The Creative Development of Fields: Learning, Creativity, Paths, Implications

Jonathan S. Feinstein

John G. Searle Professor of Economics and Management
Yale School of Management

Abstract

I present a model of the creative development of a field and analysis of the model based on an extensive set of simulations. The field begins from an initial states and grows as individuals enter the field and make new contributions; its basic structure resembles a lattice. New elements are created via combining preexisting elements, based on specific rules for combinations. Individuals working in the field follow a defined process of creative development, working to maximize the expected value of their contribution to the field. An individual selects an initial set of elements in the field to learn, then gains intuitive signals about potentially fruitful new combinations based on this learning set, selects additional elments to learn, and finally chooses a potential new element to attempt to make. If the element is viable it is added to the field, together with any subbundle elements co-created with it. The simulation analysis reveals a set of key features that characterize the development of fields through this process. There is a rich diversity of possible paths of development; this diversity is generated especially by the intuitive signals individuals receive, which lead them to attempt to make elements they might otherwise not pursue, thus shaping the development of the field in important ways. The results also reveal a high degree of path dependence, generated as individuals build on the work of their predecessors, and interesting temporal patterns for how output in one period is linked with what occurred in the previous period.

I am grateful for summer research support from the Yale School of Management. I am grateful to Soonwook (Wookie) Hong, Sharon Qian and Xuan Zhang for research assistance. I thank the staff of the Yale High Performance Computing Center for assistance and access to the Yale HPC; I am especially grateful to Andy Sherman. I thank Arthur Campbell, Constança Esteves-Sorenson, Liane Gabora, Luis Garicano, Edward Kaplan, Matthew Rabin, Peter Schott, Olav Sorenson, Bruce Weinberg, and seminar participants at the London School of Economics, Texas A&M, the University of Copenhagen, the University of Edinburgh, the University of Pennsylvania, the University of Alabama, the Yale School of Management, and the "Economics of the Arts" session of the AEA held in January, 2011 for helpful comments. I am fully responsible for the ideas and content and know that none of them wants to be.

1. Introduction

In this paper I present a model of the creative development of a field. The field is defined as an explicit knowledge structure that starts from a simple initial state consisting of a few elements, then develops through the series of creative contributions made by successive individuals who enter and work in the field over time. New elements are added by combining preexisting elements in new combinations, according to specified rules; as a result the field resembles a lattice in structure. Individuals who work in the field follow a rational process of creative development that includes learning, gaining intuitive signals about potentially fruitful new combinations, and finally choosing a new element to attempt to make guided by their signals, based on an expected value calculation. I present a thorough analysis of the model's implications based on extensive simulations and analysis of results. The results demonstate great diversity in the possible paths of development of a field starting from a given initial state, and very substantial path dependence in how a field develops, as individuals build on the work of their predecessors. The results show the importance of intuitive signals in guiding individuals and generating this diversity of paths, generating new elements that are not created in the null model in which there are no signals. I present distributions for output and the number of elements created in the field. In addition, I explore how conditional expected output in a period depends on the choices and outcome of the preceding period, generating implications about the time series properties of output in the development of a creative field.

The paper fits in the very large literature on modeling the innovation process that drives economic and cultural development. The importance of innovation to economic growth and the progress of human civilization has been recognized from the beginnings of the modern literature on growth (Solow (1957), Abramovitz (1956)) and in the Austrian focus on knowledge and individual initiative (Hayek (1960)). Indeed the central role of innovation in the development of industries was emphasized by Marshall in *Industry and Trade* (1919) and the importance of freedom of expression, creativity and experimentation is the focus of John Stuart Mill's *On Liberty* (1859) (for a recent contribution on openness see Murray, et. al. (2009)). In the modern theory of endogenous growth the production of new ideas is central to economic development (Romer (1990), Aghion and Howitt (1992)). Other branches of the literature have emphasized the importance of institutions (Acemoglu, Johnson and Robinson (2005) provide a review), including intellectual property (Mokyr (2002)) and knowledge spillovers (Griliches (1992), Romer (1986)).

