes constrained by the share of income devoted to R&D rather than the supply of ideas. Simply put, combinatorial processes are so fecund that we will never run out of ideas, only the time needed to explore them all.

Weitzman’s model is echoed in Arthur (2009), who views all technologies as hierarchical combinations of sub-components. Arthur agrees that the laptop is a combination of screen, trackpad, keyboard, hard drive, battery, and so forth, but goes further, pointing out that, say, the hard drive, is itself a combination of disks, a read/write head assembly, motors to spin and move these assemblies, and so forth. These sub-components themselves are combinations of still further subcomponents. Ridley (2010) also proposes a model akin to Weitzman’s, arguing the best innovations emerge when “ideas have sex.” This suggests ideas are combinations of genes which, when mixed, yield new offspring. This metaphor also anticipates the second premise, since organisms with similar genes have similar phenotypes.

Because clearly there is more to invention than simple combination. It does not suffice for engineers to bolt together elements at random, nor should musicians compose with a computer program that generates arbitrary lists of instruments, players and themes. Indeed, if we go back to the musings of Henri Poincaré, the full quote is:

_In fact, what is mathematical creation? It does not consist in making new combinations with mathematical entities already known. Any one could do that, but the combinations so made would be infinite in number and most of them absolutely without interest. To create consists precisely in not making useless combinations and in making those which are useful and which are only a small minority. Invention is discernment, choice._

-Poincaré (1910), p. 324-325

Later Poincaré compares mathematical creation to the jostling of atoms, which are hoped will hook onto each other in stable configurations. He explains “The mobilized atoms are… not any atoms whatsoever; they are those from which we might reasonably expect the desired solution.” In this paper I argue that we create by trying combinations similar in their composition to ideas with desirable features.

Several papers have explored this perspective, devising ways to measure the “distance” between ideas. Jovanovic and Rob (1990) represents a technology by an infinite vector, each element of which ranges between 0 and 1. The elements of this vector have an interpretation as methods, and the value of the element indexes how the method is used. Technologes are production functions and agents learn the mapping from technology vectors to productivity via Bayesian updating. Research consists in changing the values of the elements in a vector and observing the labor productivity associated with the new vector.

---

5 For example, element $k$ might be encoding “drill for oil at location $k$” and the value between 0 and 1 encodes some measure of the depth of drilling.
A related approach is developed by Kauffman, Lobo and Macready (2000) and Auerswald et al. (2000). These papers follow Jovanovic and Rob (1990) in thinking of technologies as a large combination of distinct operations, although here the length of a technology vector is finite and each element can take on one of a finite number of states (rather than ranging over a continuous interval). The mapping between each technology vector and its productivity level is called a fitness landscape. When states are interdependent, the authors show this landscape is characterized by many local maxima. Innovation in such a model consists of exploring the fitness landscape by changing different operations. The roughly correlated nature of the landscape means small changes are likely to result in productivities that are similar to current levels, and large changes are essentially a draw from the unconditional distribution of productivity values. To reach the global maximum from any given position, it may be necessary to first traverse productivity “valleys.” Auerswald et al. (2000) uses the framework to show random deviations in a production process (akin to mutations in biology) can replicate many of the features of so-called “learning curves.”

The model that I develop in this paper combines the explicitly combinatorial framework of Weitzman (1998) with the vector based learning models of Jovanovic and Rob (1990), Kauffman, Lobo and Macready (2000) and Auerswald et al. (2000). This framework permits the derivation of many stylized facts about innovation to emerge from a common set of principles. Moreover, I test some implications of the model with a greatly expanded dataset that covers hundreds of technology sectors and nearly two centuries. While there are several papers that attempt to empirically measure the relationship between combinatorial characteristics of patents or academic papers and the number of citations received, most are not direct tests of any of the above theories and tend to rely on much smaller datasets.

2 – Model Basics

2.1 – The Knowledge Production Function

We now describe formally how ideas are created in this model.

**Definition 1: Primitive Elements.** Let $Q$ denote the set of primitive elements of knowledge $q$ that can be combined with other elements to produce ideas, where $q \in Q$.

**Definition 2: Ideas.** An idea $d$ is a complete graph with at least two nodes, where nodes correspond to elements in $Q$.

---


7 For example, Nemet (2012) and Nemet and Johnson (2012) use US patents from 1976 to 2012.
In network theory, a graph is the name applied to a set of connected points. The points are the “nodes” and the connections are termed “edges.” A complete graph is a graph where every node it contains is connected to every other node via an edge.\(^8\)

**Definition 3: Edges.** An edge \( x \in X \) connects two \( q \in Q \).

Graphs formally consist of a set of nodes and a set of 2-element subsets corresponding to the edges. Hence, I will refer to ideas as either sets of elements \( q \) or sets of edges \( x \). I will often use the notation \( x \in d \) to refer to the set of edges connecting the nodes (elements) in the graph \( d \).

We now turn to three important concepts in this model.

**Definition 4: Compatibility.** The compatibility of edge \( x \) in idea \( d \) is \( c(x,d) \in \{0,1\} \). When \( c(x,d) = 1 \) then \( x \) is compatible in \( d \). When \( c(x,d) = 0 \) then \( x \) is incompatible in \( d \).

Note that \( c(x,d) = c(x,d') \) is not generally true. The compatibility of an edge may be equal to 1 in one idea and 0 in another.

**Definition 5: Affinity.** The probability edge \( x \) is compatible defines its affinity \( a(x) \in [0,1] \).

The notions of compatibility and affinity are related as follows:

\[
c(x,d) = \begin{cases} 
1 & \text{with probability } a(x) \\
0 & \text{with probability } 1 - a(x)
\end{cases}
\] (1)

Lastly, I collapse the features of an idea into a binary indicator. Ideas are either effective or ineffective, where an idea is effective if and only if the edges of all of its constituent elements are compatible.

**Definition 6: Efficacy.** An idea \( d \) is effective, represented by \( e(d) = 1 \), iff

\[c(x,d) = 1 \ \forall x \in d \].

In all other cases, represented by \( e(d) = 0 \), idea \( d \) is ineffective.

Restated, affinity determines the probability an edge is compatible, and when all edges in an idea are compatible, the idea is effective. We may imagine ideas as sets of interacting elements that must be mutually compatible for the idea to prove useful. If any two elements are incompatible, we imagine the idea suffers a catastrophic failure that renders it unfit for use.

Whereas I believe that this formulation for the structure of ideas strikes the right balance between simplicity and realism, a few caveats are in order, which I address here.

---

\(^8\) See Chartrand (2012) for more details.
First, this model assumes every connection between elements is equally important. In reality, technology is modular, with some elements tightly coupled and others only weakly interacting. A desktop computer, for example, consists of a monitor and the computer. The elements that make up the monitor are tightly interacting, but only loosely impacted by the elements in the computer. One way to capture this feature would be to assume, as in Weitzman (1998), that new combinations become elements available for combination in the future. However, as economists have long emphasized, unintended consequences are a prevalent feature of reality. Just because an inventor did not expect two elements to interact with each other does not mean they will not do so in unanticipated ways. Suppose the probability that two elements are incompatible proceeds in two steps. First, there is some probability that the two elements interact with each other. Then, if they do, there is a second probability that they interact poorly, causing the entire assemblage to break down. This is a perfectly valid way to understand what is being captured by the single parameter affinity.

Second, the model assumes that only pairwise interactions between elements matter. There are obvious counter-examples. Suppose elements $q_1$ and $q_2$ are chemical compounds that do not react unless in the presence of a catalyst $q_3$. Or imagine that a stabilizer regulates the interaction between two components that would otherwise interact in a calamitous manner. Higher order interactions are equally plausible. One impact of allowing for interactions above the pair level would be to make learning more difficult, since observing the behavior of a single pair of elements is less informative if the presence or absence of a third element is crucial. Moreover, if we determined, for example, that only interactions between sets of three elements matter, many of the results could be derived with appropriate redefinitions (for example, affinity would now apply to sets of three elements).

The above objections are of some relevance, and I hope to explore them in more detail in future research. However, as we will see, many stylized facts about research and innovation can be derived without reference to them.

2.2 – The Conduct of Research

This production function is used by a researcher who is trying to discover effective ideas. The researcher is a risk-neutral infinitely-lived profit maximizer with a discount factor $\delta \in (0,1)$. She knows every element in set $Q$, and in each period may choose to conduct a research project on some idea $d$ built from the elements in $Q$.

**Definition 7: Possible Ideas.** The set $P_{\geq 2}(Q)$ is the set of all possible ideas that can be made from elements in $Q$. It is a power set of $Q$, containing all subsets of $Q$ with two or more elements.
**Definition 8: Eligible Ideas.** A set of eligible ideas \( \tilde{D} \) is a subset of \( \mathbb{P}_2(Q) \). It is only sensible to conduct research projects on eligible ideas, and when a research project is attempted, the idea is removed from \( \tilde{D} \) at the end of the period.

The set \( \tilde{D} \) is primarily intended to indicate the set of *untried* ideas, and so it shrinks as research proceeds. I add to this set an additional element, the null set \( d_0 \equiv \{ \emptyset \} \), which represents the option not to conduct research in a period.

**Definition 9: Available Actions.** The agent’s set of available actions is \( D \equiv d_0 \cup \tilde{D} \).

Note that because \( d_0 \not\in \tilde{D} \), if the researcher chooses not to conduct research, then this option is not removed from her action set in the next period.

In principle, the researcher “knows” every idea that can be built from elements in \( Q \), in the same sense that I “know” every economics article that can be written with words and symbols I understand. However, just as I do not know whether any of these articles are good until I think more about them, or actually write them out, the researcher does not learn if an idea is effective until she decides to conduct research on it.\(^9\) Indeed, research is costly, requiring investments of time and other resources. I assume that research on any idea has cost \( k(d) \), known to the researcher, and that the option \( d_0 \), to do nothing, has \( k(d_0) = 0 \).

The return from conducting research is a reward \( \pi(d) \), also known to the researcher, which is received if the idea is eligible and discovered to be effective. This reward could indicate a prize for innovation from the government, or the sale of patent rights over the idea to a firm, or some other incentive for innovation.

**Assumption 1: The Value of Ideas.** I assume \( \pi(d) \) is fixed and known to the researcher with certainty. If a research project reveals \( e(d) = 1 \), the researcher receives \( \pi(d) \) in the same period.

Because each chosen idea is removed from the set of eligible ideas \( D \) at the end of a period, researchers cannot claim a prize for the same idea multiple times. I assume the reward value of the outside option \( d_0 \) is always zero.

Hence, a researcher who chooses to conduct research on idea \( d \) expects to receive a net value of:

---

\(^9\) Jorge Luis Borges tells a parable of an infinite library containing books with every combination of letter and punctuation mark. In this library, a book telling the truth of life must exist, since every possible book exists, but finding the book and verifying it is true amongst all the gibberish and babel is a daunting task for the library inhabitants. See Borges (1962).
\[ \pi(d)E[e(d)] - k(d) \]  

The idea is successful with probability equal to the expected efficacy \( E[e(d)] \), in which case the researcher obtains a reward \( \pi(d) \). Whether the idea succeeds or not, the researcher pays up front research costs \( k(d) \). This formulation of the innovator’s problem is not unusual, except for the term \( E[e(d)] \), which is determined by the knowledge production function described earlier.

Recall that \( e(d) = 1 \) (an idea is effective) if and only if all of its edges are compatible. This implies the probability distribution of \( e(d) \) is:

\[
e(d) = \begin{cases} 
1 & \text{with probability } \prod_{x \in d} a(x) \\
0 & \text{with probability } 1 - \prod_{x \in d} a(x) 
\end{cases}
\]  

Therefore:

\[ E[e(d)] = \prod_{x \in d} a(x) \]  

Hence, if the researcher knows the affinity of each edge, she can compute the expected efficacy of every idea. In general, we will assume the researcher does not know the true affinity of each edge and must infer its likely value from the outcomes of research projects.

Before proceeding to this more complex and realistic case, we discuss the special case where the agent knows the true affinity of each edge with certainty. This exercise is useful for understanding the behavior of researchers in settings where their knowledge of the true affinity is close to certain, or where learning processes are costly and relatively uninformative. In such settings, researchers act “as if” they know the true affinity of each edge.

3 – Special Case 1: Affinity is Known

Suppose there is a researcher with perfect knowledge of \( Q, \pi(d), k(d), \) and \( a(x) \). However, the researcher does not know whether ideas are effective or not, until she initiates a research project on them.

In any period, the researcher’s problem is to determine which ideas to attempt, and in which order. Define the expected present discounted value of the optimal strategy in period \( t \) as:

\[ V_t = \sum_{\tau=0}^{\infty} \delta^\tau \left( \pi(d_{t+\tau})E[e(d_{t+\tau})] - k(d_{t+\tau}) \right) \]  

(5)
where \( d_t \) denotes the optimal decision in period \( t \). Note that, because no relevant information is revealed by the outcome of research, the optimal choice in period \( t + \tau \) depends only on information available in period \( t \). Recall also that researchers can always choose \( d_0 \), to obtain a payoff of zero with certainty. Because the set of eligible ideas is finite, the researcher will always resort to choosing \( d_0 \) in the end.

The optimal strategy in the absence of learning is simple:

**Remark 1: Optimal strategy with certainty.** In each period, the optimal strategy is to choose the eligible idea with the highest \( E[e(d)]\pi(d) - k(d) \) so long as \( E[e(d)]\pi(d) - k(d) \geq 0 \). If no idea satisfies this, then choose \( d_0 \).

Because there is no learning in this model, the researcher’s problem collapses to choosing the order in which to consume a set of lotteries. Because the researcher is risk-neutral and discounts the future, she orders these lotteries in descending expected (net) value. Furthermore, the researcher never attempts a lottery where cost exceeds expected value. Because there is no learning, and because the researcher uses up the best lotteries first, we have the following remark.

**Remark 2: Value decreases over time.** The anticipated value of research declines over time, i.e.,

\[
V_t \geq V_{t+\tau} \quad \forall t, \tau > 0
\]  

(6)

In the absence of learning, the best research ideas are used up (“fished out”) first, so that the value of conducting research falls over time. This stands in contrast to the prevalent view that knowledge is cumulative, and that today’s researchers can accomplish more than their forebears by building on their accomplishments (“standing on the shoulders of giants”). This result is not general, of course, but a consequence of my modeling the set of ideas as fixed.

Next, it follows from equations (4) and (5) that:

**Remark 3: Value rises with affinity (certainty).** The expected value of research is rising in any \( a(x) \), i.e., \( \partial V_t / \partial a(x) \geq 0 \).

**Remark 4: Change in value rises with affinity of complementary edges:** The change in the value of research as a function of any \( a(x) \) is weakly increasing in the affinity of edges that belong to the same ideas, i.e.,

\[
\frac{\partial^2 V_t}{\partial a(x) \partial a(x')} \geq 0 \text{ if } x, x' \in d \text{ otherwise } \frac{\partial^2 V_t}{\partial a(x) \partial a(x')} = 0
\]  

(7)
Remark 3 states that research is more valuable if any two elements are more likely to be compatible, since this raises the probability any given research project will prove effective. Remark 4 states that this gain is higher if both elements belong to ideas whose other edges are also likely to be compatible. Intuitively, moving the affinity from 0 to 1 has no impact on the likelihood an idea will be effective, if the other edges are certain to be ineffective, but makes all the difference if they are certain to be effective. To summarize this section, when the researcher knows the affinity of edges with certainty, the optimal strategy is myopic, always choosing the research project with the highest reward today. The expected value of research is rising in $E[a(x)]$, but falling over time, as the best ideas are progressively fished out.

In general, however, the researcher does not know the affinity of an edge with certainty. I turn to a discussion of this case next.

4 - Research and Learning

4.1 – Information Revealed From Research

I initially supposed that a researcher’s choice of research project had one goal, namely, to discover if an idea is effective or not. I now add an additional objective – to learn which edges in an idea are compatible and incompatible. In addition to a possible reward if the idea is effective, a research project over idea $d$ now yields a packet of information. Specifically, a research project over idea $d$ reveals the efficacy $e(d)$ of the idea and the compatibility $c(x,d)$ for some, but not necessarily all, $x \in d$.

Which edges are revealed is not a trivial matter, and can easily make the model intractable or embed unrealistic features. The following procedure has two main virtues, discussed more below. First, I believe it has realistic features about what can be learned from successes and failures. Second, it keep beliefs independent, which helps maintain tractability.

The information about the compatibilities $c(x,d)$ of each edge $x \in d$, is generated by the following stochastic process.

**Assumption 2: Learning From Research (Frustrations of Failure).** The information revealed about the compatibility of edges is determined according to the following procedure:

1. An edge $x \in d$ is randomly drawn with equal probability from among the edges whose compatibility has not already been selected.
2. This compatibility is added to the research project’s revealed information.
3. If $c(x,d) = 1$ and unselected edges remain, we return to step one and repeat the above procedure. If $c(x,d) = 0$ or if no unselected edges remain, we do not add any more compatibilities to the research project’s revealed information.
When this procedure is completed, the researcher observes a packet of information $\omega(d)$. If an idea is effective, this revelation procedure will reveal that all of its edges are compatible. If the idea is ineffective, it will reveal some, but possibly not all, of the compatibilities of the edges that make up $d$. Specifically, the above revelation procedure will never reveal more than one edge is incompatible.\(^\text{10}\)

This revelation mechanism is meant to capture the frustrations of failure. Researchers often have some indication of where things began to go wrong – for example, a proof step that does not go through, or an engine part that overheats – but a full understanding of why the idea failed is often elusive. This partial revelation of information about what part of the innovation failed is captured by the fragmentary knowledge of which edges are compatible and incompatible when an idea is ineffective.

Note also the researcher is unlikely to learn much from a research project with many incompatible edges, because the probability of encountering an incompatibility that stops the revelation process early is high. This is meant to capture the notion that we do not, on average, learn as much from a project that is wrong on many levels. When an idea has only one or two incompatible edges, the researcher may learn a lot or a little by trying the research project. If she is lucky, it is the kind of idea where, although the idea is ineffective, she can get a long way before hitting a roadblock. Such a research project might be represented by one where the first incompatible edge is only reached after a long series of compatible ones. If she is unlucky, the idea is the kind in which it is very difficult to make any headway until a certain problem is cracked. This would be represented by an idea where the revelation of compatibilities is quickly stopped by an incompatible edge.

4.2 - Rational Beliefs

Though the researcher does not observe affinity $a(x)$ directly, she can make educated guesses based on the tendency of $x$ to be compatible or incompatible. Using these estimates of $a(x)$, she can estimate the probability an idea will be effective. In more formal terms, a crucial part of the discovery process is the inference of likely affinity values, given the information revelation process.

Each observation of compatibility is the outcome of a Bernoulli trial governed by the edge’s true affinity, with the two possible states being compatibility (probability $a(x)$) or incompatibility (probability $1 - a(x)$). Given $s$ instances of compatibility (“success”) from $n$ total observations, the researcher updates her beliefs according to Bayes law under the Bernoulli distribution:

---

\(^{10}\) This assumption could be made more general with the introduction of a parameter $\eta \in [0,1]$, so that when an edge with $c(x,d) = 0$ is revealed, the revelation procedure stops with probability $\eta$. The model presented here is then the special case with $\eta = 1$.  