The model in this paper focuses on modeling the creative process in the context of a field of creative activity, which could be a scientific or other intellectual field, a field of technology or design or a practice. The process has some of the features of a process of searching for the best new alternative from a distribution of possibilities (Evenson and Kislev (1976), Kortum (1997), Fleming and Sorenson (2004)). Here however creativity is modeled specifically as combining existing elements in new ways to create new elements. This approach is based on the widely accepted definition of creativity in the field of creativity studies as connecting or relating preexisting elements that have not previously been connected or related (for example, Mednick (1962), Koestler (1967), an early statement is Poincaré (1908)). My approach is connected to the important contribution of Weitzman on recombinant growth (1998) (see also (Ghiglino (2012), Feinstein (2011)), although I focus more on the creative process and less on resource limits on the development of new ideas. In a distinct but related strand of development Garicano (2000) develops an important model matching

knowledge to problems to generate solutions, one form of creative connection, and explores its implications for organizations. The model here provides a more explicit, richer learning process, that includes a role for intuitive signals generated though first round learning, and thus a richer framework for understanding the creative process and factors that influence it. The model is also connected with the model of creative development developed by Gruber (1974), Feinstein (2006), Gabora (2005), Cohen (2009) and others, which emphasizes that creativity typically emerges out of a lengthy process of guided exploration. In addition, it is a model of a field, so new ideas are generated in the context of the field (for a recent different formal model of the development of a field see Bramoullé and Saint-Paul (2010)). In turn, this allows a representation of how knowledge in a field grows over time, and reveals the structure of the field in terms of how new ideas are being generated based on and in relation to older ideas. Interestingly, the structure of the field resembles a lattice, thus also providing a link with the field of economic and social networks (Jackson (2006), Goyal (2005), Campbell (2013)). The framework allows for evaluation of the dynamics of how a field develops in relation to learning processes as well as a host of policy issues, including intellectual property, addressed in the last section and modeled as a royalty payment.

Explicitly modeling the creative process that drives innovation and knowledge creation is important. By understanding how this process works we will be better able to understand and gain some ability to predict the dynamics of how economies and fields develop. This includes how human agents respond creatively to shocks, as well as how they generate new ideas endogenously within a field. The aim is not to predict the exact next idea or innovation, but rather to build models that enable us to calibrate and appreciate the range and distribution of outcomes that may arise given the current state of a field or larger system. As I show with the results of this paper, the range is in fact large, and in fact is itself highly variable for different historical paths starting from the same initial condition, with a high amount of path dependence.

A key motivation of this paper is to present a framework that links economic models of creativity and innovation with the field of knowledge representation. Knowledge representation provides a conceptual framework for describing concepts and their relationships (Sowa (1984; 2000); Wille (1992); Ganter and Wille (1997), Ganter, Stumme and Wille (2005)). It is also useful for natural language based description (Helbig (2006); Jackendoff (1988)); in the different context of strategic interaction Feinberg (2005, 2008) has developed an important link of linguistics with game theory modeling; and recently Kaplan and Vakili (2013) have developed a text-based approach, centering around identification of topics, to analyzing patent usefulness and novelty. A knowledge structure is defined based on a set of basic concepts, which may include references to entities (for example physical elements, people), attributes, actions, and context (such as spatial and temporal locations). More complex elements are created or defined by linking basic concepts together to create new elements, for example a novel product design, new theorem, or new policy. Lancaster's model of attributes defining the value of goods (Lancaster (1965)) is a well known related approach in economics; I discuss this interpretation in the next section. However the knowledge representation framework provides a better basis for defining how new elements are created, including far more structure about the logic and rules for combining elements.

I develop a simple example of a knowledge representation framework of a field and use it to explore how the field develops. Specifically, I define elements in the field as strings made up out of basic "letters" or attributes. New strings are created by combining two preexisting strings according to defined rules, consistent

with both the basic definition of creativity and the knowledge representation focus on the rules that specify how new elements are produced. Over time the field grows as individuals create new string contributions that are added to the field. Further, new strings can be longer than any previously created strings, so that complexity of elements increases. Success is uncertain and a new element that is attempted has a probability of being viable. A viable new element has associated an output drawn from a distribution that defines its economic value; this distribution has the properties that are recognized as empirically important for the creation of new economic value, specifically a long right tail so that there is a small probability of a very high value contribution. Thus as the field grows the new contributions that are made generate a stream of economic value.

The heart of the model is a rational, optimizing model of individual creative development. The model centers on learning, intuition generated by signals individuals receive, project selection choice guided by these signals, and outcomes. The intuitive signals, central to the model of creative development, pertain to subbundles of attributes that may be created when an element that embeds them is created. Thus intuition is not necessarily about a fully defined final product, but about a "wish list" of attributes to be bundled together, and the task is to find a way to create a new element that embeds them like a template. This captures the commonsense view of the creative process as guided by partial (or as it is sometimes called "fuzzy") vision or simple insight that is then developed further, filled in and perfected. Some signals pertain to subbundles that it is not possible to create given the current state of the field. In this case - I call these *clean* runs - the individual simply optimizes based on the combinatorics and common probabilities defining what is the most attractive new element to attempt to make. An important finding is that much of the diversity of possible paths of development of the field is generated by the intuitive signals, which lead individuals to attempt to make elements they would otherwise not attempt, in turn opening up new frontiers for future development.

Each time period a single individual enters the field and goes through the creative development process outlined above. The model is inherently nonstationary in that the field grows over time, which means that the set of feasible new combinations changes over time. In general the set of potential new combinations may grow exponentially (subject to the rules of what are valid combinations) with the size of the field, in this way similar to Weitzman (1998). The field has a public history that records for each preceding individual who worked in the field in the past, his initial and final learning sets, the creative project he attempted and its outcome; the intuitive signals individuals received are not publically known and thus not part of the public history. At the start of a period, using this public history, the probabilities of viability and high output are updated for each potential new combination that may be produced in the field based on what happened in the preceding period. This updating includes indirect inference about intuitive signals, since if for example the individual working in the field in the preceding period undertook an unusual project that likely indicates something about his intuitive signals that led him to make this choice. Through tracking how the history of the field and its updating relate to the choices of subsequent individuals the model provides insight about the influence historical knowledge has on the development of a field. Model results also speak to how frequently an individual attempts to build a new element using as a building block an element created in the immediately preceding period, versus an element created further back in time, thus describing the temporal pattern of chain of linkages among elements. Overall, the model in this paper

provides a basis for describing the dynamics of development of a field that is far more rooted in learning and rational choices than psychological models, including the Darwininian model of random variation and selection (Campbell (1960), Simonton (2002)).

I analyze the model through extensive simulations. I explore several different paramater cases, for each running three simulations, as well as three companion runs with no signals, which I call *clean* runs. The simulations within a case differ in their stochastic probability draws determining which new elements are viable and their values if viable; in particular for each simulation I generate a *masterlist* that specifies for each potential new element whether or not it is viable and its value (output) if it is viable. I run each simulation out five periods, identifying every possible path of development of the field assuming that individuals follow the optimal creative development strategy; paths differ in the set of intuitive signals individuals receive, which guides their choices about learning and which new element to attempt to create.