12
where \( \left( \frac{n}{s} \right) \tilde{a}^s (1 - \tilde{a})^{n-s} = P(s, n \mid a(x) = \tilde{a}) \). To solve equation (8), we need to know the prior distribution of \( a(x) \). Making the following assumption is necessary for model tractability:

**Assumption 3: Independence of Affinity.** The researcher believes \( a(x) \) is independently distributed for all \( x \).

As long as this assumption stands, the updating of beliefs about any \( a(x) \) depends only on observations on the edge \( x \) alone. If \( a(x) \) were not independently distributed, it would be necessary to also take into account the observations on correlated edge, greatly complicating the problem.

In the remainder of the paper, I will assume the researcher’s beliefs follow a beta distribution, which has the useful property of being the conjugate family of a Bernoulli distribution. The conjugate family for a distribution defines a class of distributions such that, if the prior distribution belongs to the conjugate family, then the posterior distribution will as well. In this case, if the prior distribution for an affinity belongs to the beta distribution, then after updating the researcher’s beliefs by observing a set of Bernoulli trials, the posterior distribution will also be a (different) beta distribution. This structure maintains a constant form for the beliefs of the researcher as her information varies. Moreover, the form of a beta-distribution is sufficiently flexible to enable us to explore many kinds of assumptions about the prior beliefs of the researcher.

The beta distribution of \( a(x) = \tilde{a} \) takes the form

\[
\Pr(a(x) = \tilde{a}) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \tilde{a}^{\alpha-1} (1 - \tilde{a})^{\beta-1}
\]

where \( \Gamma(x) \) is the gamma function. The distribution’s domain is over the \([0,1]\) interval with its shape governed by the parameters \( \alpha > 0 \) and \( \beta > 0 \). Changing \( \alpha \) and \( \beta \) can yield a centered bell shape, highly skewed distributions, and U-shaped distributions. It can be shown that, given a beta distribution: \(^{11}\)

\[
\int_0^1 \left( \frac{n}{s} \right) \tilde{a}^s (1 - \tilde{a})^{n-s} P(a(x) = \tilde{a}) d\tilde{a} = \left( \frac{n}{s} \right) \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(s + \alpha)\Gamma(n-s + \beta)}{\Gamma(n + \alpha + \beta)}
\]

\(^{11}\) See Casella and Berger (2002) p. 325.
Combining equations (8), (9), and (10), the updated beliefs of the researcher given \( s \) instances of compatibility from \( n \) total observations is:

\[
Pr(a(x) = \tilde{a} | s, n) = \frac{\Gamma(n + \alpha + \beta)}{\Gamma(s + \alpha)\Gamma(n - s + \beta)} \tilde{a}^{s+\alpha-1}(1-\tilde{a})^{n-s+\beta-1}
\]  

(11)

Note that this is equivalent to a beta distribution with \( \alpha' = s + \alpha \) and \( \beta' = n - s + \beta \). Hence, defining \( \alpha(x) \) and \( \beta(x) \) to be the initial parameters governing the prior beliefs about of \( a(x) \), after observing \( s(x) \) instances of compatibility and \( n(x) - s(x) \) instances of incompatibility, the researcher believes \( a(x) \) to be governed by a beta distribution with parameters \( \alpha(x) + s(x) \) and \( \beta(x) + n(x) - s(x) \). The expected value of such a distribution\(^{12}\) is given by:

\[
E[a(x) | n(x), s(x)] = \frac{\alpha(x) + s(x)}{\alpha(x) + \beta(x) + n(x)}
\]  

(12)

Note that as the number of observations grows large, the expectation converges to \( s(x) / n(x) \), which is simply the proportion of observations where compatibility is observed. The sum \( \alpha(x) + \beta(x) \) determines the relative weight put on new observations and the initial beliefs, and is a measure of initial certainty.

When a researcher knows the affinity of all edges with certainty, we noted:

\[
E[e(d)] = \prod_{x \in d} a(x)
\]  

(13)

and given that the probability distributions corresponding to the affinity parameters for each edge are always independent of each other, this can be expressed as:

\[
E[e(d)] = \prod_{x \in d} E[a(x)]
\]  

(14)

where \( E[a(x)] \) is given by equation (12).

To highlight some of the features the addition of uncertainty in affinity adds to this model, I now present a special case whose solution can be characterized analytically.

5 – Special Case 2: Research as a Multi-armed Bandit Problem

To isolate the effects of learning in the researcher’s problem, I reformulate the problem as a multi-armed Bernoulli bandit problem. While this imposes strong restrictions on the model, the advantage

is that there exists a large literature on such problems. In such problems an agent must choose between $n$ options, each of which offers a reward with a fixed probability unknown to the agent. Over time, as the researcher observes the frequency with which she receives a reward from any given option, she obtains a progressively better estimate of the underlying probability of receiving a reward from that option. The agent’s problem is to balance a myopic strategy that selects the option currently believed to be most favorable, and a far-sighted strategy which seeks to gather information on other options so that the true best option can be found. Essentially, she must make a trade-off between exploitation and exploration.

Such models have a well-known solution technique, called a Gittins index (discussed below). My model takes the form of a multi-armed Bernoulli bandit problem under the following conditions:

1. The set of elements $Q$ is infinite.
2. Only ideas of size 3 are eligible.
3. $\pi(d) = 1$ and $k(d) = k$ for all eligible $d$.
4. The elements of $Q$ can be partitioned into $n+1$ subsets. Denote the first $n$ subsets by $H_i$, where $i = 1,\ldots,n$. Each $H_i$ contains just two elements, and no $H_i$ has any element in common. Denote the last set $Q_{-H}$.
5. Let the elements $q_0(x)$ and $q_i(x)$ be connected by edge $x$. The researcher correctly believes:
   a. $a(x) = 0$ if $q_0(x), q_1(x) \in Q_{-H}$
   b. $a(x) = 0$ if $q_0(x) \in H_i$ and $q_1(x) \in H_j$ where $i \neq j$.
   c. $a(x) = 1$ if $q_0(x) \in H_i, \forall i = 1,\ldots,n$ and $q_1(x) \in Q_{-H}$ or vice-versa.
   d. $a(x)$ follows a beta distribution with parameters $\alpha_i$ and $\beta_i$ if $q_0(x), q_1(x) \in H_i$.

These conditions are illustrated in Figure 1, where I have labelled the elements that form an idea $q_i$, with $i = 1,\ldots,9$ pictured. The sets $H_1$, $H_2$, and $H_3$ each consist of 2 elements, with the remaining elements in set $Q_{-H}$ (condition 4). Since the edge connecting $q_8$ and $q_9$ both belong to the set $Q_{-H}$, the researcher believes it has an affinity of zero (condition 5a). Similarly, because the edge connecting $q_1$ and $q_4$ spans $H_1$ and $H_2$, the researcher believes it to have affinity of zero (condition 5b).

Lastly, consider the idea composed of $\left(q_5, q_6, q_7\right)$. The edges spanning $\left(q_5, q_7\right)$ and $\left(q_6, q_7\right)$ connect a point in $Q_{-H}$ to points in $H_3$, and so the researcher believes these have expected affinity.

---

equal to 1 (condition 5c). Lastly, since the edge connecting \((q_5, q_6, q_7)\) is entirely contained in set \(H_3\), the researcher believes its affinity is governed by a beta distribution (condition 5d).

Under these conditions, the agent would like to choose ideas with the highest expected efficacy, since all ideas have the same prize value and cost. Moreover, the only ideas he believes have a non-zero efficacy are those like \((q_5, q_6, q_7)\) which connect an element in the set \(Q_{-H}\) to the edge contained inside some set \(H_i\). Indeed, any element would work just as well: \((q_5, q_6, q_8)\) or \((q_5, q_6, q_9)\) would both have the same expected efficacy as \((q_5, q_6, q_7)\). Connecting an element to the edge \(x\) belonging to \(H_i\) has expected efficacy \(E[e(d)] = E[a(x_i)]\), where \(x_i\) denotes the edge in \(H_i\).

**Figure 1: Example Set**

This simplifies his problem to the choice of which \(x_i\) to attempt, balancing the expected net reward of

\[
E[a(x_i)] - k
\]

against the value of learning more accurately the true value of each \(a(x_i)\). Specifically, the researcher’s problem is characterized by a Bellman equation of the form:

\[
V(B) = \max \left\{ \max_{x_i} \left[ \frac{\alpha_i}{\alpha_i + \beta_i} \{1 + \delta V(B_{i+})\} + \frac{\beta_i}{\alpha_i + \beta_i} \delta V(B_{i-}) - k \right], 0 \right\} \tag{16}
\]

where

\[
B = ((\alpha_1, \beta_1), \ldots, (\alpha_i, \beta_i), \ldots, (\alpha_n, \beta_n))
\]

\[
B_{i+} = ((\alpha_1, \beta_1), \ldots, (\alpha_i + 1, \beta_i), \ldots, (\alpha_n, \beta_n))
\]

\[
B_{i-} = ((\alpha_1, \beta_1), \ldots, (\alpha_i, \beta_i + 1), \ldots, (\alpha_n, \beta_n)) \tag{17}
\]
Taking this equation from left to right, the researcher’s problem is first to choose whether or not to conduct research. If she does not, she obtains 0 with certainty. Moreover, if the researcher ever chooses to quit research in one period, she will do so in all subsequent periods, since her information and action set will be unchanged in the following period.

If the researcher does choose to conduct research, she must decide the best edge $x_i$ to select. With probability $\alpha_i / (\alpha_i + \beta_i)$, the idea is successful and the researcher obtains a reward equal to 1. Moreover, in the next period, she will update her beliefs in accordance with equation (12), so that her beliefs are described by the vector $B_{i+}$. Therefore, in the next period, she obtains $V(B_{i+})$, discounted by $\delta$. Conversely, with probability $\beta_i / (\alpha_i + \beta_i)$ the idea is ineffective and she obtains no reward this period. Moreover, if an idea is ineffective, the revelation mechanism will reveal one (and only one) edge to be incompatible. Because the researcher’s beliefs are correct, she will observe edge $x_i$ is incompatible, and update her beliefs to vector $B_{i-}$. In the next period, she will obtain $V(B_{i-})$, discounted by $\delta$. Finally, in either case, she pays $k$ to conduct research.

This formulation is equivalent to a standard multi-armed bandit problem with expected affinity of each edge corresponding to the fixed Bernoulli probability of receiving a reward. Using the Gittins Index approach, we can obtain some clear insights, summarized in the following remarks.

**Remark 5: Optimal strategy with learning.** The optimal strategy in every period is to choose the option with the highest Gittins Index $\lambda_i(\alpha_i, \beta_i)$ where

$$\lambda_i(\alpha_i, \beta_i) \equiv \sup \{ \lambda_i : v_i(\alpha_i, \beta_i, \lambda_i) = 0 \}$$

and $v_i(\alpha_i, \beta_i, \lambda_i)$ given by:

$$\max \left\{ \frac{\alpha_i}{\alpha_i + \beta_i} (1 + \delta v_i(\alpha_i + 1, \beta_i, \lambda_i)) + \frac{\beta_i}{\alpha_i + \beta_i} \delta v_i(\alpha_i, \beta_i + 1, \lambda_i) - k - \lambda_i , 0 \right\}$$

A proof is presented in the appendix. See Gittins, Glazebrook, and Weber (2011) for more discussion.

Note that $v(\alpha_i, \beta_i, \lambda_i)$ only depends on the beta parameters of one edge, $x_i$. This decomposes the $n$-dimensional choice problem into $n$ one-dimensional problems. The Gittins index $\lambda_i(\alpha_i, \beta_i)$ can be thought of as a riskless payment the researcher can receive in lieu of the reward from choosing $x_i$, chosen so the researcher is exactly indifferent between the two options. It accounts for the expected value in this period, equal to $\alpha_i / (\alpha_i + \beta_i) - k$, plus the prospects of achieving a better or worse outcome in subsequent periods, as the researcher’s beliefs are updated. Choosing the highest Gittins index in every period therefore accounts for the rewards in the current period, plus the potential gains from better information.
Note, however, that the researcher’s expected payoff is not equal to the Gittins index. This is because a Gittins index is computed with reference only to one edge. The index tells the researcher what to do, but it does not say how much she should expect to make.

However, we may make a few remarks about the value of research in this setting.

**Remark 6: Value rises with affinity (value of learning).** The expected value of research is nondecreasing in \( \alpha_i \).

In this model, value only comes from obtaining rewards, which occur with probability \( \alpha_i / (\alpha_i + \beta_i) \). If the researcher fixes an optimal strategy, and then one \( \alpha_i \) is increased, the researcher cannot be worse off. Neither will she be worse off if we let her re-select the optimal strategy.

Multi-armed bandit problems also have the following feature:

**Remark 7: Stick with the winner.** The optimal strategy follows a “stick-with-the-winner” formulation, i.e.,

\[
\lambda_i(\alpha_i + 1, \beta_i) > \lambda_i(\alpha_i, \beta_i) > \lambda_i(\alpha_i, \beta_i + 1)
\]  

A proof is presented in Bellman (1956).

Once an idea has been found successful, in this model, the probability it will be successful in the next period as well is increased, which makes the choice still more favorable in the next period and the opportunity cost of trying something else higher. Therefore, the researcher always sticks with a winning edge, at least until it stops working (although this is not sufficient for him to switch his strategy either).

Besides preferring winners, the researcher also prefers the edge about which less is known:

**Remark 8: Favor uncertainty.** When the expected reward is the same, an optimal strategy chooses the option where more is learned, i.e.,

\[
\lambda_i(\alpha_i, \beta_i) > \lambda_i(\alpha_i, \beta_i + 1)
\]  

if \( m > 1 \).

A proof is presented in Gittins and Wang (1992).

Note that the expected value of a beta distributed variable with parameters \( (\alpha, \beta) \) and \( (ma, mb) \) is the same, since:

\[
\frac{ma}{ma + mb} = \frac{\alpha}{\alpha + \beta}
\]  

(22)
However, an edge with \((\alpha + 1, \beta)\) has a higher expected value than one with \((m\alpha + 1, m\beta)\). In the next period, the potential benefits are higher for the more uncertain idea. At the same time, the potential downsides also looms larger for the more uncertain choice. However, since the researcher always has the option to quit research, downside risks are capped at 0. This leads agents to prefer ideas from which they can learn more, that is ideas where they have less certainty about the affinity of their edges.

Finally, the above results imply the following.

**Remark 9: Value is rising and concave in success:** The expected value of research is increasing in the number of times any given edge is found to be effective, denoted \(s(x)\), and bounded from above by \((1 - k)/(1 - \delta)\).

That \(V(B)\) is increasing in \(s(x)\) is simply a reformulation of Remark 8, since the researcher’s \(\alpha_i\) parameter is updated to \(\alpha_i + s(x)\) after observing \(s(x)\) instances of compatibility. Moreover, since researchers stick with winners, as \(s(x)\) continues to increase, the expected payoff begins to resemble the payoff from simply playing the same edge in each period. This is a concave function of \(s(x)\), bounded from above by \((1 - k)/(1 - \delta)\).

To summarize, in the multi-armed bandit formulation of the researcher’s problem, the optimal strategy is a mix of myopic and far-seeing strategies, since the Gittins index takes into account both the immediate payoff and the distant future payoffs. Researchers, somewhat paradoxically, prefer both proven research paths (they stick with winners) and unproven and untested research projects (they favor uncertainty). The value of research increases when research is successful, so that research is cumulative and has a standing-on-the-shoulders-of-giants effect. However, eventually, the payoff from success stops increases as it approaches a ceiling, given by the present discounted value of a successful research project in every period.

However, to derive these results, we had to rely on some extreme assumptions that simplified the combinatorial dynamics of this model. In the next section, we will incorporate uncertainty about the true affinity into a more general setting.

**6 – The General Case**

**6.1 – Definitions**

To implement the Bayesian belief updating presented earlier, I summarize the researcher’s beliefs by the vector \(B\) where:

\[
B = \left( (\alpha_1, \beta_1), \ldots, (\alpha_i, \beta_i), \ldots, (\alpha_{m(N)}, \beta_{m(N)}) \right)
\]  

(23)
and \( m(n) = n(n-1)/2 \). These beliefs are now updated via the stochastic vector \( \omega(d) \) of information revealed by a research project over idea \( d \). This vector has the same number of elements as \( B \) and is defined so that after conducting a research project on \( d \), the updated beliefs vector \( B' \) is given by:

\[
B' = B + \omega(d)
\]  
(24)

For example, if a research project on some \( d' \) reveals edge \( x_1 \) is compatible and edge \( x_{m(N)} \) is incompatible, and does not reveal any other information, then \( \omega(d) \) takes the form:

\[
\omega(d') = ((1,0),(0,0),\ldots,(0,0),(0,1))
\]  
(25)

Of course, \( \omega(d_0) = ((0,0),\ldots,(0,0)) \) by assumption: agents learn nothing when they choose not to do a research project.

In this paper I limit attention to the case where the researcher has no competition, thereby evading strategic considerations. Hence the researcher’s problem is to find an optimal policy function \( d^*(D,B) \) mapping from available actions \( D \) and beliefs \( B \) to an optimal choice.

The policy function \( d^*(D,B) \) maximizes:

\[
V(D,B) = E\left[ \sum_{t=0}^{\infty} \delta^t \left( e\left(d^*(D_t,B_t)\right)\pi\left(d^*(D_t,B_t)\right) - k\left(d^*(D_t,B_t)\right) \right) \right]
\]  
(26)

where

\[
D_{t+1} = d_0 \cup \left( D_t \setminus d^*(D_t,B_t) \right)
\]

\[
B_{t+1} = B_t + \omega\left(d^*(D_t,B_t)\right)
\]  
(27)

The payoff function in (26) is the discounted sum of expected per-period returns from choosing idea \( d^*(D_t,B_t) \) in period \( t \). The researcher obtains \( \pi\left(d^*(D_t,B_t)\right) \) if \( d^*(D_t,B_t) \) is effective, which occurs with probability \( E[e\left(d^*(D_t,B_t)\right)] \), but pays \( k\left(d^*(D_t,B_t)\right) \) either way. As noted earlier, if the researcher chooses \( d_0 \), then \( \pi(d_0) = E[e(d_0)] = k(d_0) = 0 \).

It is instructive to write equation (26) as a Bellman equation:

\[
V(D,B) = \max_{d \in D} \left\{ E[e(d)]\pi(d) - k(d) + \delta E[V(D',B')] \right\}
\]  
(28)

where
This formulation makes clear that the choice of idea has a payoff in the current period, but also an impact on the future, through the $E[V(D', B')]$ term. Agents may prefer to take a loss in the current period, in order to learn and increase their payoffs in the future. This formulation also makes clear that if it is ever optimal for the researcher to choose $d_0$ at some stage, then it is optimal for her to do so in every subsequent period because such a choice ensures $D' = D$ and $B' = B$.