The results generate a set of interesting results. A first striking finding is the diversity of possible paths of development starting from a given initial state. A typical simulation generates several thousand paths. Further there is a great deal of diversity, although less, in the set of field structures generated over these paths: many simulations generate more than 100 distinct field structures through the first five periods. Even when multiple paths generate the same field structure they still differ in their histories, including elements that have been attempted but turned out not to be viable; hence different paths have different probabilities of viability and value for potential new elements going forward. In addition to the large number of structures for any single simulation, the number of distinct structures itself varies widely across simulations. Thus there is variability in the potential variability of development of the field, depending on parameter values and outcomes early in the history of the field. Linked to the diversity of structures, there is also substantial variance in the number of new elements created and generated output. It is important to note that the *masterlist* of which elements are created and their value is held constant and thus is not the source of the variability across paths within a simulation; however variations in masterlists contributes to variability across simulations for a given set of parameter values. Interestingly, there are many fewer paths and structures for clean runs. Thus the intuitive signals are central in generating the diversity in possible paths of development, leading individuals to attempt elements that would otherwise not be attempted, in turn opening up new frontiers in the field.

A second, related finding is the high degree of path dependence. Dividing paths based on the element created (or attempted but not viable) in the first period, the differences across these paths sets are striking. While there is overlap, in nearly every case when I compare two paths each path leads to new elements that are not created along the other path, and in a fair proportion of comparisons many elements are distinct. This result fits naturally with the cumulative nature of creative development of the field, as individuals build on the work of their predecessors. A noteworthy feature, at least through five periods, is that once two paths diverge they do not tend to reconverge, but generate substantially different sets of elements going forward, in most cases with differences growing over time.

A third set of findings concerns conditional output calculations and the dynamics of output, in particular how productive individuals are in terms of creating new elements and output based on what occurred in the preceding generation. A basic distinction is whether a new element was successfully created or not in the preceding period. Concerning this distinction there are in fact two effects working in opposite

directions. On the one hand, when a new element is created it opens up new possibilities to the extent it can be combined with other elements in the field - how many such combinations are feasible to attempt will vary depending on the element and the existing elements in the field. On the other hand, since this element was a potential new element to attempt to make, and has now been attempted and made, it is no longer available to be made in the future, which reduces the set of available elements within the pool of potential new elements. To explore conditional output further I further subdivide the previous period based on whether the element that was attempted was chosen because a live signal led the individual to choose it or whether it was a *clean* run. It stands to reason that the second effect will be especially important in the latter case since the set of elements available when there are no live signals, which occurs relatively often, is reduced, whereas when the new element that is attempted is based on a particular signal, this is more likely to be an element that would not typically be chosen. Results generally confirm this intuition: for most simulations it is the case that the difference between conditional expected output when a new element has been made in the preceding period and conditional expected output when no new element has been made is greater when the individual in the preceding period was guided by a live signal than when he was not. Thus the model generates implications about the time series of output, including how the dependence varies depending on the nature of the creative process in the preceding period.

Finally, a fourth set of results explores the extension to the basic model when individuals earn a royalty if an element they create is used in the following period. A royalty set at thirty percent of the expected value of a contribution only slightly modifies the main results, suggesting fairly large incentives would be needed to strongly influence the path of development of the field in this setting.

The remainder of the paper is organized as follows. The next section introduces basic terminology about the field, its knowledge structure, the creation of new elements, and valuation of new created elements. Section 3 describes the process of creative development of individuals working in the field, including their learning process, intuitive signals, and their optimal strategy. In section 4 I describe the development of the field including its structure and how the history of the field is updated after each generation. In section 5 I outline the simulation methodology and an extensive set of simulation results are presented and discussed. In section 6 the extension of the model in which individuals earn royalties if elements they create are used in the next time period is described and simulation results for this extension are presented. Finally, section 7 concludes. An appendix contains some additional formulas.

2. The Field: Elements, Creativity, Values

The field consists of elements. There are a variety of kinds of elements and different fields will have different elements and different proportions of the different kinds of elements. Elements may be of many kinds; they include definitions, attributes, heuristic rules, values, facts (empirical phenomena) and data, objects, designs, and theorems. Elements thus include both tangible and intangible kinds; they also may include events, places, and people. More complex elements are built up out of simpler elements as the field grows; I focus in particular on the creative process through which individuals working in the field create new elements. As these complex elements are created they create links among existing elements, building relationships among elements; this is discussed in detail below. The approach follows the formal concept

analysis approach (Wille (1992); Ganter and Wille (1997)) and more broadly the knowledge representation framework discussed further below as for example outlined by Sowa (1984; 2000).

Textbook accounts of a field give a sense for the elements in a field, up to some level of complexity, and including some but not all elements. For example the well known graduate text *Microeconomic Theory* by Mas-Colell, Whinston and Green (1995) includes 23 Chapters. Chapter 1 is entitled "Preferences and Choice" and introduces a set of definitions about preferences and preference orderings and propositions about these defined elements. Chapter 2 is entitled "Consumer Choice" and applies material from Chapter 1 to the problem of consumption choices from a consumption set subject to budget constraint, working to definitions of demand functions and comparative statics and the weak axiom of revealed preference. Chapter 3 is entitled "Classical Demand Theory" and includes utility maximization, expenditure minimization and duality, integrability, and the strong axiom of revealed preference. In separate work Arthur Campbell and I (Campbell and Feinstein (2013)) have essentially decoded the set of concepts in this book, on the order of 1,000 concepts in total; however we have not taken the step of organizing the material as a formal knowledge structure. A related notion is the WordNet lexicon for the English language (Miller, et. al. (1990); Princeton (2013 website); see also Helbig (2006) for a linguistic knowledge structure framework). WordNet provides information about nouns, verbs, adjectives and adverbs and their inter-relationships. It groups words into what are called synsets and builds hierarchies from more general to more specific concepts within a family, an approach that is compatible with the knowledge structure approach in this paper.

Clearly a knowledge structure for a well-developed field is rich, filled with many elements. In this paper the knowledge structure will be quite simple relative to these applications, so as to allow focus on development of a formal model of how a field develops.

2.1. Attributes and Elements

The most primitive units in the field are *letters*. There are N letters; a_i denotes letter i. We assume letters cannot be subdivided further.[1] Letters can be thought of as primary attributes as one interpretation of the model.[2]

Elements in the field are comprised of strings of letters linked together. An element in the field is an ordered string of letters. Order matters. Thus $a_1 - a_2 - a_3$ is taken to be a distinct element from $a_2 - a_1 - a_3$. However, which end of a string is given first does not matter, so that $a_1 - a_2 - a_3$ is the same as $a_3 - a_2 - a_1$. New elements are new strings created by combining two parent strings. There are rules for how strings are created and thus which strings can be created given the current state of the field. These rules are described in subsection 2.3. Not all strings are viable: some strings can be made and others cannot be. Viability is discussed in subsection 2.4. When a string is viable it has a value drawn from a distribution - this is also specified in 2.4. The overall value of a string is the sum of the value of the main string and the values of any subbundles co-created with the main string, also discussed below.

[1] The number of letters or attributes is ultimately not critical, since, as discussed in the mathematical theory of communication (Shannon and Weaver (1971)) and theory of complexity and algorithms (Li and Vitanyi (1997)) alphabets can be translated into one another. From this viewpoint the most basic alphabet is the 0/1 binary alphabet and any list of attributes can be redescribed as strings of zeroes and ones. However, a particular alphabet or list of attributes typically has a natural interpretation. Further, the rules for which strings can be combined described in Section 2.3, which are transparent for the given alphabet, would be more complicated form if translated to hold for another alphabet such as the binary.

[2] In a more general model sets of letters are linked to attributes, so that for example a string of letters can producer an emergent attributes not present in any of the letters individually.