6.2 – Solving the General Case

There is no general closed-form solution for this problem but one can apply the principle of backwards induction. Because the set of possible ideas is finite, and (as noted above) the researcher never pauses then restarts research, we can be certain that from period $|D|$ on the researcher will choose $d_0$ in every stage. Thus, $V(D, B) = 0$ in period $|D|$ for certain, no matter what the researcher does in period $|D| - 1$. Since the researcher knows the next period payoff with certainty, she can find the best choice in period $|D| - 1$, and work backwards toward period 0.

Although a closed form solution is not feasible, I would like to demonstrate that an optimal solution is characterized by many of the same features detailed in Remarks 1-8. Unfortunately, this model is also beset by the “curse of dimensionality,” so that even a general numerical approximation is difficult to obtain. The curse of dimensionality is a name applied to problems where the computational resources needed to solve them grow very quickly.\textsuperscript{14} For example, consider the number of parameters needed to characterize the state-space for a problem with 3, 4, and 5 elements. Given 3 elements, there are 3 affinities, the beliefs of which are described by 2 parameters each, so that $B$ has six dimensions. Three elements also implies there are 4 possible ideas. Since each of these can be available or unavailable, there are $2^4$ different sets $D$ associated with every $B$. If we increase the number of elements to 4, then $B$ has 12 dimensions and there are $2^{11}$ possible sets of $D$ for each vector of beliefs. If we increase the number of elements to 5, then $B$ has 20 dimensions and there are $2^{26}$ different sets $D$ associated with each one. Obtaining a good approximation of a state space with 20 dimensions and $2^{26}$ discrete action states is very challenging.

Given the foregoing, my approach is to instead solve a manageable (small) version of the problem 100 times, and then to use a regression analysis to see if characteristics of Remarks 1-8 hold for these optimal solutions. Essentially, I will project a linear approximation onto a highly complex and non-linear solution, to check for the validity of Remarks 1-8 outside of the special cases for which

\textsuperscript{14} Powell (2011) provides a good overview of the problem and possible approximation methods.
they were derived. If they remain valid, then I can interpret similar regressions on actual patent data with more confidence. It will be important that the choice of problem to solve is sufficiently rich that it captures the complexity of the model, but remains solvable. The basic structure of the problem I will solve takes the following form:

1. There are four elements $q \in Q$ that can be combined into ideas.
2. There are ten eligible ideas. Six ideas are composed of pairs of elements, and four ideas are composed of three elements.\(^{15}\)
3. The cost of ideas is normalized to 1.
4. The researcher’s discount factor is 0.95.
5. The researcher’s beliefs about the affinity of each edge follow a beta distribution. Each edge has unique beta parameters. I discuss this more below.
6. Each idea has a unique prize value $\pi(d)$, known to the researcher. I discuss this more below.

These six conditions capture several key features of my model. First, I model the beliefs of the researcher (as in Special Case 2), while allowing ideas to depend on multiple edges and to have different reward values (as in Special Case 1). Second, each edge can be used up to three times (once in a two-element idea, and twice in a three-element idea), so that updating of beliefs can happen more than once. Third, information revealed from one idea spills over to other ideas. Fourth, some ideas are subsets of others. Finally, I do not impose symmetries in beliefs or prize values.

Most importantly, this problem can be solved in a reasonable amount of time. I have written a computer program in python to solve this problem. To begin, I define the possible sets $D$ in which the researcher may find himself. Since there are 10 ideas, and each idea may be either eligible or ineligible, there are $2^{10} = 1,024$ distinct sets of eligible ideas. For each of these sets, I define the potential belief vectors of interest. Since we know the initial beta parameters of every $a(x)$, the program can exhaustively list the vectors $B$ that might be attained in any given set.

Once I have a set of $(D, B)$ states, I work backwards. The program begins by evaluating the null set $D = \{\emptyset\}$, where the only available option is to quit research and earn zero with certainty (for every belief vector $B$). Next, using this result, it evaluates the best action for each set with just one eligible idea remaining. When there is just one eligible idea, the problem simplifies to:

$$V(D, B) = \max\{E[e(d)]\pi(d) - k(d), 0\}$$  \hspace{1cm} (30)

Using these results, the program evaluates the best action for each set with two eligible ideas remaining, which has the form of equation (28). At each stage, it uses the researcher’s beliefs and the

\(^{15}\) Including the 11th idea, composed of all four elements, dramatically increases the computational time to solve, without adding much insight, so I omit it. Alternatively, we might assume this idea has prohibitively high costs, but is theoretically eligible.
revelation procedure to compute the probabilities associated with each state the researcher may find herself in next period.

After working backwards, the program obtains a mapping from every \((D,B)\) state to a best action. With four elements and ten ideas, this program still takes approximately an hour to solve depending on computer processing power. See the appendix for a more detailed description of this program.

6.3 – 100 Variations

I solve 100 variations of the above problem which differ in researcher beliefs and rewards \(\pi(d)\). To obtain the researcher’s beliefs about a given edge, I derive \(\alpha_i\) and \(\beta_i\) from the following equations, where \(\tilde{a}\) is drawn from a beta distribution with \(\alpha = 6\) and \(\beta = 2\):

\[
\tilde{a} = \frac{\alpha_i}{\alpha_i + \beta_i} \\
0.4 = \alpha_i + \beta_i
\]

(31)

I chose to draw the initial expected affinity from a beta distribution with \(\alpha = 6\) and \(\beta = 2\) because such a distribution has an expected value of 0.75, but all values above 0.4 occur with relative frequency, as can be seen in Figure 2. This yields a wide array of initial beliefs about edges.

Figure 2: PDF of Beta Distribution with \(\alpha = 6\) and \(\beta = 2\)

I calibrated \(\alpha_i + \beta_i = 0.4\) because at this level of certainty, the inherent difficulty of creating more complex ideas can be overcome by learning, at least for a typical case. At this level of certainty,
\[ E[\bar{a}] = 0.75 \], so that \( \alpha = 0.3 \) and \( \beta = 0.1 \). Initially, a 3-element idea with \( E[a(x)] = 0.75 \) for each edge has \( E[e(d)] = 0.75^3 \approx 0.42 \). Thus, 3-element ideas are initially less likely to succeed than 2-element ideas. However, if the researcher observes one compatibility on each edge, each edge’s beta parameters are increased to \( \alpha = 1.3 \) and \( \beta = 0.1 \), so that \( E[a(x)] = 1.3 / 1.4 \approx 0.93 \). This increases the expected efficacy of the idea to \( E[e(d)] \approx 0.93^3 \approx 0.8 \), so that such ideas are more likely to succeed than a 2-element idea with \( E[a(x)] = 0.75 \).

This means a researcher investigating pairs of ideas can learn enough to make a 3-element idea just as attractive as a 2-element idea with no information, in the typical case. If the certainty was much higher, learning would not convey much information. Nevertheless, variation in the initial draws of \( \bar{a} \) means we will observe many cases where it is not possible to learn enough to make the efficacy of a three element idea higher than a two-element one (with no information).

Next, I select the value of prizes. There is evidence, both theoretical\(^{16}\) and empirical\(^{17}\) that important traits about ideas, such as their value, are Pareto distributed. Other studies, however, indicate the distribution of the value of ideas, while fat-tailed and highly skewed, is not Pareto.\(^{18}\) To address both possibilities, I use two distributions for \( \pi(d) \) with the same mean and variance, but where one is a Pareto distribution and the other a log-normal distribution. In practice, I find neither has a meaningful impact on the optimal strategy.

For half the cases, I draw prizes from a Pareto distribution with:

\[
\begin{align*}
\alpha_{\min} &= 1 \\
\alpha &= 2.41
\end{align*}
\] (32)

These values imply the median prize has value approximately equal to \( 4/3 \), so that, on average, half of the two-element ideas will satisfy the condition \( E[e(d)] \pi(d) > 1 \) and therefore be myopically rational to attempt (since the average two-element idea will have \( E[e(d)] = 3/4 \)).

For the other half of the cases, I use a log-normal distribution tuned to have the same median and variance as the Pareto distribution (so that it is primarily the behavior of the tails that differs between the distributions). This implies the log of this distribution follows a normal distribution with \( \mu = 0.288 \) and \( \sigma^2 = 0.633 \). Both distributions are plotted in Figure 3.

To illustrate features common across many variations, I simulate many actual decisions by a researcher, using the policies that emerge from the solutions to the above problems. To simulate the researcher’s problem, I use the researcher’s initial beliefs in each model to draw “true” affinities.

\(^{16}\) Jones (2005) and Kortum (1997) both show many stylized facts about aggregate R&D and growth can emerge if traits of ideas are Pareto distributed.

\(^{17}\) See Jones (2005), pg. 533-535 for discussion of this evidence.

\(^{18}\) See Scotchmer (2004), pg. 275-282 for a discussion.
With the true affinities, I generate the edge compatibilities and efficacies of each idea, as well as the information revealed to the researcher if that idea is attempted. I then have the researcher follow her strategy, observing her choice in each stage. She makes a choice, observes new information, updates her beliefs, and then follows her optimal strategy in the new information and action state. I will perform 1,000 such simulations for each model, and in each simulation the researcher makes 10 choices over 10 periods (after the tenth period she always chooses to quit research).

**Figure 3:** $\pi(d)$ probability density functions

![Probability Density Functions](image)

*Note: The two pdfs cross again around $\pi(d) = 29$.*

### 7 – Numerical Analysis of the General Case

So far, I have made a number of remarks about the characteristics of optimal innovation behavior under simplified settings. In one instance, the learning aspect of the model was suppressed, and in the other, combinatorial features of the model were suppressed. In the general setting, optimal innovation behavior has the characteristic of each.

#### 7.1 – Optimal Strategy: Probability A Pair of Elements Are Combined

We first consider the probability a researcher will attempt to combine two elements as part of a new idea (either a 2-element or 3-element idea). My approach is to model the probability an edge is used in any given decision as a function of characteristics of the edge. In this way, we obtain a profile of the features held by the numerically-derived optimal strategies. Specifically, I run a probit regression with the following form over all the simulated researcher decisions:

$$\Pr(u_{x,t} = 1) = \Phi(\beta_0 + X_{x,t}'\beta)$$

(33)
where \( u_{x,t} \) is a dummy variable equal to 1 if edge \( x \) is used as part of the idea attempted in period \( t \), when the researcher follows an optimal strategy. Each observation corresponds to one edge in one (simulated) researcher’s decision. I omit edges that do not belong to an eligible idea because the probability that \( u_{x,t} = 1 \) falls to zero in this case. The explanatory variables are traits for each edge, where I choose what traits to include based on the analysis of Remarks 1-8.

Turning first to Section 3, I showed the optimal strategy when affinities are known is straightforward: always choose the idea with the highest expected net value, so long as it exceeds 0 (Remark 1). Therefore, whether or not an edge belongs to the idea with the highest expected net value is a key explanatory variable. I capture this with the dummy variable \( Myopic_{x,t} \), which is equal to 1 if edge \( x \in d^* \) and \( \arg \max_{d \in D} \{ \pi(d)E[e(d)] - k(d) \} = d^* \) in period \( t \).

Of course, when affinities are not known, the optimal strategy in Section 3 is no longer appropriate. Sometimes an idea is selected that is not the myopic best choice, but which provides useful information for future periods. However, when choosing between rival ideas that will provide equally useful information in the future, researchers will still prefer ideas with higher expected net value in the current period, using up the most highly valued ideas first. To account for the fact that the best ideas associated with an edge are fished out over time, I create two variables, \( Attempts_{x,t} \) and \( Pos.value_{x,t} \). The variable \( Attempts_{x,t} \) counts the number of times an idea with edge \( x \) has been attempted up to period \( t - 1 \). As \( Attempts_{x,t} \) rises, there ought to be fewer good ideas left that contain edge \( x \). The variable \( Pos.value_{x,t} \), conversely, counts the number of remaining eligible ideas that contain edge \( x \) and also satisfy \( \pi(d)E[e(d)] - k(d) > 0 \). The intuition here is that, when the researcher deviates from a myopic strategy, she still wants to minimize her losses. One way to do this is to choose ideas that do not have the highest net expected value, but which still have high net expected value. Edges that belong to many eligible ideas that will be eventually attempted under a pure fishing out strategy are more likely to have high, if not the highest net expected value.

The optimal strategy for the special case in Section 5 is less straightforward than in the Section 3 case. Researchers adopt a stick-with-the-winner strategy (Remark 7), which I capture by counting the number of times the edge has been observed compatible by the researcher up to period \( t \). I denote this variable \( Compatible_{x,t} \). In Section 5, I also showed that researchers prefer edges with greater uncertainty (Remark 8). I also capture the degree of certainty about an edge with the variable \( Attempts_{x,t} \) since researchers obtain better information about an edge (usually) when more ideas with it have been attempted.

Unfortunately, a Gittins index strategy does not work in the general setting. Firstly, all ideas are not equally valued, so that the payoff from any one strand of research is declining over time (because of fishing out effects), unless this is offset by the researcher’s continual upward assessment of each
remaining idea’s efficacy. Secondly, in the general setting, all ideas with more than two elements depend on multiple edges. Knowledge that an edge has a high affinity is useless if all other edges have low affinity, since the researcher can’t combine the edge with any others to generate effective ideas. Conversely, if an edge is embedded in a network of many of other edges with high affinity, learning it too has high affinity is very rewarding, since it can be combined with many other edges. To measure this effect, I again use the variable \( Pos.value_{x,t} \). Intuitively, any idea with \( \pi(d)E[e(d)] - k(d) > 0 \) will be tried eventually, and so learning about edges contained in the idea will have an impact on the value of research. Edges with a high value of \( Pos.value_{x,t} \) belong to many ideas that would benefit from learning the idea is effective.

The final regression takes the form:

\[
Pr(u_{x,t} = 1) = \Phi(\beta_0 + \beta_1 Myopic_{x,t} + \beta_2 Attempts_{x,t} + \beta_3 Pos.value_{x,t} + \beta_4 Compatible_{x,t})
\]  

(34)

I anticipate the optimal strategy is characterized by the following:

**Conjecture 1: Probability of Pairwise Combination.** When the probability a researcher will optimally combine two elements as part of a research project is modeled by (34), then \( \beta_1, \beta_3, \beta_4 > 0 \) and \( \beta_2 < 0 \).

This is indeed the case, as Table 1 indicates.

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Constant</th>
<th>Myopic_{x,t}</th>
<th>Attempts_{x,t}</th>
<th>Pos.value_{x,t}</th>
<th>Compatible_{x,t}</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-2.096</td>
<td>2.765</td>
<td>-0.459</td>
<td>0.151</td>
<td>0.428</td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td>(0.003)</td>
<td>(0.008)</td>
<td>(0.002)</td>
<td>(0.008)</td>
</tr>
</tbody>
</table>

Observations: 3,160,024  
Psuedo R\(^2\): 0.626  
Akaike Inf. Crit.: 1,183,283

*Note: Standard errors are reported in parentheses.*

All the signs are in the anticipated direction, and all parameters are significantly different from zero.

Note that \( Pos.val \geq 1 \) whenever \( Myopic = 1 \) (otherwise quitting research would be the myopic best choice), that \( Pos.val \leq 3 - Attempts \) (because the maximum number of eligible ideas associated with an edge is 3), and \( Compatibility \leq Attempts \) (since we can only observe an edge is compatible by attempting an idea containing it). With these restriction, and because these variables are discrete over a small range, we can exhaustively list every feasible combination of edge traits, as well as the probability of selection in Table 2.
Clearly choosing the idea with the highest net expected value is usually the preferred strategy, with the probability of selecting an edge that is the myopic best choice typically on the order of 60-85%. Even when the researcher has twice tried the idea, and never observed a compatibility, such an edge is played 46.1% of the time. Note also that the probability of play is declining in Attempts, although this can be almost perfectly offset by observing compatibilities. For instance, the probability of choosing an edge with $\text{Pos.val} = 1$, $\text{Myopic} = 0$, and $\text{Compatible} = 0$ falls from 2.6% to 0.2% as Attempts rises from 0 to 2, but that it only drops to 2.2% if each attempt reveals the edge to be compatible.

In the general case, the optimal strategy is a mix between the two strategies discussed in Section 3 and 5. In this model with just four elements, fishing out effects are very strong – at most, any edge only belongs to 3 eligible ideas. Nonetheless, researchers are more likely to return to a pair of elements that has been compatible in the past and they favor learning about edges that are connected to other good ideas.

These results suggest a natural explanation for how knowledge accumulation effects and fishing out effects can coexist. Within any given set of primitive knowledge elements, the set of ideas that can be created is finite, and if knowledge is perfect, fishing out effects dominate. This is one reason why the coefficient on the number of ideas attempted using a pair of elements is negative. However,
since knowledge is generally not perfect – especially at the outset – knowledge accumulation effects kick in, since learning that elements are compatible tends to expand the set of ideas that can be profitably attempted. This is why the coefficient on the number of compatible observations is positive. Thus, if the set of elements is fixed, knowledge accumulation effects can increase the value of R&D, but only up to an upper bound, given by the perfect knowledge setting. Thereafter, fishing out effects dominate. The only way out of this long-run trap is to expand the set of primitive elements, something this paper does not address (but see Weitzman 1998 for an optimistic take).

Knowledge accumulation in this model is also related to the concept of technological path dependence. Technological path dependence refers to the idea that certain strands of technology obtain market dominance and hence the attention of future innovators. For example, Acemoglu et al. (2012) present a model where two kinds of technology – carbon neutral and carbon emitting – are substitutes, and in the absence of government policy innovators devote most of their attention to whichever technology has greater market share. Through this mechanism, the transition costs from a carbon emitting to carbon neutral production scheme rise over time, as carbon emitting technology improves at a faster rate than carbon neutral. My model exhibits a similar feature purely through the production function. Since combinations that have worked in the past are more likely to work in the future, researchers optimally base subsequent research on these combinations, rather than searching for alternatives.