Beyond the direct interpretation of the model as combining components to form larger elements, there are connections with two important literatures. Defining elements as strings of attributes is reminiscent of Lancaster's model of consumer demand in which products are bundles of attributes, and the actual value consumers place on a product is based on its underlying attribute bundle (Lancaster (1966)). Lancaster and the extensive literature that follows does not focus on the ordering of a set of attributes in defining the value of a product. More fundamentally, neither Lancaster nor the subsequent literature consider the creative process through which attributes are bundled together to create new products. The approach in this paper can be viewed as building on the product attributes literature, focusing on the creative process through which new products or more generally cultural elements are created by bundling attributes or subelements together.

Another interpretation is provided by viewing the model from the perspective of the knowledge representation paradigm as noted above. Knowledge representation is the study and development of frameworks for classifying knowledge and describing elements and how they relate to one another and change. It can be used to describe the set of elements in a field, in a systematic fashion. An element in a knowledge representation for a given field has a set of components and attributes associated with it that define the element. The framework is flexible and can be applied to many fields. For example, in electronics integrated circuits are often described in terms of components, the properties of those components, and circuits that define how a set of components are interconnected, together with properties of the circuit as a whole (see Dorf and Svoboda (2006) for textbook treatment). This is a more complex structure and representation than the string formulation in this paper, especially in being inherently two dimensional and with a sharper distinction between components and properties or attributes; but it is consistent with the general approach adopted here. It is also interesting, given the focus of this paper, to note that the field of integrated circuits has developed enormously. Complexity has also increased dramatically, including the development of Very Large Scale Integrated circuits (VLSI) (Mead and Conway (1980) is an important early statement of this shift to more complex systems). This fits the approach in this paper, in which strings get longer as the field develops, representing more complex elements. Other examples of application include visual art, in which often a two dimensional representation of form and color is relevant, architecture, fields of design, chemical engineering, and musical composition. In all cases a framework can be developed that incorporates basic elements, attributes, more complex compound elements built up out of simpler elements, ultimately leading to the creation of quite large elements or systems incorporating many elements in rich structures. While the model in this paper is quite simple relative to all of these examples, it provides an initial bridge between economic or decision-based models of learning, creativity, and knowledge creation and the knowledge representation framework.

Initially the field is comprised of N_0 elements each of length two that have been created previously. We do not consider how these short strings have been created but take the initial state of the field as given.[3]

2.2. Creating New Elements

2.2.1. Definition of Creativity

[3] These elements may be viewed as having been created in another field.

The model of creativity is based on the definition of creativity as *connecting, combining or relating two or more elements that have not previously been connected, combined or related*. Specifically, individuals create new elements in the field by combining two previously created strings that have not previously been combined.

This definition is widely accepted in the field of creativity studies (see for example Mednick's idea of remote associates (Mednick (1962)); Koestler's definition of creativity as bisociation between two networks of thought (Koestler (1967)); Ward, Smith and Vaid (1997)); in economics it has been introduced in Weitzman (1998) and is also pivotal in Feinstein's (2006) framework of creative development. Creativity as a novel connection of two or more elements has a far wider application than might at first appear. For example, it incorporates connecting a novel solution with a problem (Fauconnier and Turner (1998); Poincare (1909)), metaphor (Gentner (1983); Ward, Smith and Vaid (1997)), connecting a theoretical framework with a novel application, employing a conceptual schema to make sense of an experience or sensory data, and connecting a concept with its physical manifestation via an explicit technology.

Despite how fundamental this definition of creativity is, it has not found its way widely into the economics literature on creativity and innovation. Most commonly in economic models, as discussed in the Introduction, a formal model of creativity is not specified. Often the focus is on how many resources are devoted to innovation with the basic assumption that more resources increases the likelihood of an innovation being made, but without specifying an explicit underlying model of the creative process. An important literature has emerged that describes the creative process as search drawing from a distribution of possibilities (Evenson and Kislev (1976), Kortum (1997); see also Jones (2010)). While this literature does provide a more explicit model of how innovations are generated it is not rooted directly in the creative process as making novel connections and combinations, and does not incorporate a knowledge representation structure of the field in which the innovations are sought. Three exceptions to the general rule that creativity is not explicitly modeled as novel combinations are Weitzman, Feinstein (2011), and Ghiglino (2012), but even these models do not incorporate a very explicit knowledge structure.

2.2.2. Conceptual Bridges and the Creation of New Strings

Not all combinations make sense. In this paper I specify rules for which combinations are valid based on the cognitive science principle of conceptual blending (Fauconnier and Turner (1998; 2002)). According to conceptual blending conceptually valid combinations are made by connecting two concepts that share a bridge or overlap that enables their distinct elements to be linked. Thus the combination of two technologies requires that they are able to be fit together via a bridge element, such as matching states or a viable physical linkage. As Fauconnier and Turner discuss at length, a valid solution to a problem requires that some conceptual frame be overlapping between the problem and the proposed solution, so that the solution fits the problem. Likewise, a statistical model "fits" with a dataset if the data satisfies the conditions for correct application of the model: for example if the model is a binary choice model then the data must also have the property that its dependent variable has two outcomes in order for the model to fit and be used for estimation. At a practical level, in fields of design and technology, for example integrated circuits, new circuits are created by linking existing circuits, and to be linked the components must fit together at the bridge connection. In the case of metaphor the metaphor "source" domain must fit or overlap with the

problem or "target" domain (Gentner (1983)) to generate a valid insight into the problem or a sensible link. More broadly a creative association of any kind is possible only when the two elements being linked share some kind of point of connection, which could be experiential (such as overlapping in space and time) or conceptual (sharing terms or variables in common).

I apply this principle here with the condition that *two strings can be combined into a new larger string when they share a letter or substring of letters in common, each having the letter or substring on an end; they can then be combined via their overlapping ends.* Figure 1 illustrates this combining process. Since each string has two ends there are four possible ways to connect two strings, which I denote LL (the left edge of the first string connects with the left edge of the second), LR, RL and RR. I keep track of these directionality issues in the model and simulation, so that two strings can be combined in more than one way and these different ways may lead to different new strings.[4]

Strings $a_1 - a_2 - a_3$ and $a_3 - a_4 - a_2$ can be combined via the right edge of string 1 overlapping with left edge of string 2, a RL connection, to form the new string $a_1 - a_2 - a_3 - a_4 - a_5$. Figure 1a illustrates this combination. Note that when the strings combine only a single a_3 is included as this element overlaps between the two strings. Strings $a_1 - a_2 - a_3$ and $a_4 - a_5$ cannot be combined. Strings $a_1 - a_2 - a_3$ and $a_1 - a_3$ can be combined in two ways: as an LL connection producing new string $a_3 - a_2 - a_1 - a_3$, and as a RR connection producing new string $a_1 - a_2 - a_3 - a_1$. Strings $a_1 - a_2 - a_3$ and $a_1 - a_2 - a_1$ can be combined in two ways, as a LL and a LR connection, however both lead to the same final string: $a_3 - a_2 - a_1 - a_2 - a_1$. Overlaps of more than a single element are also allowed. Thus, strings $a_1 - a_2 - a_3$ and $a_2 - a_1 - a_4$ can be combined in the LL direction with two elements, $a_1 - a_2$, overlapping, producing string $a_4 - a_1 - a_2 - a_3$. Figure 1b depicts this combination, with string two placed under string one and flipped.[5]

New strings are always longer than either parent string. Thus over time longer more complex elements are created in the field. In this important sense the field is non-stationary. This feature is consistent with the growth of many fields and human culture (as for example in Jones (2009)), which clearly becomes more complex over time, notwithstanding that important simple ideas also continue to be added to the pool of creativity and innovations, here as new lateral strings.[6] Thus the field can grow both in depth or complexity, through longer elements being added, and also in breadth, with more elements of a given length also being added over time.