The above model also provides a clear mechanism for how knowledge spillovers might happen. As we have noted above, if the researcher observes some edge $x$ is compatible, this increases the expected efficacy of all other ideas that also include edge $x$. In this way, positive developments in one idea can spill over to related ideas. There is also a second channel of knowledge spillover though. The probability that edge $x$ forms part of a research project is positively related to the number of ideas containing edge $x$ with $E[e(d)]\pi(d) - k(d) \geq 0$ (captured by the explanatory variable $\text{Pos.value}$). Suppose the researcher observes edge $x'$ is also compatible. This raises the expected efficacy of all ideas that contain edge $x'$, including some ideas that also contain edge $x$. If the expected efficacy of such an idea rises by a sufficient amount, it may flip the expected net value of the idea from negative to positive. This increases the probability research projects containing edge $x$ will be attempted. For example, suppose researchers invent a new kind of turbine for power plants. It is known that such a turbine can be reconfigured into a jet engine. This may stimulate research on projects related to jets, but not to turbines at all. In this way, entire technological paradigms can be locked in, since the rise in $\text{Pos.value}$ can be temporarily self-reinforcing. Greater knowledge about the affinity of combinations in a subset of elements raises the value of $\text{Pos.value}$ for edges in the subset, which in turn increases the probability of research that increases knowledge about these same edges.
7.2 – Value of Research Dynamics

In Section 3 and 5, the value of research is higher when affinity is higher (Remarks 3, 4, and 6). At the same time, in the certainty setting, the expected value of research declines over time (Remark 2). To check these assumptions, I model the trajectory of $E[V(D,B)]$ over time follows:

$$\ln(E[V(D_t,B_t)]) = \beta_1 + \beta_2 \cdot \ln(Average.affinity_t) + \beta_3 t + \epsilon_t$$  \hspace{1cm} (35)

Where $t$ refers to the period and $Average.affinity_t = \frac{1}{6} \sum_x E[a_t(x)]$ is the mean value of affinity in period $t$. Each observation corresponds to the decision made by a simulated researcher following an optimal strategy. At each decision point, we observe $E[V(D,B)]$ for a researcher in this state, as well as the average affinity of all six edges, at that point.

**Conjecture 2: Research Value over Time.** When the value of research is modeled by equation (35), then $\beta_2 > 0$ and $\beta_3 < 0$.

As indicated in Table 3, this is indeed the case.

<table>
<thead>
<tr>
<th>Table 3: Simulated Research Value Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependent variable:</td>
</tr>
<tr>
<td>($t$)</td>
</tr>
<tr>
<td>ln($Average.affinity_t$)</td>
</tr>
<tr>
<td>Fixed Effects?</td>
</tr>
<tr>
<td>Observations</td>
</tr>
<tr>
<td>$R^2$</td>
</tr>
</tbody>
</table>

In the second model in Table 3, I include fixed effects for all 100 variations, but these have little impact. The general trend is clear: the value of research is declining over time, as ideas are fished out, but this effect can be offset by favorable technological opportunity (represented here by the average expected affinity between the elements available for combination).
This dynamic gives us a potentially bumpy path for the value of research. New discoveries that raise \textit{Average.affinity} from time to time are necessary to prevent the value of research from being dominated by fishing out effects. In this model, however, such effects cannot last long. Unless researchers are revising their beliefs upwards, fishing out effects will cause the value of research to decline. Moreover, the gain to research productivity from revising beliefs upwards plateaus over time (Remark 8), implying a long-term decline in the value of research for any fixed set of elements. Of course, all this is predicated on the maintained assumption that there is a finite set of primitive knowledge elements.

The above results can be interpreted in terms of radical versus incremental innovation, or in terms of general purpose technologies. The distinction between innovation that generates radically new types of processes and products, and innovation that makes improvements to existing technologies while leaving the basic framework unchanged has roots in economic history. Examples of radical innovation might include the steam engine, electricity, and the computer, while examples of incremental innovation might be a new model of a car or smart phone. The importance of radical innovation for establishing a platform for subsequent improvement is also emphasized in the general purpose technology literature (see Helpman 1998).

In terms of this model, we may identify radical innovations as those which are composed of very novel combinations of elements. Such a combination would be characterized by edges with a high degree of uncertainty about their true affinity, since they have rarely been tried before. In contrast, we might identify incremental innovations as those which are composed of relatively common combinations of elements. Such combinations are characterized by edges with a high degree of certainty about their true affinity, since they have been frequently tried in the past and there is a lot of data to estimate their affinity.

When radical ideas turn out to be effective, the researcher’s beliefs are significantly impacted, and the expected affinity of each edge rises by a comparatively large degree (since beliefs are most responsive to new information when they are characterized by a lot of uncertainty). This has the effect of raising \textit{Average.affinity} by a larger amount than would an incremental innovation with the same number of edges. In this model, radical innovations have a large and positive impact on $E[V(D,B)]$ when successful, precisely because they provide a new platform for subsequent innovation (which will be comparatively incremental). Radical innovation provides one way to temporarily reverse the declining value of research that fishing out effects imply.

At the same time, it does not follow that a dearth of radical innovation signals a poor outlook for innovation. An assortment of recent works advocate a form of “technological pessimism” (Cowan 2011). This notion is generally based on a raft of arguments, including an intuitive appeal to the reader that technology hasn’t lived up to our dreams. For example, Gordon (2012) asks his readers to consider the impact on their life of losing either (1) all innovation since 2002 or (2) running water.

---

and indoor toilets, to demonstrate the paucity of recent innovation. Meanwhile, Thiel (2011) points to the broken promises of technology in speed of transport, alternative energy, and cancer treatment and Stephenson (2009) asks “Where’s my donut-shaped space station? Where’s my ticket to Mars?” These complaints can be read as frustration with the current incremental state of technological advance, as against the radical innovations that have occurred in the past or which were expected soon.

However, radical innovation, because they rely on untried combinations, are typically riskier and do not benefit much from knowledge accumulation effects. They are always out there, as a backstop research agenda, when other avenues are fished out. Innovations that are very different from the kind around them may be more memorable than another iteration on an existing paradigm, and they may herald promising new avenues of research. But it may be a mistake to complain that there is not more interest in attempting radical innovation. When researchers are disproportionately trying for radical innovations, this is a signal that technological opportunity is low.

To summarize, researchers overcome the fishing out effect, temporarily, by directing their efforts towards ideas known to succeed. This “stick with the winners” approach is always exhausted in the long run, so that researchers branch out and experiment with combinations that are less studied and which are characterized by uncertainty. If one of these experiments pans out, the research community enjoys another temporary boom. Some ideas flip from negative to positive net expected value as $E[a(x)]$ and $E[e(d)]$ are revised up, which raises $Pos.value$. By inducing research on related edges, the benefits of success can spill over past the initial research project, further sustaining the boom. If none of the experiments pan out though, researchers eventually give up on research and consider the useful ideas which can be pulled from the set of elements exhausted.

We now turn to some empirical applications of this model.

8 – An Application to Patent Data

The model presented in this paper is very stylized. There is one researcher, so that strategic interaction and knowledge spillovers between people and firms do not exist. I assume researchers start with an endowment of knowledge about what elements are available for combination, as well as a set of beliefs, whereas, in reality, a researcher’s knowledge stock is determined endogenously. Moreover, the supply of researchers is itself responsive to the opportunities outside of research. I have modeled the supply of elements to be combined as fixed rather than allowing it to grow through Weitzman (1998) channels. Finally, I limit researcher to pursue one idea per period, when firms may try several research project simultaneously.

Nevertheless, in this section I will assume the underlying model presented in this paper is sufficiently representative of actual research conduct so that its effects can be measured in the real world.
8.1 – Patents as Effective Ideas

To test my model with data, the first step is to determine how to measure ideas. There are many plausible candidates, but I use patents granted by the US Patent and Trademark Office (USPTO). The USPTO provides a useful filter, since patents are only granted if they pass the inspection of a patent examiner. To do this, the innovation must be novel relative to the prior art and nonobvious to someone with ordinary skills in the field, suggesting that each granted patent is indeed a new idea. Furthermore, the innovation must be useful, in the sense that it solves some problem. I identify this attribute with an idea being effective.\(^\text{20}\) I therefore identify each patent with the successful realization of an effective idea.

To be sure, patents are not a perfect measure of ideas. Not all ideas are patented, or even patentable. Abstract ideas, for instance, cannot be patented. Moreover, ideas that are patentable may not be patented, even if they are novel, nonobvious, and useful, because patenting is not costless and requires divulging the details of the innovation.\(^\text{21}\) Furthermore, patents can also be inappropriately granted if the patent office is too lenient.\(^\text{22}\)

Another disadvantage of patent data is the unavailability of related metrics such as the value and costs of different ideas. Worse, patents provide no record of failed and ineffective ideas. A variety of methods have been used to infer the value of patents,\(^\text{23}\) generally finding the value of ideas is heavily skewed. However, these methods, which often rely on citations, or firm level data, are not applicable to my current data set. Obtaining the research costs of each patent would be even more difficult. While it is possible to match R&D spending at the firm level to the patent portfolio of a firm, it is difficult to allocate this spending among patents. Moreover, because we do not observe ideas that were attempted but not patented, we cannot even be sure what share of aggregate R&D spending should be assigned to the aggregate patent portfolio.

These problems, however, are not unique to patents, but rather characteristics of many efforts to measure innovation at large scales. Given that most measures of innovation are problematic, patents at least have several major positive features, among which is coverage. Patents provide an extensive source of microdata on individual innovations across a huge array of industries and years. For example, I have data on millions of patents, across hundreds of sectors, issued over close to two centuries. Despite their shortcomings, patents remain one of the best sources of innovation data we have, and, as we will see, we can still learn a lot from them.

My data is the full set of US utility patents granted between 1836 and 2012. There were 8.3 million patents granted over this time frame. The year 1836 marks the beginning of the current patent

\(^{21}\) See Clancy and Moschini (2013).
\(^{22}\) See Bessen and Meurer (2008) for a study related to software patents.
numbering system, so that my dataset includes patent #1. I have plotted the number of patent granted in each year in Figure 4. The number of patents issued per year has grown over time, averaging a growth rate of 2.7% per annum over the last 100 years.

8.2 – Technology Mainlines as Elements

The second major decision that must be confronted in order to test the model with data is what to use as a proxy for the elements that are combined to yield a new idea. There are a number of candidates, such as the words used in patent documents or the citations to other patents, but I choose to use the technology classifications assigned to each patent. This has the advantage of being available for all US utility patents since 1836.

Figure 4: Patent Granted Per Year, 1836-2012

The USPTO has developed the US Patent Classification System (USPCS) to organize patent and other technical documents by common subject matter. Subject matter can be divided into a major component called a class, and a minor component, called a subclass. The USPTO states “A class generally delineates one technology from another. Subclasses delineate processes, structural features, and functional features of the subject matter encompassed within the scope of a class.” These classifications are meant to be exhaustive and non-overlapping, and are therefore a natural candidate for elements of combination.

---

24 Prior to the 1836 numbering scheme, an additional 9,957 patents were issued, which are not in my dataset. See US Patent and Trademark Office (2014a).
There are more than 450 classes and more than 150,000 subclasses in the USPCS. To take an example, class 014 corresponds to “bridges,” class 301 corresponds to “land vehicles (wheels and axles),” and class 706 corresponds to “data processing (artificial intelligence).” A complete list of the current classes can be found on the USPTO website. The subclasses are nested within each class, and correspond to more fine-grained technological characteristics. For example, subclass 014/8 corresponds to “bridge; truss; arrangement; cantilever; suspension.” The subclass 301/108.5 corresponds to “land vehicles (wheels and axles); wheel; hub; hub cap; retained by threaded means; central-threaded means.” And the subclass 706/29 corresponds to “data processing (artificial intelligence); neural network; structure; architecture; lattice.”

Subclasses are nested and hierarchical. The uppermost subclass is called a mainline subclass, hereafter simply “mainline.” For example, the subclasses “bridge; truss,” and “land vehicles (wheels and axles); wheel,” and “data processing (artificial intelligence); neural network,” are all mainlines. The subclass nested one level down is said to be “one indent” in from the mainline. For example, the subclass “bridge; truss; arrangement” is one indent in from the mainline “bridge; truss,” and the subclass “land vehicles (wheels and axles); wheel; hub” is one indent in from the mainline “land vehicles (wheels and axles); wheel.” Within these one indent subclasses will be still further subclasses, called two indent subclasses, and so on.

A classification is assigned to a patent by the patent office with the following methodology. The examiner has some portion of the patent, called the subject matter, he would like to assign to a subclass. Scanning through the list of mainlines in a class, the examiner stops when he finds a mainline that corresponds to the subject matter. The examiner then scans through the list of subclasses one indent in from the mainline. If none of the one indent subclasses apply to the subject matter, the examiner assigns the mainline to the patent. If one of the subclasses does apply, then the examiner repeats this process for the two indent subclasses that lie within the one indent subclass. The examiner then repeats the above process for three indent subclasses and so forth, until he arrives at a point where no deeper subclasses apply to the subject matter. At this point, the highest indent subclass (which will correspond to the most specific and narrow definition) found to be applicable is assigned to the patent. The USPCS is continually updated to reflect new technological categories, and patent classifications are updated as part of this process.

For every patent in my dataset, I observe both the year it was granted and the technology subclasses to which it is assigned. However, simply using the technology subclasses as elements to be combined is problematic for a few reasons. First, the categories may not correspond to the same level of specificity, since they are nested. For example, consider three subclasses, that all belong to class 706, “data processing (artificial intelligence).”

- 706/29 – Data processing (artificial intelligence); neural network; structure; architecture; lattice.

---

26 http://www.uspto.gov/web/patents/classification/selectnumwithtitle.htm
• 706/15 – Data processing (artificial intelligence); neural network.
• 706/45 – Data processing (artificial intelligence); knowledge processing system.

Classes 706/29 and 706/15 are both associated with neural networks, but at different levels of specificity, while 706/45 is not associated with neural networks at all. Without looking at the USPC index, we would not know there is a relationship between some of the subclasses, but not others. The second major problem is simply that 150,000 technological subcategories is computationally too difficult to manage.

Instead, I use technology mainlines as my primary elements of combination. This identifies a set comprising approximately 17,000 elements. Of these, approximately 13,000 are assigned to utility patents in my dataset (from here on, I restrict attention to the set of mainlines actually assigned). These mainlines are meant to be exhaustive and nonoverlapping. The mean number of mainlines used per class is 29.6, with a median of 21. However, the number per class varies widely. The maximum is 246 mainlines in one class, while 18 classes have just 1 mainline each. If each mainline is assumed to proxy for an element of combination, we would like each mainline to cover roughly the same scope of technologies.

It might then be objected that, because classes vary so much in how finely they divide their technologies, mainlines are a poor choice of proxy. However, we must remember that the technology classification system itself differs in how many technologies are encompassed in one class. For instance, the class 002, which corresponds to “apparel,” has 33 mainlines in it. Class 004, which corresponds to the group “baths, closets, sinks, and spittoons,” has 70. Because class 004 appears to tie together a more disparate set of technologies, more mainlines does not necessarily indicate that a mainline from class 004 covers a wider set of technologies than a mainline from class 002. Moreover, in some cases, the reverse happens and one type of technology is split into many classes. For example, classes 532-570 all correspond to organic compounds, and classes 520-528 all correspond to synthetic resins or natural rubbers. Of the 18 classes with one mainline each, 13 belong to one of these two series. In these cases, classes already divide up the space of technologies very finely, so that additional division into many mainlines is not necessary. Because defining the scope of what constitutes an element across different technologies is bound to be somewhat arbitrary, using mainlines as a proxy seems to me an appropriate first step.

8.3 – Assigning Each Patent A Combination of Mainlines

The USPTO makes available a large text document listing the technology subclass assigned to each patent, under the most recent classification scheme. Each line of this text document contains a patent number, a subclass code, and an indicator for whether the subclass is the primary subclass (discussed more in Section 10). By reading this document line by line, I extract the subclasses assigned to each patent, as well as the identity of the primary subclass. I can then use the patent

---

number to infer the year of the patent’s grant.\textsuperscript{29} For the reasons discussed above, I next collapse each technology subclass down to the mainline to which it belongs. For example, US patent 7,640,683 is titled “Method and apparatus for satellite positioning of earth-moving equipment” and describes a method of attaching antenna to the arm of an earthmoving machine in such a way that using satellite positioning systems is possible. This patent was assigned to four technological subcategories:

1. 37/348 – Excavating; ditcher; condition-responsive.
2. 414/699 – Material or article handling; vertically swinging load support; shovel or fork type; tilting; control means responsive to sensed condition.
3. 701/50 – Data processing; vehicles, navigation and relative location; vehicle control guidance, operation, or indication; construction or agricultural vehicle type.
4. 37/382 – Excavating; road grader-type; condition responsive.

Using the USPC index\textsuperscript{30} I coded a program to reassign each subclass to its associated mainline. Applying this program to the above patent, I reclassify it as consisting of the following elements:

1. 37/347: Excavating; ditcher
2. 414/680: Material or article handling; vertically swinging load support
3. 701/1: Data processing; vehicles, navigation and relative location
4. 37/381: Excavating; road grader-type

Ideally, we would like every patent to be comprised of two or more mainlines, because the model I developed is premised on ideas as sets of at least two elements. In practice, after collapsing all patent assignments to mainlines, only 62.5\% of patents are assigned more than one mainline. This share varies over time, averaging 40\% over the period 1836-1935 and 69\% between 1936 and 2012. A potential explanation is that new classification schemes consolidate commonly used pairs of mainlines into a single technology subclass. If this is the case, then we do not observe as many combinations of elements for older technologies, because combinations frequently used are reclassified as a single technology.

Over the entire period, the mean number of mainlines per patent is 2.29. The share of patents in a given year assigned more than one mainline, as well as the mean number of mainlines per patent, are each plotted in Figure 5. Over time, the number of elements in a patent has grown. While this may be the consequence of consolidation of subclasses, as discussed above, an alternative interpretation is that the complexity of patents has risen over time. The behavior of complexity in the context of my model is the subject of another research project.

Out of approximately $13,000 \cdot 13,000 / 2 = 84.5$ million possible mainline-pair combinations, 1.98 million pairs are actually assigned to at least one patent over the period 1836-2012. Put another way,

\textsuperscript{29} This can be inferred from US Patent and Trademark Office (2014a).
of the 84.5 million possible edges, we observe about 2% of the edges as belonging to effective ideas. The mean number of patents each edge belongs to over the entire period is 10.9, but the distribution is highly skewed. Some 50.6% of observed edges are only ever assigned to one patent, but the maximum patents assigned to an edge is 22,113.

To summarize, my dataset comprises 8.3 million patents granted between 1836 and 2012. Each of these patents is represented as a combination of mainlines, with 62.5% of patents being assigned more than one mainline. We now turn to the first test of the model.

**Figure 5: Mainlines per Patent**

9 – Probability An Edge Is Used

In section 7.1, I showed that the probability a researcher uses a given edge as part of a research project can be modeled by a probit regression, as in equation (34). The analogous model, suitable to the patent data at hand, is:

\[
Pr(\tilde{u}_{x,t} = 1) = \Phi(\beta_0 + X'_{x,t-1}\beta)
\]

where

\[
X'_{x,t-1}\beta = \beta_1 \cdot Related_{x,t-1} + \beta_2 \cdot Age_{x,t-1} + \sum_{i=2}^{9} \phi_i 1(Comp_{x,t-1} = i) + \phi_{10} 1(Comp_{x,t-1} \geq 10)
\]
Each observation corresponds to data on one edge in a given year. The dependent variable $u_{x,t}$ equals 1 if edge $x$ is used by at least one patent granted in year $t$, and 0 otherwise. As in Section 7.1, I use information on the edge to predict the probability it is used in a given year.

All of the independent variables are lagged by $l$ years. The choice of lag is motivated by two distinct factors. First, we observe the year a patent was granted, rather than the year a researcher decided to conduct research. Data on patents from 1967-2000 from Hall, Jaffe, and Trajtenberg (2001) suggests the typical lag between a patent application and grant is 2 years. If we presume the typical research project takes 1-3 years to complete, then information available 3-5 years before a patent is granted is most relevant.

Suppose there is a $l$-year lag between the decision to conduct research and the realization of a patent grant. The second factor is when information about the outcome of research becomes available to other researchers. We may consider two extremes. First, if research outcomes are only revealed by patent grants, then a researcher with a patent granted in year $t$ made his decision to innovate in period $t-l$, based on patents granted up to year $t-l$. Second, suppose research outcomes are revealed within 1 year through channels outside the patent system. Then a researcher with a patent granted in year $t$ made his decision to innovate in period $t-l$ based on research projects completed up to period $t-l-1$. These projects will be (typically) granted patents by period $t-1$.

Our preferred specification lags the explanatory variables by 3 years, but I will experiment with alternatives as a robustness check.

9.1 – Explanatory Variables

In section 7.1, I showed the variable Compatible was positively associated with use of an edge, where Compatible counted the researcher’s observations of an edge being compatible. The analogous variable in this model is $Comp_{x,t-1}$, a count of how many patents have been assigned edge $x$ in any year up to and including year $t-1$. We do not observe instances where an idea was ineffective but the researcher observed an edge to be compatible. To allow for a flexible non-linear relationship between Comp and the probability an edge is used, I separately estimate one coefficient for each of the first 9 values of Comp, and then a final coefficient for the indicator $Comp \geq 10$.

Our earlier analysis gives reason to believe the relationship between edge use and prior observations of compatibility is increasing but concave. For instance, when affinities are believed to be beta distributed, researchers believe:

$$E[a(x)] = \frac{\alpha + s(x)}{\alpha + \beta + s(x) + f(x)}$$

(38)
where $s(x)$ is the number of observed compatibilities ("success") and $f(x)$ the number of observed incompatibilities ("failures"). When this is the case, $\partial E[a(x)]/\partial s(x) > 0$ and $\partial^2 E[a(x)]/\partial s(x)^2 < 0$. Since the informational content of more observations of compatibility are less valuable, I expect the coefficients $\phi_i$ for $i = 1, \ldots, 9$ to map out an increasing and concave curve.

In section 7.1, I also measured the number of times an edge had been attempted with the variable Attempts, as a way of controlling for fishing out effects and a preference for uncertainty. Since we do not observe research attempts that are not patented, it is impossible to create this variable with the patent data. As an alternative proxy, I use the variable Age which is the number of years that have elapsed between year $t - l$ and the first year the edge was used. I am here assuming that as soon as researchers start trying to combine two elements as part of an idea, subsequent attempts are positively correlated with time. Since this fishes out the best ideas, I expect the coefficient on Age to be negative.

In section 7.1, I also measured the number of eligible ideas with positive net value with the variable Pos.Value. Since we do not observe the value or costs of patents, I cannot construct such a variable from the patent data. As a proxy for Pos.Value, I instead construct the variable Related. The construction of $\text{Related}_{x,t-1}$ is based on the observation that the probability any given idea satisfies $\pi(d)E[e(d)] - k(d) > 0$ is increasing in $E[e(d)]$. Therefore, the number of ideas that use edge $x$ and have "high" values of $E[e(d)]$ is likely to be correlated with the number of ideas that satisfy $\pi(d)E[e(d)] - k(d) > 0$. With 13,000 mainlines in use, the number of possible sets using any one edge is astronomical$^{31}$ and therefore infeasible to compute. However, since the average idea only contains 2.29 mainlines, counting only small ideas should be sufficient to get a reasonable estimate for the number of ideas with positive net value. To construct the variable $\text{Related}_{x,t-1}$, I therefore count all possible combinations of three mainlines that (1) include the edge $x$ and (2) have at least one patent assigned to the other two edges as of time $t - l$. If there are 13,002 mainlines, this variable ranges from 0 to a theoretical maximum of 13,000.

In some specifications of the model, I will also be interested in whether there is a differential effect for older or newer information. Accordingly, I also construct two variables called 5-year $\text{Related}_{x,t-1}$ and 10-year $\text{Related}_{x,t-1}$. These variables are constructed in the same way as $\text{Related}_{x,t-1}$, except with the additional restriction that I only count 3-mainline patents from the 5 or 10 years preceding period $t - l$ with 1 observation of compatibility for each edge.

---

$^{31}$ If there are 13,002 mainlines, then the number of ideas that include any two mainlines is $2^{13,000}$, since after subtracting the two mainlines that must be present, the remaining 13,000 can take on one of two states, “in the idea” or “not in the idea.”
Whereas I would like to include a variable analogous to *Myopic* from Section 7.1, I am unable to construct a good proxy from the available data, given the lack of information on values and costs of ideas, as well as the choice set presented to researchers. I am therefore forced to leave this explanatory variable out.

9.2 – Data

Because computing *Related*$_{x,t-1}$ is computationally intensive, I conduct regression (36) on a random subset of edges. Specifically, I draw 10,000 edges (without replacement) from the set of 1.98 million edges that are ever used. For each edge, I include as observations each year after the edge was first assigned to any patent, for a total of 564,531 edge-year observations. A summary of the data used is presented in Table 4.

**Table 4: Summary of Sampled Data Used in Edge Regression**

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>Median</th>
<th>Mean</th>
<th>Max</th>
<th>Standard Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u_{x,t}$</td>
<td>0</td>
<td>0</td>
<td>0.085</td>
<td>1</td>
<td>0.279</td>
</tr>
<tr>
<td><em>Related</em>$_{x,t}$</td>
<td>0</td>
<td>74</td>
<td>111.5</td>
<td>2,113</td>
<td>124.5</td>
</tr>
<tr>
<td>10-year <em>Related</em>$_{x,t-l}$</td>
<td>0</td>
<td>14</td>
<td>25.93</td>
<td>686</td>
<td>36.15</td>
</tr>
<tr>
<td>5-year <em>Related</em>$_{x,t-l}$</td>
<td>0</td>
<td>7</td>
<td>14.01</td>
<td>491</td>
<td>22.19</td>
</tr>
<tr>
<td><em>Age</em>$_{x,t}$</td>
<td>0</td>
<td>33</td>
<td>39.67</td>
<td>176</td>
<td>31.22</td>
</tr>
<tr>
<td><em>Comp</em>$_{x,t}$</td>
<td>1</td>
<td>1</td>
<td>6.429</td>
<td>4,098</td>
<td>35.08</td>
</tr>
<tr>
<td>$1(<em>Comp</em>$_{x,t}$ = 1)</td>
<td>0</td>
<td>1</td>
<td>0.556</td>
<td>1</td>
<td>0.497</td>
</tr>
<tr>
<td>$1(<em>Comp</em>$_{x,t}$ ≥ 10)</td>
<td>0</td>
<td>1</td>
<td>0.102</td>
<td>1</td>
<td>0.317</td>
</tr>
</tbody>
</table>

Only 8.5% of the edge-year pairs are used in any given year. However, in the typical year, the median edge has 74 related 3-mainline ideas, where each edge has been assigned to a patent at least once, though the distribution is quite skewed, with the variance larger than the mean. The typical age of an edge is 39.67, indicating a total life-span of approximately 80 years for most edges. Finally, as with the full set of edges, the median number of times an edge is assigned to a patent is just once, with a mean of 6.4 (compared to 10.9 for all edges). Only 10% of edges are assigned to 10 or more patents.

9.3 – Results

The results of regression (36) are presented in Table 5, where I have omitted the $\phi_i$ coefficients because of space reasons. The $\phi_i$ coefficients are plotted in Figure 6. All are significant at the 0.1% level, and a clearly concave shape is visible.
All results are in the direction predicted, and are highly significant. This is consistent with an innovation framework where fishing out effects conflict with (declining) knowledge accumulation effects for any one pair of elements, and spillovers from related ideas. In columns (2) and (3), I specify models that include the explanatory variables 5-year $\text{Related}_{x,t-1}$ and 10-year $\text{Related}_{x,t-1}$. In each case, I find a positive impact on the short-term $\text{Related}_{x,t-1}$ variable, while the coefficient on the general $\text{Related}_{x,t-1}$ variable turns negative.

**Figure 6:** Estimated Coefficients on $\text{Comp}_{x,t}$ with 95% confidence interval

One potential explanation for this is that, once I control for the most recent set of related ideas, the older related ideas are picking up a variable like $\text{Attempts}_{x,t}$ from section 7.1. Instead of counting the number of related “good” eligible ideas, $\text{Related}_{x,t-1}$ may be counting the number of good ideas that have been attempted, and have therefore been fished out.

A second potential explanation recognizes that mainlines are only proxies for the elements in the model. Suppose each mainline is more like a basket of elements in the model, and an edge represents the combination of two elements drawn from two different baskets. For example, suppose mainline A consists of the basket of elements $(q_1, q_2)$ and mainline B consists of the basket of elements $(q_3, q_4)$. Whether a patent combines $q_1$ and $q_3$, or $q_1$ and $q_4$, etc., we only observe mainline A combined with mainline B. However, researchers may observe the true elements of combination. If they only care about connections between $q_1$ and $q_3$ then, a researcher may appear to ignore some combinations of mainlines A and B and pay attention to others. If the elements drawn from a mainline are correlated across time, then more recent observations are more likely to be informative.
for researchers. And indeed, the marginal effects of 5-year $\text{Related}_{x,t-1}$ and 10-year $\text{Related}_{x,t-1}$ on the latent variable are an order of magnitude larger than $\text{Related}_{x,t-1}$.

<table>
<thead>
<tr>
<th>Table 5: Probability An Edge Assigned to a Patent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependent Variable:</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>1($\text{Comp}_{x,t-1} \geq 10$)</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>$\text{Related}_{x,t-1}$</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>$\text{Age}_{x,t-1}$</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>$t - l$</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>5-year $\text{Related}_{x,t-1}$</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>10-year $\text{Related}_{x,t-1}$</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Observations</td>
</tr>
<tr>
<td>Log Likelihood</td>
</tr>
<tr>
<td>Akaike Inf. Crit.</td>
</tr>
</tbody>
</table>

Note: Standard errors are in parentheses. All coefficients are statistically significant at $p = 0.1\%$

Finally, in column (4), I also control for a general time trend that is consistent across all edges. Age remains relevant in the presence of a general time trend, and has a larger impact on the latent variable.

It is instructive to consider the magnitude of these effects. I compute the probability an edge is used as a function of explanatory variables in Table 6.
Table 6: Probability Edge is Used

<table>
<thead>
<tr>
<th>$\text{comp}_{x,t}$</th>
<th>$\text{related}_{x,t-l}$</th>
<th>$\text{age}_{x,t-l}$</th>
<th>Probability of Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.055</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>33</td>
<td>0.016</td>
</tr>
<tr>
<td>1</td>
<td>74</td>
<td>0</td>
<td>0.061</td>
</tr>
<tr>
<td>1</td>
<td>74</td>
<td>33</td>
<td>0.019</td>
</tr>
<tr>
<td>1</td>
<td>74</td>
<td>176</td>
<td>0.000</td>
</tr>
<tr>
<td>1</td>
<td>2,113</td>
<td>33</td>
<td>0.239</td>
</tr>
<tr>
<td>2</td>
<td>74</td>
<td>33</td>
<td>0.048</td>
</tr>
<tr>
<td>3</td>
<td>74</td>
<td>33</td>
<td>0.077</td>
</tr>
<tr>
<td>10</td>
<td>74</td>
<td>33</td>
<td>0.563</td>
</tr>
</tbody>
</table>

In the first period after an edge has been observed to be compatible, the probability it will be assigned to a patent in three years is 5.5%, if it has no related good ideas, and 6.1% if it has 74 related ideas (the median amount). These probabilities fall to 1.6% and 1.9% respectively, however, if 33 years (the median age of an edge) elapses without another observed compatibility.

An edge with median values for all three explanatory variables has just a 1.9% chance of being assigned to a patent in three years. If we increase the age to its maximum level, the probability of use falls to essentially zero, while if we increase the number of related “good” ideas to 2,113 (the maximum in the dataset), the probability of use approaches one quarter. If we add a second observation of compatibility, the probability of use more than doubles to 4.8%. A third observation increases the probability of use to 7.7%. Finally, if the edge-year is one of the 10% with more 10 or more observations, then it has a better than 50% chance of being assigned to a patent in three years.

9.4 - Robustness Checks

Appendix A3 reports a series of robustness checks. I test alternative lag structures, and add fixed effects to a logistic regression, but neither has a significant impact on the above results. I also show the explanatory variables discussed in the preceding section can predict the number of patents an edge is assigned to in a given year, conditional on it being assigned to at least one edge.

10 – Patent Prediction

The previous section established that researcher behavior in the real world is broadly consistent with this paper’s model. Researchers are more likely to patent ideas that draw on previously successful
combinations of elements, but this probability increases at a decreasing rate and is decreasing over time. Moreover, the probability of patenting is also positively related to the number of related ideas. In this section, I show that variables whose construction is guided by this model have measurable impacts on aggregate inventive activity.

In section 7.2, I showed that the value of research \( E[V(D, B)] \) can be modeled as a function of \( \text{Average.affinity} \) and time \( t \) according to equation (35). We do not observe the correct variables to directly estimate such an equation with patent data, but I will estimate an analogous regression with proxies.

### 10.1 – Technology Class Panels

In section 7.2, we had panel data on different sets of elements (in the sense of having different affinities and yielding ideas of different value) over time. While we do not have an exact analogue in the patent data, I exploit the technology classification system to obtain a similar panel. If we observed \( E[V(D, B)] \) and \( \text{Average.affinity} \), I could apply equation (35) to different domains of technology. For example, if computer innovations are built from one set of elements and apparel innovations are built from another set, we could model the expected value of computer research or apparel research as a function of the average expected affinity of elements used in each domain, as well as the length of time innovation has proceeded in each domain.

It is not necessary for the domains to draw on different elements. For example, innovation in biofuel and beer production both draw on some of the same elements related to alcohol production, but they are not identical. An increase in the expected affinity of the elements used only in biofuel production might not have an impact on the value of beer research, while an increase in the expected affinity of the elements common to each might increase the value of research in both fields.

In this section, I use 429 technology classes in the USPCS as proxies for different technological domains. This gives us a panel of 429 classes, each with up to 176 years of innovative activity. Data corresponding to class \( cl \) in year \( t \) is subscripted with \( cl, t \).

### 10.2 – Patent Attempts As Dependent Variable

While we do not observe \( E[V(D_i, B_i)] \) in the patent data, the number of patents granted to a class in a year is a reasonable proxy for this variable. In the model, when the researcher decides to quit research, \( E[V(D_i, B_i)] = 0 \). Therefore, anytime \( E[V(D_i, B_i)] > 0 \) the researcher is conducting research on new ideas, rather than quitting research. We can apply similar logic to the real world as well. Suppose we adjust the model so that researcher \( i \) with the knowledge necessary to innovate in class \( cl \) earns a present-discounted value \( v_i \) by using time and resources on non-innovation activities, or \( E[V(D_i, B_i)] \) by conducting research. Whenever \( E[V(D_i, B_i)] > v_i \), the researcher will pursue innovation activities, and otherwise he will pursue his outside option. If \( v_i \) varies across
researcher, then a higher level of $E[V(D_t, B_t)]$ will lead more researchers to choose to conduct research.

Of course such a basic model ignores strategic and equilibrium conditions, but the general argument stands. Indeed, this is just a special case of the general principle that, in a competitive market, resources will flow into activities with higher returns. This should also hold in the model, with a higher expected value of research in class $cl$ leading more firms to enter the class or deploy more resources to R&D. By this logic, it would be desirable to look directly at research inputs such as R&D spending rather than outputs like patent grants, but this will have to await future work, as my dataset is inadequate for this purpose. Fortunately, studies have consistently found patent grants and R&D spending are highly correlated across both industries and firms.\(^{32}\)

To determine the number of patents granted to a class in a year, I use the primary classification assigned to each patent. Each patent is assigned one, and only one, primary classification, which is based on the main inventive concept. The primary classification is generally used in economics to assign patents to different technology classes (see for example, Hall, Jaffe, and Trajtenberg 2001). The dependent variable $n_{d,t}$ corresponds to the number of patents with primary class $cl$ granted in year $t$. The path of three $n_{d,t}$ variables over time is plotted in the left column of Figure 7.

As can be seen in Figure 7, some classes of technology have already hit their peak and since gone into a long-run decline, such as class 178, “telegraphy.” Other classes are very old, but innovations continue, such as class 310, “electrical generator or motor structure.” Finally, some classes are relatively new, such as computer related classes, which only began patenting in the second half of the 20th century. Moreover, the scale of $n_{d,t}$ varies considerably by class, from peaks near 150 in telegraphy to 3,000 for digital memory.

### 10.3 – Explanatory Variables

We would like a measure of the average expected affinity of all the edges in a given class. The first step in constructing such a measure is to determine the edges that are used by a class. To do this, I tally the mainlines assigned to each patent in a class. For example, recall that patent 7,640,683, for “Method and apparatus for satellite positioning of earth-moving equipment” was assigned the mainlines:

5. \(37/347\): Excavating; ditcher
6. \(414/680\): Material or article handling; vertically swinging load support
7. \(701/1\): Data processing; vehicles, navigation and relative location
8. \(37/381\): Excavating; road grader-type

---

\(^{32}\) See Griliches (1990) for a survey. He cites an R-squared between R&D and patents of around 0.9.
**Figure 7:** Patent Class Trends

**Left Column:** Patents Granted per Year  
**Right Column:** Average Affinity Proxies

**Class 178:** Telegraphy

**Class 310:** Electrical Generator or Motor Structure

**Class 711:** Electrical Computers and Digital Processing Systems: Memory
The primary class for this patent is class 37, excavating, and it was granted in year 2010. I therefore assign mainline-pairs between 37/347, 414/680, 701/1, and 37/381 as belonging to class 37 from 2010 onwards. After doing this for every patent in the class, I obtain the list of all mainline-pairs used by any patent in the class, in each year.

The number of mainlines used by a class is different from the number of mainlines that are nested under a class by the USPCS, because most patents assigned to a class draw on mainlines from outside the class (just as the above patent is assigned to class 37 but is also assigned mainlines from classes 414 and 701). The mean number of mainlines nested under one class is 29.6, but the mean number of mainlines used by patents in a class is 1,161. Moreover, the minimum number of mainlines nested under one class was 1, while the minimum used by patents in a class is 5. The maximum number of mainlines nested under a class was 246, while the maximum used by a class is 5,126.

The next step is finding a way to proxy the average expected affinity between edges in a class, when we observe neither failed ideas nor the beliefs of the researchers. I take three alternative approaches towards measuring the expected affinity of an edge.

In all three cases, I assume the expected affinity of an edge which has never been assigned to a patent is small enough that it can be represented as zero. Thereafter, the expected affinity of an edge is rising in the number of times it has been assigned to a patent.

1. \[ \text{Affinity.proxy.1}_{x,t} = \frac{S_t(x)}{1 + S_t(x)} \] where \( S_t(x) = \sum_{\tau=0}^{t} s_\tau(x) \) and \( s_t(x) \) is the number of patents edge \( x \) is assigned to in year \( t \). This formulation is equivalent to the expectation of a beta distributed bernoulli variable, when \( \alpha = \epsilon, \epsilon \) is very close to 0, \( \beta = 1 - \epsilon \), and there are no observed incompatibilities. It can be understood as an upper-bound on the true \( E[a(x)] \), when the researcher’s beliefs are beta with \( \alpha = \epsilon \) and \( \beta = 1 - \epsilon \).

2. \[ \text{Affinity.proxy.2}_{x,t} = \frac{S_t(x)}{1 + N_t(x)} \] where \( N_t(x) = \sum_{\tau=0}^{t} n_\tau(x) \), and

\[ n_t(x) = \max\{s_t(x),(1 + g_t)s_{t-1}(x)\} \] \hspace{1cm} \text{(39)}

and

\[ 1 + g_t = \sum_x s_t(x) / \sum_x s_{t-1}(x) \] \hspace{1cm} \text{(40)}

This formulation attempts to infer the number of times researchers tried to use an edge using a simple rule. In each period, researchers expect the number of attempts to use edge \( x \) is equal to the number of times the edge was assigned to patents in the last period, multiplied
by an aggregate growth term $1 + g_t$, or the number of patents it was actually assigned to this period, whichever is larger. The growth term $1 + g_t$ is the overall growth rate of patent edge assignments. This formulation is also equal to the expectation of a beta distributed bernoulli variable, so long as the rule for inferring research attempts $N_t(x)$ is correct. In practice, this proxy penalizes edges that fail to “keep up” with the aggregate growth rate of all edge assignments, by assuming their failure to keep up reflects a failure of research ideas to be effective, rather than an absence of research attempts.

3. $\text{Affinity.proxy}_{3,t} = \frac{S^{0.95}_t(x)}{1 + S^{0.95}_t(x)}$ where $S^{0.95}_t(x) = \sum_{\tau=0}^t 0.95^{t-\tau} s_\tau(x)$. This formulation is equivalent to Proxy 1, but where more recent assignments are accorded more weight. This formulation has a similar justification as the inclusion of 5-year related$_{x,t-1}$ and 10-year related$_{x,t-1}$ from the previous section. If mainlines are imperfect proxies of elements, then more recent observations may be more relevant. In practice, this proxy builds in a tendency for $E[a(x)]$ to decay over time if an edge is not continuously assigned to new patents.

I compute the above proxies for every pair of mainlines in my dataset, for every year. As a first test of these three proxies, I include them in the probit model from section 8, in place of the cumulative count variables. That is, I estimate the model:

$$\Pr(\mu_{x,t} = 1) = \Phi(\beta_0 + \beta_1 \cdot \text{Affinity.proxy}_{x,t-3} + \beta_2 \cdot \text{Related}_{x,t-3} + \beta_3 \cdot \text{Age}_{x,t-3})$$

(41)

with the sample of year-edge observations used in Section 8. The results of this exercise are presented in Table 7.

In all three cases the proxies are positive and highly significant predictors that an edge will be used. Proxy 2 performs the best in terms of the magnitude of its effect and by AIC criteria. Note, however, that the sign on Related turns negative when I include Proxy 2. It turns out this is only the case for Related, and not when I instead use the variables 5-year Related or 10-year Related (or even 20-year Related). For specifications using these variables instead of Related the coefficient remains positive (and other coefficients do not change significantly). Both Proxy 2 and Proxy 3 perform significantly better than the models with cumulative counts of prior observations, by AIC criteria.

Given my proxies for $E[a(x)]$, and the assignment of mainline-pairs to classes, we can now compute a measure for the average expected affinity for each class, in each year:

$$\text{Average.affinity.proxy}_{c,t} = \frac{1}{N_{c,t}} \sum_{x \in c,t} \text{Affinity.proxy}_{x,t}$$

(42)

We plot the Average.affinity.proxy in the right column of Figure 7, for three example classes. As can be seen, my measure for Average.affinity.proxy is only defined from the point at which the first
patent in the class is granted. Moreover, it exhibits variation over time. Finally, although the three proxies do not move in sync with each other, they are correlated.

Table 7: Testing Proxies for $E[a(x)]$

<table>
<thead>
<tr>
<th>Dependent Variable:</th>
<th>$u_{x,t}$</th>
<th>Proxy 1</th>
<th>Proxy 2</th>
<th>Proxy 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Affinity.proxy_{t-3}$</td>
<td>4.549</td>
<td>4.796</td>
<td>(0.020)</td>
<td>(0.020)</td>
</tr>
<tr>
<td></td>
<td>(0.020)</td>
<td>(0.046)</td>
<td>(0.020)</td>
<td></td>
</tr>
<tr>
<td>$Related_{x,t-3}$</td>
<td>0.001</td>
<td>-0.0001</td>
<td>0.0001</td>
<td>(0.0002)</td>
</tr>
<tr>
<td></td>
<td>(0.00002)</td>
<td>(0.00002)</td>
<td>(0.00002)</td>
<td></td>
</tr>
<tr>
<td>$Age_{x,t-3}$</td>
<td>-0.015</td>
<td>-0.014</td>
<td>-0.010</td>
<td>(0.0001)</td>
</tr>
<tr>
<td></td>
<td>(0.0001)</td>
<td>(0.0001)</td>
<td>(0.0001)</td>
<td></td>
</tr>
</tbody>
</table>

Observations          | 564,531   | 564,531 | 564,531 |
Log Likelihood         | -127,338.2| -78,639.6| -101,210.8|
Akaike Inf. Crit.      | 254,684.5 | 157,287.2| 202,429.7|

Note: All coefficients significant at $p = 0.1\%$

10.4 – Regression Specification

We now have enough data to run a regression analogous to (35). However, such a regression would be inappropriate for several reasons. First, since the dependent variables are logs, I drop the 6.3% of observations where $n_{c,t} = 0$, so that the results should be viewed as being conditional on positive research activity taking place. In my robustness checks, I add 1 to the observations so that these dropped observations can be included. This causes the magnitude of the estimated elasticity to drop, but it remains significant.

Second, it is clear from Figure 7 that the number of patents granted per year and $Average.affinity.proxy$ are trending variables, so that a simple regression of one onto the other would mostly pick up trends. Since these trends may well vary by class, I estimate a fixed effect model on the differenced data:

$$\Delta \ln n_{c,t} = \alpha_{c} + \beta_{2} \cdot \Delta \ln Average.proxy.affinity_{c,t-1} + X'_{c,t}\beta + \varepsilon_{c,t}$$  \hspace{1cm} (43)
Note the term $\alpha_{cl}$ can account for class-specific growth rates.

Third, to account for higher order trends and omitted variables, I include as controls several lags of the dependent variable $\Delta \ln n_{cl,t}$. Moreover, to control for omitted variables that influence aggregate patenting, I include $\Delta \ln N_t$ where $N_t = \sum_{cl} n_{cl,t}$. Finally, I also add $\text{Age}_{cl,t}$, defined as the number of years since the first patent was assigned to class $cl$, to control for life-cycle effects for a given technology. This model, without the proxy for $\text{Average.proxy.affinity}$ will form the baseline for comparison.

Last, simply plotting $\text{Average.proxy.affinity}$ over time conflates two sources of variation, namely the sample of edges and the value of $\text{Affinity.proxy}$ for each edge. In period $t$, the set of edges used by a class $cl$ is defined as the set of edges used by all patents awarded to class $cl$ in any period prior to or including period $t$. However, defining classes in this manner means the set of edges used is growing over time, rather than constant (as in Section 7.2). To make sure the change in expected affinity is due to changes in affinity, rather than the definition of the set, in each period, I compute $\Delta \ln \text{Average.proxy.affinity}_{cl,t}$ using the same set of edges. Specifically, when comparing $\text{Average.proxy.affinity}$ in period $t$ and period $t-1$, I always use the set of edges defined for period $t$.

10.5 – Data Description

After computing the above metrics, we have an unbalanced panel of 429 classes with up to 176 years of data per class (the panel is unbalanced, because I exclude years before a class has its first granted patent), for a total of 59,592 observations. Some summary statistics are presented below in Table 8.

**Table 8: Summary Patent Class Statistics**

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>Median</th>
<th>Mean</th>
<th>Max</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{Age}_{cl,t}$</td>
<td>1</td>
<td>81</td>
<td>82.8</td>
<td>176</td>
<td>47.39</td>
</tr>
<tr>
<td>Patents Granted ($n_{cl,t}$)</td>
<td>0</td>
<td>57</td>
<td>132.9</td>
<td>8,348</td>
<td>276.1</td>
</tr>
<tr>
<td>$\text{Average.proxy.affinity.1}$</td>
<td>0.071</td>
<td>0.722</td>
<td>0.705</td>
<td>0.999</td>
<td>0.092</td>
</tr>
<tr>
<td>$\text{Average.proxy.affinity.2}$</td>
<td>0.053</td>
<td>0.455</td>
<td>0.454</td>
<td>0.805</td>
<td>0.060</td>
</tr>
<tr>
<td>$\text{Average.proxy.affinity.3}$</td>
<td>0.064</td>
<td>0.543</td>
<td>0.552</td>
<td>0.996</td>
<td>0.090</td>
</tr>
<tr>
<td>$\Delta \ln n_{cl,t}$</td>
<td>-2.984</td>
<td>0.001</td>
<td>0.031</td>
<td>4.414</td>
<td>0.426</td>
</tr>
<tr>
<td>$\Delta \ln \text{Average.proxy.affinity.1}$</td>
<td>0.000</td>
<td>0.012</td>
<td>0.034</td>
<td>2.881</td>
<td>0.090</td>
</tr>
<tr>
<td>$\Delta \ln \text{Average.proxy.affinity.2}$</td>
<td>-1.216</td>
<td>0.010</td>
<td>0.030</td>
<td>3.105</td>
<td>0.110</td>
</tr>
<tr>
<td>$\Delta \ln \text{Average.proxy.affinity.3}$</td>
<td>-0.044</td>
<td>0.008</td>
<td>0.030</td>
<td>2.877</td>
<td>0.097</td>
</tr>
</tbody>
</table>
Note that the mean values for the change in log-transformed variables are all of a similar magnitude, although the $\Delta \ln n_{cl,t}$ has more variance than any of the $\Delta \ln \text{Average.affinity.proxy}$ metrics. Some correlations are presented in Table 9:

**Table 9: Correlations Among Proxies**

<table>
<thead>
<tr>
<th>Average.affinity.proxy.1</th>
<th>Average.affinity.proxy.2</th>
<th>Average.affinity.proxy.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.919</td>
<td>0.661</td>
</tr>
<tr>
<td>0.919</td>
<td>1</td>
<td>0.786</td>
</tr>
<tr>
<td>0.661</td>
<td>0.786</td>
<td>1</td>
</tr>
</tbody>
</table>

Note that these three measures are much more correlated in their differences than in their levels. This stems from the fact that the difference between $\text{Average.affinity.proxy}$ in one period to another is most prominently driven by the edges which go from 0 to 1 assigned patents, which all three proxies measure as a change from 0 to 0.5.

### 10.6 – Results

My baseline model, which does not include the variable $\Delta \ln \text{Average.affinity.proxy}$, is:

$$\Delta \ln n_{cl,t} = a_{cl} + \sum_{i=1}^{8} \phi_i \Delta \ln n_{cl,t} + \sum_{i=1}^{3} \phi_i \Delta \ln N_t + \beta_3 \cdot \text{Age}_{cl,t} + \epsilon_{cl,t}$$

This model forecasts changes in class patenting activity with a class-specific fixed effect, lags of changes in class patents, lags of the change in total patents, and the age of the class. The estimated coefficients for this model are presented in Table 10.

I chose to include 8 lags of $\Delta \ln n_{cl,t}$ and 3 lags of $\Delta \ln N_t$ since the statistical significance levels of further lags is lower, and also reduces the unadjusted $R^2$. The baseline model has a few notable features. First, there is a tendency for the growth rate of patents in a class to converge to the aggregate growth rate of patents, which can be seen from the approximately equal but opposite coefficients on $\Delta \ln n_{cl,t}$ and $\Delta \ln N_t$. Second, the growth rate of patents slows over time, as indicated by the negative coefficient on $\text{Age}_{cl,t}$. Finally, a Hausman test strongly rejects the hypothesis that class fixed effects can be ignored. Different classes have different growth rates.
Unless explicitly stated, all the explanatory variables in equation (44) and Table 10 are included in all the following regression, although I do not report their estimated coefficients to save space.

**Table 10: Coefficients of Baseline Model**

<table>
<thead>
<tr>
<th>Lags</th>
<th>( \Delta \ln n_{ct,t} )</th>
<th>( \Delta \ln N_t )</th>
<th>( Age_{ct,t} )</th>
<th>Class Fixed Effects?</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-0.421 (0.009)</td>
<td>0.446 (0.016)</td>
<td>-0.001 (0.000)</td>
<td>Yes</td>
</tr>
<tr>
<td>1</td>
<td>-0.246 (0.008)</td>
<td>0.248 (0.016)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-0.178 (0.008)</td>
<td>0.130 (0.016)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>-0.120 (0.007)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>-0.099 (0.005)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>-0.066 (0.006)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>-0.033 (0.006)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>-0.021 (0.005)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: All coefficients are significant at \( p = 0.1\% \). White standard errors clustered by class in parentheses.

In Section 8, my preferred specification was a lag of 3 years, but in my robustness check I found the choice of lag had little impact on the estimated results. In this model, the choice of lag is not inconsequential. My first investigation examines all three proxies for \( \Delta \ln Average.affinity.proxy \) across two lag specifications. The results are displayed in Table 11.

As discussed earlier, the significance of explanatory variables with a lag greater than two is consistent with patent grants as a primary vehicle for knowledge diffusion, while the significance of variables with smaller lags indicates knowledge can spread before the patent is granted. I find evidence for both effects, with \( \Delta \ln Average.affinity.proxy \) usually positive and statistically significant for lags of 1 year and 3-5 years, although significance is stronger for the 3-5 year period. Since research may take different lengths of time for different projects, it is not surprising that multiple years of explanatory variable are statistically significant.

The coefficients on all three proxies are similar in magnitudes, but as found in Table 7, Proxy 2 has the strongest predictive power. It is the only proxy with statistical significance for a lag of 5 years, and the adjusted R-squared is usually highest for this proxy. I also conduct an F-test for the joint hypothesis that all coefficients on lagged values of \( \Delta \ln Average.affinity.proxy \) are statistically indistinguishable from zero, but this is rejected at \( p = 0.1\% \) in every specification. Again, the F-statistic is largest for Proxy 2.

Table 11 broadly supports the model presented in this paper. The average expected affinity between elements in a technology field is a positive predictor of the change in patents granted in the future,
with an elasticity between 0.1 and 0.2, but applying over several periods. This occurs even though I control for lagged behavior, and despite the imprecision in the proxies for elements, ideas, and expected affinity.

10.7 – Robustness Checks

Appendix A4 reports a series of robustness checks. As shown in Figure 5 in the main text, the fraction of patents with more than one mainline varies substantially over the period 1836-2012. Since the model is premised on ideas consisting of combinations of at least 2 elements, the model is a better fit for the data after 1936, when 69% of patents were assigned 2 or more mainlines, as opposed to 40% before 1936. This suggests the pre-1936 data may be subject to greater measurement error, which would lead to attenuation bias. Indeed, I find that no coefficients on lags of $\Delta \ln \text{Average.affinity.proxy}$ remain significantly different from zero when estimating regression (43) using data from 1836-1935. Conversely, estimating regression (43) with data from 1936-2012, when 69% of patents contain at least a pair of mainlines, increases the estimated coefficients of columns (4)-(6) in Table 11 by 32%.

Another potential source of bias in my estimates is the construction of proxies. Since these rely on patent grants, they will be correlated with the growth rate of patents over time. If the growth rate of patents is correlated over time, this could introduce bias into the model. While I have attempted to control for this by adding 8 lags of $\Delta \ln n_{cl,t}$, it may be that the proxy is picking up the effect of lags beyond 8. To check for this, I double the number of lags in the last three columns of Table A5. This leads to a strengthening of the results, compared to those found in Table 11, column (4)-(6).

A third potential source of bias may stem from the USPTO’s ongoing updating of its classification system. If commonly used pairs of mainlines are eventually consolidated into single mainlines, then the only pairs we will observe in earlier periods will be pairs of mainlines that were *not* subsequently combined many times. This would tend to bias the results, with the bias more severe for earlier periods. Indeed, if I break the period 1936-2012 into two more periods from 1936-1986 and 1986-2012, the results for the earlier period remain very strong, but the size of estimated coefficients for 1986-2012 grows substantially (the coefficient on $L^3 \Delta \ln \text{Average.affinity.proxy}$ is over 2).
Table 11: Different Lag and Proxy Specifications

<table>
<thead>
<tr>
<th>Dependent Variable:</th>
<th>( \Delta \ln n_{cl,t} )</th>
<th>( \Delta \ln n_{cl,t} )</th>
<th>( \Delta \ln n_{cl,t} )</th>
<th>( \Delta \ln n_{cl,t} )</th>
<th>( \Delta \ln n_{cl,t} )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.001***</td>
<td>-0.001***</td>
<td>-0.001***</td>
<td>-0.001***</td>
<td>-0.001***</td>
</tr>
<tr>
<td></td>
<td>(0.0001)</td>
<td>(0.0001)</td>
<td>(0.0001)</td>
<td>(0.00005)</td>
<td>(0.00005)</td>
</tr>
<tr>
<td>( L\Delta \ln \text{Average.affinity.proxy} )</td>
<td>0.167**</td>
<td>0.127**</td>
<td>0.120*</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.077)</td>
<td>(0.053)</td>
<td>(0.064)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( L^2\Delta \ln \text{Average.affinity.proxy} )</td>
<td>-0.058</td>
<td>0.016</td>
<td>-0.054</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.079)</td>
<td>(0.053)</td>
<td>(0.066)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( L^3\Delta \ln \text{Average.affinity.proxy} )</td>
<td>0.191***</td>
<td>0.142***</td>
<td>0.152***</td>
<td>0.202***</td>
<td>0.149***</td>
</tr>
<tr>
<td></td>
<td>(0.065)</td>
<td>(0.043)</td>
<td>(0.054)</td>
<td>(0.061)</td>
<td>(0.043)</td>
</tr>
<tr>
<td>( L^4\Delta \ln \text{Average.affinity.proxy} )</td>
<td>0.173***</td>
<td>0.167***</td>
<td>0.152***</td>
<td>0.186***</td>
<td>0.179***</td>
</tr>
<tr>
<td></td>
<td>(0.059)</td>
<td>(0.041)</td>
<td>(0.050)</td>
<td>(0.059)</td>
<td>(0.041)</td>
</tr>
<tr>
<td>( L^5\Delta \ln \text{Average.affinity.proxy} )</td>
<td>0.082</td>
<td>0.116***</td>
<td>0.077</td>
<td>0.088</td>
<td>0.122***</td>
</tr>
<tr>
<td></td>
<td>(0.060)</td>
<td>(0.041)</td>
<td>(0.051)</td>
<td>(0.056)</td>
<td>(0.040)</td>
</tr>
<tr>
<td>( L^6\Delta \ln \text{Average.affinity.proxy} )</td>
<td>-0.040</td>
<td>0.011</td>
<td>-0.036</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.049)</td>
<td>(0.034)</td>
<td>(0.042)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( L^7\Delta \ln \text{Average.affinity.proxy} )</td>
<td>0.040</td>
<td>0.015</td>
<td>0.029</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.050)</td>
<td>(0.034)</td>
<td>(0.043)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( L^8\Delta \ln \text{Average.affinity.proxy} )</td>
<td>0.002</td>
<td>-0.022</td>
<td>0.004</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.044)</td>
<td>(0.032)</td>
<td>(0.038)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Controls</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lags of ( \Delta \ln n_{cl,t} )</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Lags of ( \Delta \ln N_t )</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Class Fixed Effects</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>F-Test( ^\dagger )</td>
<td>4.925</td>
<td>5.346</td>
<td>4.000</td>
<td>11.086</td>
<td>11.952</td>
</tr>
<tr>
<td>Adjusted ( R^2 )</td>
<td>0.147</td>
<td>0.147</td>
<td>0.146</td>
<td>0.146</td>
<td>0.147</td>
</tr>
</tbody>
</table>

Note: \( ^\dagger \) The null hypothesis is the joint insignificance of all coefficients on lags of \( \Delta \ln \text{Average.affinity.proxy} \).
White standard errors clustered by class are in parentheses. *\( p<0.1 \); **\( p<0.05 \); ***\( p<0.01 \)
I find a similar effect when I change the measure change in expected affinity for a class with reference to the edges used in the previous period, rather than the current period. That is, when comparing $Average.proxy.affinity$ in period $t$ and period $t-1$, I now use the set of edges defined in $t-1$. This has the effect of omitting all edges that were used for the first time in period $t$, and restricts attention to edges that have already been used at least once. As reported in Appendix A4, this leads to a negative and statistically significant coefficient for the 1836-1935 data, but familiar (and strong) positive and significant coefficients for the 1936-2012 period.

In another specification, I restore observations where $n_{cl,t} = 0$ by transforming the dependent variable to $\Delta \ln(1 + n_{cl,t})$. All variables remain positive and significant in this specification. Lastly, when I omit all classes with that contain less than two mainlines, results are largely unaffected.

11 - Discussion of the Results

The preceding empirical analysis shows that accounting for the combinations of mainlines used in a class of patents does help predict the growth rate of patents by class, as compared to a model composed of lags, fixed effects, and class age. Patenting increases are correlated with new combinations made before the patent was granted. Where the data better fits the model, results are stronger, and modifications to the proxy and data used do not dramatically impact the results. The implied elasticity is on the order of 0.1-0.2 for several years, although this may attenuated by measurement error. If we rely only on data from 1936 onwards, the elasticity rises.

According to the model presented in this paper, a rise in expected affinity for some edges in a class increases the expected value of research, because ideas using these edges are more likely to be effective. The rise in the expected value of research is, in turn, reflected in an increase in class patent output. We would anticipate the rise in class patent output to come disproportionately from patents that include edges whose increase led to an increase in $Average.affinity.proxy$. While I have not tested for this effect directly, this story is consistent with the results from Table 7, which showed the probability a pair of mainlines is assigned to a patent is also increasing in the expected affinity of the edge, and of the edges in related ideas.

Across all specifications, I also find the age of a class has negative and significant impact on the growth rate of patents granted. This is also consistent with the presence of fishing out effects, implying a long-run decline in the growth rate of any one class. As a back of the envelope calculation, the growth rate of 2.7% for all patents over the last 100 years falls to zero in 27 years all else equal, when the coefficient on age is -0.001 (a common result). I find a similar time-scale at work when estimating the probability an edge will be used in any given year. In both cases, I find a story consistent with the model, where technological progress in any given class must constantly reinvigorate itself by discovering new edges have a high affinity, or else research productivity falls to zero.
12 – Concluding Remarks

This paper established that two principles – knowledge is combinatorial and elements of knowledge that work well together in one setting are likely to work well in another – are sufficient to build a model of the knowledge production function that, when embedded into a simple economic framework, generates several stylized traits of the innovation process (knowledge accumulation, technological lock-in, spillovers, etc.). Some stylized predictions about the conduct of research appear to be borne out for US innovation over the previous two centuries.

Empirically, there is additional scope to improve the combinatorial measures created in this paper. Alternative proxies for ideas and elements could be considered, as well as specifications for the estimation of expected affinity. Moreover, the measures of underlying research productivity frontier suggested by this model may be useful in other applications (for example, the measurement of R&D policy effects).

Several lines of future work could emerge from this paper. There are other predictions of this model that were not discussed in this paper. For example, the model makes some specific claims about the evolution of complexity, and the trade-off between basic and applied research. Developing and testing these implications is the source of ongoing work. Furthermore, as discussed briefly in the introduction to Section 8, the simple model of a lone researcher who uses the knowledge production function developed in this paper is woefully incomplete. Incorporating the interaction of multiple agents and the endogenous formation of knowledge over a set of elements would be desirable.
References


Appendices

Appendix A1 – Gittins Index

Appendix A2 – Computer Program Details

Appendix A3 – Robustness Checks for Section 9

Appendix A4 – Robustness Checks for Section 10
Appendix A1 – Gittins Index

The following proof is based on Weber (1992). Suppose we are in the setting described in Section 5 of this paper.

Optimal strategy with learning: The optimal strategy in every period is to choose the option with the highest Gittins Index $\lambda_i(\alpha_i, \beta_i)$ where

$$\lambda_i(\alpha_i, \beta_i) = \sup \{ \lambda : v_i(\alpha_i, \beta_i, \lambda_i) = 0 \}$$

and $v_i(\alpha_i, \beta_i, \lambda_i)$ given by:

$$\max \left\{ \frac{\alpha_i}{\alpha_i + \beta_i} (1 + \delta v_i(\alpha_i + 1, \beta_i, \lambda_i)) + \frac{\beta_i}{\alpha_i + \beta_i} \delta v_i(\alpha_i, \beta_i + 1, 0) \right\}$$

Proof:

1 – A Single Edge

Suppose edge $x_i$ is the only edge available to choose, so that the researcher’s problem collapses to the choice between conducting a 3-element research project that includes edge $x_i$ or to quit research. Thus, the researcher’s problem can be written as a Bellman equation of the form:

$$v_i(\alpha_i, \beta_i) = \max \left\{ \frac{\alpha_i}{\alpha_i + \beta_i} (1 + \delta v(\alpha_i + 1, \beta_i)) + \frac{\beta_i}{\alpha_i + \beta_i} \delta v(\alpha_i, \beta_i + 1) - k, 0 \right\}$$

Next, suppose there is an additional charge to conduct research, which we will call the prevailing charge, denoted $\lambda_i$. The prevailing charge is the same each time the researcher chooses to conduct research on edge $x_i$. The researcher’s problem can then be written as a Bellman equation of the form:

$$v_i(\alpha_i, \beta_i, \lambda_i) = \max \left\{ \frac{\alpha_i}{\alpha_i + \beta_i} (1 + \delta v_i(\alpha_i + 1, \beta_i, \lambda_i)) + \frac{\beta_i}{\alpha_i + \beta_i} \delta v_i(\alpha_i, \beta_i + 1, \lambda_i) - \lambda_i, 0 \right\}$$

Define the fair charge $\lambda_i(\alpha_i, \beta_i)$ as the maximum prevailing charge selected so that, in expectation, the optimal strategy makes zero profit:

$$\lambda_i(\alpha_i, \beta_i) = \sup \{ \lambda : v_i(\alpha_i, \beta_i, \lambda_i) = 0 \}$$

Suppose the researcher is facing a fair charge, so that she is indifferent between conducting research and quitting, since both earn expected profit of zero. If the researcher decides to conduct research, she observes the compatibility of edge $x_i$. If she observe edge $x_i$ to be compatible, her expected profit will be positive going forward, since the charge is fixed but the expected probability of
winning a reward in each period is increased. If she observes edge $x_i$ to be incompatible, the prevailing charge will be too high in the next state, so that the researcher would prefer to quit research, earning zero profit in expectation.

Now suppose the prevailing charge is always lowered to the fair charge rate, whenever the researcher finds herself in a position where it would be optimal to quit research. This does not affect the researcher's expected profit, since she expects to earn zero under a fair charge, but would have earned zero anyway by quitting research. If the prevailing charge is always reduced in this way, so as to always keep the researcher indifferent when she would otherwise prefer to quit, then the researcher need never stop conducting research. Her expected lifetime profit from such a strategy is zero.

This procedure for reducing the prevailing charge generates a stochastic sequence $\{\lambda_i, n\}_{n=0}^\infty$ which is nonincreasing in the number of times $n$ the researcher chooses to conduct research.

2 – Many Edges

Suppose now that there are many edges available for research, each of which has its own prevailing charge that is periodically reduced in the manner discussed above. Suppose the researcher adopts the following strategy:

**Gittins Strategy:** In every period, choose the edge with the highest prevailing charge.

The Gittins strategy insures an edge is chosen in every period (since prevailing charges are always lowered when the researcher would otherwise quit research). Such a strategy has zero expected profit. Moreover, there can be no strategy that yields strictly positive profit in expectation, since this would require strictly positive profit for at least one edge.

Next, note that the sequences $\{\lambda_i, n\}_{n=0}^\infty$ associated with each edge are independent of the strategy chosen, since they depend only on the number of times $n$ an edge has been chosen. The Gittins strategy interleaves the many edge sequences into a single nonincreasing sequence of prevailing charges that maximizes the expected present discounted cost of prevailing charge paid. However, this strategy also yields the maximum expected profit of zero, which means the expected present discounted value of net rewards (absent prevailing charges) must exactly equal the expected cost of charges. Since cost was maximized, this strategy also maximizes rewards, and is therefore an optimal policy.

Since the prevailing charge is periodically lowered to the fair charge, and the fair charge only depends on the state of one edge, an equivalent strategy is to always choose the edge with the highest fair charge, which is given by equation (49).
Appendix A2 – Program Details

This section gives some more details on how I solve the general problem presented in Section 6. The program is in the Python language. To begin, I define the possible sets $D$ in which the researcher may find himself. Since there are 10 ideas, and each idea may be either eligible or ineligible, there are $2^{10} = 1,024$ distinct sets of eligible ideas. For each of these sets, I next define the potential belief vectors of interest. Since I know the initial beta parameters of every $\alpha(x)$, the program can exhaustively list the vectors $B$ that might be attained in any given set.

For example, suppose we are considering the following set

$$D = \left[ (q_1, q_2, q_3), (q_1, q_2, q_4), (q_1, q_3, q_4), (q_2, q_3, q_4) \right]$$  \hspace{1cm} (50)

In this set, all of the ideas composed of three elements are eligible, but all of the ideas comprised of two elements are ineligible. In the model, the only way ideas can become ineligible is if they are tried. Therefore, this state can only be arrived at by the researcher after she has conducted research projects on all the two-element ideas. Specifically, we know the researcher has attempted:

$$\text{attempted} = \left[ (q_1, q_2), (q_1, q_3), (q_1, q_4), (q_2, q_3), (q_2, q_4), (q_3, q_4) \right]$$  \hspace{1cm} (51)

These are the six ideas composed of two elements. Each of these research projects yields information. For each edge, the researcher now has an observation of either one compatibility, or one incompatibility. Therefore, the beta parameters of each edge can take on one of two states: $(\alpha_i + 1, \beta_i)$ or $(\alpha_i, \beta_i + 1)$. Since there are six edges, and each can take on two states, there are $2^6 = 64$ potential belief vectors associated with the set of eligible ideas.

Often, I can simplify matters by ignoring some edges. For example, suppose we are considering the following set

$$D = \left[ (q_1, q_2) \right]$$  \hspace{1cm} (52)

In this set, only one idea is eligible – all other ideas have already been attempted. This implies a large number of potential belief vectors. For instance, once every idea has been attempted, any given edge can take on 9 states, implying potentially millions of different $B$ vectors. However, most of this information is irrelevant in this case. The only parameters we care about are the ones that describe

\[\text{edge values: } (\alpha + 3, \beta), (\alpha + 2, \beta), (\alpha + 2, \beta + 1), (\alpha + 1, \beta), (\alpha + 1, \beta + 1), (\alpha + 1, \beta + 2), (\alpha, \beta + 1), (\alpha, \beta + 2), (\alpha, \beta + 3)\]

---

33 For each edge there are two ideas with three elements containing the edge, each of which may reveal compatible, incompatible, or nothing, and one idea with two elements which may reveal compatible or incompatible. Thus, the potential parameter values are: $(\alpha + 3, \beta), (\alpha + 2, \beta), (\alpha + 2, \beta + 1), (\alpha + 1, \beta), (\alpha + 1, \beta + 1), (\alpha + 1, \beta + 2), (\alpha, \beta + 1), (\alpha, \beta + 2), (\alpha, \beta + 3)$
edges in eligible ideas. In this case, there is just one edge left in an eligible idea, so we do not have to compute the millions of different $B$ vectors that apply to irrelevant edges.

Once it has a set of $(D, B)$ states, the program works backwards. It begins by evaluating the null set $D = \{\emptyset\}$, where the only available option is to quit research and earn zero with certainty (for every belief vector $B$). Next, using this result, the program evaluates the best action for each set with just one eligible idea remaining. When there is just one eligible idea, the problem simplifies to:

$$V(D, B) = \max \left[ E[e(d)] \pi(d) - k(d), 0 \right]$$

(53)

Using these results, the program evaluates the best action for each set with two eligible ideas remaining, which has the form of equation (28). At each stage, it uses the researcher’s beliefs to compute the probabilities associated with each state the researcher may find herself in next period. For example, suppose the researcher has:

$$D = \{(q_1, q_2), (q_1, q_2, q_3)\}$$
$$B = [(0.2, 0.2), (1.2, 0.2), (1.2, 0.2), \ldots]$$

(54)

Where the belief parameters apply to pairs $(q_1, q_2)$, $(q_1, q_3)$ and $(q_2, q_3)$ respectively.

If the researcher chooses research project $(q_1, q_2, q_3)$, then her possible outcomes are:

<table>
<thead>
<tr>
<th>effective?</th>
<th>$V(D, B)$</th>
<th>$B' = B + \omega(d)$</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>$\pi(d) + \delta V(D, B') - k(d)$</td>
<td>$[(1.2, 0.2), (2.2, 0.2), (2.2, 0.2), \ldots]$</td>
<td>0.37</td>
</tr>
<tr>
<td>no</td>
<td>$\delta V(D, B') - k(d)$</td>
<td>$[(0.2, 0.2), (1.2, 0.2), (1.2, 1.2), \ldots]$</td>
<td>0.05</td>
</tr>
<tr>
<td>no</td>
<td>$\delta V(D, B') - k(d)$</td>
<td>$[(0.2, 0.2), (1.2, 1.2), (1.2, 1.2), \ldots]$</td>
<td>0.05</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>no</td>
<td>$\delta V(D, B') - k(d)$</td>
<td>$[(1.2, 0.2), (1.2, 1.2), (2.2, 0.2), \ldots]$</td>
<td>0.02</td>
</tr>
</tbody>
</table>

In fact, there are 12 potential updated belief vectors that may be attained if the idea is ineffective, reflecting the many different ways an idea can be ineffective (compared to the single way it can be effective). To see where these probabilities come from, consider first the probability the idea is effective, given by the first row of Table A1. Since:

$$E[\Pr(e(d) = 1)] = \prod_{x \in d} E[\alpha(x)]$$

(55)

And since
\[ E[a(x_i)] = \frac{\alpha_i}{\alpha_i + \beta_i} \]  

The probability \((q_1, q_2, q_3)\) is effective is

\[ E[e(d)] = \frac{0.2}{0.4} \left( \frac{1.2}{1.4} \right)^2 \approx 0.37 \]

When the idea is effective, each edge is compatible, and so the belief vector next period is given by \([1.2, 0.2, 2.2, 0.2, \ldots]\), where I have added 1 to the \(\alpha\) parameter of each edge in \((q_1, q_2, q_3)\).

If the idea is ineffective, computing the probability of \(B'\) is more involved. Consider the second row, where

\[ B' = [(0.2, 0.2), (1.2, 0.2), (1.2, 1.2), \ldots] \]

Equation (58) indicates the researcher observed edge \((q_2, q_3)\) to be incompatible, but did not observe any other edges. The probability of such a revelation requires using the revelation procedure outlined on page 13, and is the joint product of (1) selecting edge \((q_2, q_3)\), which occurs with 1/3 probability, and (2) finding the edge is incompatible, which occurs with probability \(0.2 / 1.4\). Since \(1/3 \cdot 0.2 / 1.4 \approx 0.05\), the researcher attaches probability 0.05 to this outcome.

Alternatively, consider the final row, where

\[ B' = [(1.2, 0.2), (1.2, 1.2), (2.2, 0.2), \ldots] \]

Equation (59) indicates the researcher observed edges \((q_1, q_2)\) and \((q_2, q_3)\) to be compatible, and edge \((q_1, q_3)\) to be incompatible. There are two ways the revelation procedure could have generated this particular set of observations.

1. It could have (1) drawn edge \((q_1, q_2)\) and found it to be compatible (probability of being drawn is 1/3, probability of being compatible is 1/2), (2) drawn edge \((q_2, q_3)\) and found it to be compatible (probability of being drawn is 1/2, probability of being compatible is 1.2/1.4) and (3) drawn edge \((q_1, q_3)\) and found it to be compatible (probability of being drawn is 1 and probability of being incompatible is 0.2/1.4). The joint probability of this sequence is approximately 0.01.

2. It could have (1) drawn edge \((q_2, q_3)\) and found it to be compatible (probability of being drawn is 1/3, probability of being compatible is 1.2/1.4), (2) drawn edge \((q_1, q_2)\) and found it to be compatible (probability of being drawn is 1/2, probability of being compatible is 1/2)
and (3) drawn edge \((q_1,q_3)\) and found it to be compatible (probability of being drawn is 1 and probability of being incompatible is 0.2/1.4). The joint probability of this sequence is approximately 0.01.

Taken together, the probability of observing this set of observations is approximately 0.02. Similar calculations are performed for each state.

Conversely, if the researcher chooses research project \((q_1, q_2)\), then her possible outcomes are:

**Table A2: Choosing \((q_1, q_2)\)**

<table>
<thead>
<tr>
<th>effective?</th>
<th>(V(D,B))</th>
<th>(B' = B + \omega(d))</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>(\pi(d) + \delta V(D,B') - k(d))</td>
<td>([(1.2,0.2),(1.2,0.2),(1.2,0.2),\ldots])</td>
<td>(\frac{0.2}{0.2+0.2} = 0.5)</td>
</tr>
<tr>
<td>no</td>
<td>(\delta V(D,B') - k(d))</td>
<td>([(0.2,1.2),(1.2,0.2),(1.2,0.2),\ldots])</td>
<td>(\frac{0.2}{0.2+0.2} = 0.5)</td>
</tr>
</tbody>
</table>

In this case, because there is just the one edge, the researcher observes either the edge is compatible (with probability \(\frac{1}{2}\)) or that it is incompatible (also with probability \(\frac{1}{2}\)).

After working backwards, we have a mapping from every \((D,B)\) state to a best action. With four elements and ten ideas, this program still takes approximately an hour to solve depending on computer processing power.
Appendix A3 – Robustness Checks for Section 9

This appendix contains the results from a series of robustness checks on the regressions discussed in section 9.3.

A3.1 – Different Lag Structures

Section 9.3 presented results using a lag of 3 years, but I also experimented with 1, 5, and 10 year lags. The results are presented in Table A3. Once again I omit the estimation of the $\phi_i$ coefficients for space. The choice of lag made no difference to my estimation results.

<table>
<thead>
<tr>
<th>Table A3: Probability An Edge Assigned to a Patent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependent Variable:</td>
</tr>
<tr>
<td>$1(\text{Comp}_{x,t-1} \geq 10)$</td>
</tr>
<tr>
<td>$\mu_{x,t}$</td>
</tr>
<tr>
<td>Lag = 1</td>
</tr>
<tr>
<td>Lag = 5</td>
</tr>
<tr>
<td>Lag = 10</td>
</tr>
<tr>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>2.242</td>
</tr>
<tr>
<td>(0.0096)</td>
</tr>
<tr>
<td>2.242</td>
</tr>
<tr>
<td>(0.0096)</td>
</tr>
<tr>
<td>2.242</td>
</tr>
<tr>
<td>(0.0096)</td>
</tr>
<tr>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>$\text{Related}_{x,t-1}$</td>
</tr>
<tr>
<td>0.001</td>
</tr>
<tr>
<td>(0.00002)</td>
</tr>
<tr>
<td>0.001</td>
</tr>
<tr>
<td>(0.00002)</td>
</tr>
<tr>
<td>0.001</td>
</tr>
<tr>
<td>(0.00002)</td>
</tr>
<tr>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>$\text{Age}_{x,t-1}$</td>
</tr>
<tr>
<td>-0.016</td>
</tr>
<tr>
<td>(0.0001)</td>
</tr>
<tr>
<td>-0.016</td>
</tr>
<tr>
<td>(0.0001)</td>
</tr>
<tr>
<td>-0.016</td>
</tr>
<tr>
<td>(0.0001)</td>
</tr>
<tr>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>Observations</td>
</tr>
<tr>
<td>Log Likelihood</td>
</tr>
<tr>
<td>Akaike Inf. Crit.</td>
</tr>
</tbody>
</table>

Note: Standard errors are in parentheses. All coefficients are statistically significant at $p = 0.1\%$

A3.2 - Fixed Effects

To see if unobserved variables impact my results, I run a fixed effects model. In a non-linear setting, computing fixed effects by de-meaning the data is not appropriate. Moreover, simply estimating a coefficient for dummy variables associated with each edge leads to the incidental parameters problem, which can bias coefficients.\(^{34}\) Fortunately, the Chamberlain estimator is an alternative to demeaning data that can be used in a logistic regression framework. For every edge $x$, the

\(^{34}\) See Greene (2008), pg 800-801.
Chamberlain estimator conditions estimation on the sum of $u_{x,t}$ over edge $x$’s lifecycle, and it can be shown that this approach strips out fixed effects, just as de-meaning the data does in a linear model.\textsuperscript{35} I report the results of a logistic regression with and without conditioning on the sum of $u_{x,t}$ in columns (1) and (2) of Table A4, and the estimated $\phi_i$ coefficients in Figure A1 (left panel). As can be seen, including fixed effects has only modest impacts on my results.

\textbf{Figure A1:} Estimated Coefficients on $\text{Comp}_{x,t}$ with 95\% confidence interval

\textbf{Left Panel:} Logistic Regression with and without fixed effects

\textbf{Right Panel:} Negative Binomial Regression, conditional on $u_{x,t} = 1$

9.4.2 – Count Data

Next, I investigate whether the same explanatory variables can predict the number of patents in a given year that are assigned a particular edge, conditional on at least one observation. Because most observations are small (in fact, “one” is the most common observation), I use a negative binomial model, a common method for studying count data. A negative binomial model is like a generalized Poisson model,\textsuperscript{36} and estimates a conditional mean as:

$$E[ n_{x,t} \mid X_{x,t-1}, n_{x,t} > 0 ] = \exp\left( \beta_0 + X'_{x,t-1} \beta \right)$$

(60)

\textsuperscript{35} See Greene (2008), pgs. 803-805.

\textsuperscript{36} Specifically, a negative binomial distribution relaxes the assumption that the conditional mean of the distribution is equal to the conditional variance. See Long (1997), pgs 230-237 for more discussion. Moreover, as Table A4 indicates, we reject the null hypothesis that theta is equal to one, which would imply a poisson distribution is appropriate, rather than a negative binomial.
where $n_{x,t}$ is the number of patents assigned edge $x$ and granted in year $t$ and $X'_{x,t-1}\beta$ is again given by equation (37). While the coefficients no longer have the same interpretation, the direction of the effects remains unchanged, as can be seen in Table A4, column (3). Moreover, the coefficients $\phi_i$, plotted in Figure A1 (panel right) continue to exhibit an essentially concave shape for $i = 1,...,8$ and I cannot rule out concavity for $\phi_9$, given the width of the confidence interval.

<table>
<thead>
<tr>
<th>Table A4: Robustness Checks</th>
<th>Logistic $u_{x,t}$ (1)</th>
<th>Negative Binomial $n_{x,t}$ (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependent Variable</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$1(\text{Comp}_{x,t-1} \geq 10)$</td>
<td>4.016 (0.018)</td>
<td>3.707 (0.038)</td>
</tr>
<tr>
<td>$\text{Related}_{x,t-1}$</td>
<td>0.001 (0.00003)</td>
<td>0.003 (0.0001)</td>
</tr>
<tr>
<td>$\text{Age}_{x,t-1}$</td>
<td>-0.029 (0.0002)</td>
<td>-0.042 (0.0004)</td>
</tr>
<tr>
<td>Fixed Effects</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Observations</td>
<td>564,531</td>
<td>564,531</td>
</tr>
<tr>
<td>Log Likelihood</td>
<td>-126,446.5</td>
<td>-100,841.0</td>
</tr>
<tr>
<td>Akaike Inf. Crit.</td>
<td>252,917.1</td>
<td>201,704.1</td>
</tr>
<tr>
<td>Estimated Theta</td>
<td>2.397 (0.024)</td>
<td></td>
</tr>
</tbody>
</table>

Note: Standard errors are in parentheses. All coefficients are statistically significant at $p = 0.1\%$
Appendix A4 – Robustness Checks for Section 10

This appendix contains the results from a series of robustness checks on the regressions discussed in section 10.6.

A4.1 – Different Time Frames

As shown in Figure 5 in the main text, the fraction of patents with more than one mainline varies substantially over the period 1836-2012. Since the model is premised on ideas consisting of combinations of at least 2 elements, the model is a better fit for the data after 1936, when 69% of patents were assigned 2 or more mainlines, as opposed to 40% before 1936. This suggests the pre-1936 data may be subject to greater measurement error, which would lead to attenuation bias. Accordingly, I re-estimate my model for two time periods, 1836-1935 and 1936-2012. The results are presented in Table A5.

When we restrict our attention to the period 1836-1935, the results are no longer statistically significant. However, for the second period, encompassing much of the 20th century, the results are significantly strengthened relative to the complete sample (compare to Table 11, column 4-6). This supports the argument that attenuation bias due to measurement error has likely reduced the size of the estimated coefficients. While I do not report the results, if I break the period 1936-2012 into two more periods from 1936-1986 and 1986-2012, the results for the earlier period remain very strong, but the size of estimated coefficients for 1986-2012 grows substantially (the coefficient on \( L^3 \Delta \text{ln Average.affinity.proxy} \) is over 2).

Another potential source of bias in the estimates is the construction of the proxies. Since these rely on patent grants, they will be correlated with the growth rate of patents over time. If the growth rate of patents is correlated over time, this could introduce bias into the model. While I have attempted to control for this by adding 8 lags of \( \Delta \ln n_{dt} \), it may be that the proxy is picking up the effect of lags beyond 8. To check for this, I double the number of lags in the last three columns of Table A5. This leads to a strengthening of the results, compared to those found in Table 11, column (4)-(6).

10.7.2 – Measuring Changes from the Prior Set

In Table A6, I measure the change in expected affinity for a class with reference to the edges used in the previous period, rather than the current period. That is, when comparing \( \text{Average.proxy.affinity} \) in period \( t \) and period \( t - 1 \), I now use the set of edges defined in \( t - 1 \). This has the effect of omitting all edges that were used for the first time in period \( t \), and restricts attention to edges that have already been used at least once. As can be seen in the first three columns of Table A6, this has a significant impact on the coefficients attached to \( \Delta \ln \text{Average.affinity.proxy} \). For all three proxies, some of the lagged changes now have a negative and statistically significant coefficient, in violation of the model assumptions.
Table A5: Robustness over time periods

<table>
<thead>
<tr>
<th>Years</th>
<th>Proxy</th>
<th>1844-1935</th>
<th>1944-2012</th>
<th>1852-2012</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td>0.002***</td>
<td>0.002***</td>
<td>0.002***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.0002)</td>
<td>(0.0002)</td>
<td>(0.0002)</td>
</tr>
<tr>
<td>$L^3 \Delta \ln Average.\text{affinity.proxy}$</td>
<td>0.034</td>
<td>0.023</td>
<td>0.020</td>
<td>1.132***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.067)</td>
<td>(0.047)</td>
<td>(0.057)</td>
</tr>
<tr>
<td>$L^4 \Delta \ln Average.\text{affinity.proxy}$</td>
<td>0.061</td>
<td>0.055</td>
<td>0.055</td>
<td>0.168</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.061)</td>
<td>(0.044)</td>
<td>(0.051)</td>
</tr>
<tr>
<td>$L^5 \Delta \ln Average.\text{affinity.proxy}$</td>
<td>-0.021</td>
<td>0.030</td>
<td>0.013</td>
<td>0.145</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.060)</td>
<td>(0.043)</td>
<td>(0.052)</td>
</tr>
</tbody>
</table>

**Controls**

| Lags of $\Delta \ln n_{cl,t}$ | 8     | 8     | 8     | 8     | 8     | 8     | 16    | 16    | 16    |
| Lags of $\Delta \ln N_t$      | 3     | 3     | 3     | 3     | 3     | 3     | 6     | 6     | 6     |

| Class Fixed Effects | Y     | Y     | Y     | Y     | Y     | Y     | Y     | Y     | Y     |
| Adjusted $R^2$      | 0.181 | 0.181 | 0.181 | 0.126 | 0.126 | 0.126 | 0.140 | 0.140 | 0.140 |

**Note:** *p<0.1; **p<0.05; ***p<0.01
White standard errors clustered by class in parentheses.
<table>
<thead>
<tr>
<th>Years Proxy</th>
<th>1844-2012</th>
<th>1844-1935</th>
<th>1944-2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>-0.001***</td>
<td>-0.001***</td>
<td>-0.001***</td>
</tr>
<tr>
<td></td>
<td>(0.0001)</td>
<td>(0.00005)</td>
<td>(0.00004)</td>
</tr>
<tr>
<td>$L^3\Delta \ln \text{Average.affinity.proxy}$</td>
<td>-0.264</td>
<td>-0.463***</td>
<td>-0.841***</td>
</tr>
<tr>
<td></td>
<td>(0.466)</td>
<td>(0.121)</td>
<td>(0.292)</td>
</tr>
<tr>
<td>$L^4\Delta \ln \text{Average.affinity.proxy}$</td>
<td>0.655</td>
<td>-0.274**</td>
<td>-0.200</td>
</tr>
<tr>
<td></td>
<td>(0.477)</td>
<td>(0.124)</td>
<td>(0.320)</td>
</tr>
<tr>
<td>$L^5\Delta \ln \text{Average.affinity.proxy}$</td>
<td>-1.031**</td>
<td>-0.068</td>
<td>-1.092***</td>
</tr>
<tr>
<td></td>
<td>(0.465)</td>
<td>(0.108)</td>
<td>(0.302)</td>
</tr>
<tr>
<td>Controls</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lags of $\Delta \ln n_{ct}$</td>
<td>8 8 8</td>
<td>8 8 8</td>
<td>8 8 8</td>
</tr>
<tr>
<td>Lags of $\Delta \ln N_t$</td>
<td>3 3 3</td>
<td>3 3 3</td>
<td>3 3 3</td>
</tr>
<tr>
<td>Class Fixed Effects</td>
<td>Y Y Y</td>
<td>Y Y Y</td>
<td>Y Y Y</td>
</tr>
<tr>
<td>Observations</td>
<td>53,930</td>
<td>53,930</td>
<td>23,766</td>
</tr>
<tr>
<td>Adjusted R²</td>
<td>0.145</td>
<td>0.146</td>
<td>0.146</td>
</tr>
</tbody>
</table>

Note: *p<0.1; **p<0.05; ***p<0.01

White standard errors clustered by class in parentheses.
Further investigation reveals this result is not consistent over time. When we restrict our data to the first century, the negative coefficients remain statistically significant, and generally increase in magnitude (see the middle three columns in Table A6). However, when we restrict our attention to the 1936-2012 period (see the last three columns in Table A6), the sign on two of the proxies flips from negative to positive and statistically significant, and the coefficients associated with Proxy 2 are positive but insignificant. As noted above, I believe the results from the later period are more reliable, since the data better fits the model during this period.

Turning first to the later period, the change in coefficients is primarily due to the decrease in variation in $\Delta \ln \text{Average.affinity.proxy}$ when measured with the prior set of edges. For example, $\Delta \ln \text{Average.affinity.proxy.1}$ has a mean of 0.005 when measured with the prior set of edges, compared to 0.034 when measured with the most recent set of edges. The means of $\Delta \ln \text{Average.affinity.proxy.2}$ and $\Delta \ln \text{Average.affinity.proxy.3}$ actually become negative, as well as closer to zero, when measured with the prior period’s edge set. This is primarily because about half of edges are only assigned to a patent once, and therefore much of the period-to-period change in $\text{Average.affinity.proxy}$ comes from edges being assigned their first and only patent. Such changes are only picked up by measuring $\Delta \ln \text{Average.affinity.proxy}$ with respect to the second period. Thereafter, these edges either do not change (in Proxy 1), decline for one period (in Proxy 2), or decline by a small amount every period (in Proxy 3). While Proxy 2 appears incapable of picking up the positive impact of $\text{Average.affinity.proxy}$ for the 1944-2012 period, Proxies 1 and 3 do, in support of the model.

A potential explanation for the negative coefficients in the earlier sample may be that the USPTO’s updated classification has induced bias into the definition of mainlines. As noted in Section 8.3, one explanation for the increase in mainlines per patent over the sample period may be that the USPTO consolidates commonly used pairs of mainlines into a single mainline as time goes on. If this is the case, then combinations of mainlines that were commonly used together do not show up in the early period, because they are later re-classified as a single mainline. Instead, only combinations of mainlines that were not subsequently used remain. These pairs of unconsolidated edges represent failed programs of combination, which may account for the negative coefficient attached to them.

10.7.3 – Alternative Sample Selection

In the first three columns of Table A7, I restore the $n_{cl,t} = 0$ observations by transforming the dependent variable to $\Delta \ln (1 + n_{cl,t})$. This leads to a weakening of the coefficients, but they remain positive and statistically significant. In the last three columns of Table A7, I include only classes with more than 1 mainline nested under the class, in case such classes are unusual or the mainlines they draw on are poor proxies. This does not have a significant impact on estimated coefficients, since there appear to be few patents assigned to such classes.
### Table A7: Alternative Samples

<table>
<thead>
<tr>
<th>Dependent Variable: $\Delta \ln (1 + n_{cl,t})$</th>
<th>$\Delta \ln n_{cl,t}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Omitted Classes</td>
<td>&lt;2 Mainlines</td>
</tr>
<tr>
<td>Proxy</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age</strong></td>
<td>-0.001***</td>
<td>-0.001***</td>
<td>-0.001***</td>
<td>-0.001***</td>
<td>-0.001***</td>
<td>-0.001***</td>
</tr>
<tr>
<td></td>
<td>(0.00004)</td>
<td>(0.00004)</td>
<td>(0.00004)</td>
<td>(0.00005)</td>
<td>(0.00005)</td>
<td>(0.00005)</td>
</tr>
<tr>
<td>$L^3 \Delta \ln \text{Average.affinity.proxy.2}$</td>
<td>0.080**</td>
<td>0.058**</td>
<td>0.065**</td>
<td>0.210***</td>
<td>0.155***</td>
<td>0.163***</td>
</tr>
<tr>
<td></td>
<td>(0.035)</td>
<td>(0.026)</td>
<td>(0.030)</td>
<td>(0.062)</td>
<td>(0.043)</td>
<td>(0.052)</td>
</tr>
<tr>
<td>$L^4 \Delta \ln \text{Average.affinity.proxy.2}$</td>
<td>0.163***</td>
<td>0.136***</td>
<td>0.141***</td>
<td>0.198***</td>
<td>0.191***</td>
<td>0.168***</td>
</tr>
<tr>
<td></td>
<td>(0.033)</td>
<td>(0.025)</td>
<td>(0.028)</td>
<td>(0.060)</td>
<td>(0.041)</td>
<td>(0.050)</td>
</tr>
<tr>
<td>$L^5 \Delta \ln \text{Average.affinity.proxy.2}$</td>
<td>0.105***</td>
<td>0.113***</td>
<td>0.100***</td>
<td>0.078</td>
<td>0.113***</td>
<td>0.067</td>
</tr>
<tr>
<td></td>
<td>(0.032)</td>
<td>(0.024)</td>
<td>(0.027)</td>
<td>(0.056)</td>
<td>(0.041)</td>
<td>(0.049)</td>
</tr>
</tbody>
</table>

**Controls**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lags of $\Delta \ln n_{cl,t}$</td>
<td>8†</td>
<td>8†</td>
<td>8†</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Lags of $\Delta \ln N_t$</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Class Fixed Effects</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
</tbody>
</table>

| Observations | 61,678 | 61,678 | 61,678 | 52,358 | 52,358 | 52,358 |
| Adjusted $R^2$ | 0.168  | 0.168  | 0.168  | 0.145  | 0.145  | 0.145  |

**Note:** † - Controls are lags of $\Delta \ln (1 + n_{cl,t})$. White standard errors clustered by class in parentheses.

*p<0.1; **p<0.05; ***p<0.01