The model of string elements can be generalized in several ways. One important generalization is to allow attributes to be distinct from the building blocks (letters) used to construct elements. In this approach, there are two kinds of elements, letters and attributes, and each string of letters has a set of attributes associated with it. The main advantage is greater flexibility, specifically in how attributes change as strings are combined to form new elements. In particular, some attributes may drop out and new "emergent" attributes may be associated with a new element. For example, a chemical compound created out of two component chemicals may possess new attributes neither parent possesses.

[4] There are two ways to write any given string as noted above. I assume that one way is chosen and fixed so that whether its left or right end is being used for a given pairing is well-determined. All subsequent strings that are then checked against this string are checked starting both from its left edge moving right and also in the reverse order, starting from the right and moving left. If a string matches in either direction it is considered to be the same string.

[5] For a valid new combination there must be at least one element from each parent that is not contained in the other parent, so that each parent contributes at least one new letter to the new string.

[6] There is a related literature on the growth of complexity in the theory of evolution; see Adam, Ofria and Collier (2000) and Heylighen (1996)).

A second generalization is to specify elements using functions, rather than strings. Rules can then be specified that specify which building block elements can be linked to create new elements, more flexibly than the simple edge conceptual overlap approach I use. This is the approach used in many areas of knowledge representation. For example in natural language representation rules of grammar are defined as functions that specify which kinds of elements can be combined to construct valid new statements, such as a noun with a verb with a direct object; see Helbig (2006) for an excellent discussion and examples. In electronics rules for the creation of new circuits out of existing circuits and components employ circuit diagrams to specify which elements can be combined to produce new circuits and in what ways they can be combined. These rules mandate certain types of interconnections between parts as valid, and may also impose other constraints that need to be satisfied, such as system balance. Typically in the function approach each element has a set of attributes that define valid roles it may play in the creation of new elements and functions are defined that determine how elements with given roles fit together (or not). In general, for two building block elements g_1 and g_2 a function $f(g_1, g_2)$ specifies whether g_1 and g_2 can be combined, and if so also specifies the set of attributes associated with the new element. This process is recursive: the new element itself can become a building block, for example: $f(f(g_1, g_2), f(g_3, g_4))$. The model of this paper fits with this approach, but only as one simple case, within a large family of possibilities.[7]

2.2.3. The Creation of Subbundles

An important feature of the model, linked to the intuitive signals that guide creative development, is the creation of subbundles consisting of subsets of letters created jointly with the creation of a main string. For example, linking $a - 1 - a_2$ with $a_2 - a_3$ to create string $a_1 - a_2 - a_3$ may also create as a byproduct the subbundle consisting of a_1 and a_3. These two letters are not adjacent in the final string, but they are implicitly connected since they are embedded in the larger string and linked via the bridging element a_2. Therefore if they have value as a joint pair that value can be realized via the creation of the parent string. In fact, of the four attribute bundles created by forming $a_1 - a_2 - a_3$ only two are novel: the full string and a_1 with a_3 (the other two are the parent strings, $a_1 - a_2$ and $a_2 - a_3$). Thus this subbundle may be quite important for the total value created by the new string. The perspective on the creative process consistent with this model is that when a new product or element is created, the main source of value created is not necessarily the full product (though it may be in reality and in the model) but rather can as well be a subbundle of attributes that have never been jointly produced before. I do not impose a specific assumption about which bundle is more important, but rather define the total value realized when a new string is created to be the sum of the value of the main string plus the values of all co-created subbundles. Since these values are all different, any one of them can be the predominant value created. I impose the condition that the subbundles are not realized unless the main string is viable.

The creation of a given subbundle of attributes or elements may be the principle aim in creating a larger parent element. Consider for example a new consumer product. This product may incorporate many elements or attributes, but many of these may primarily serve as bridges or connectors that make the overall

[7] Formally, the conceptual overlap model with strings states that two elements g_1 and g_2 can be combined for example in the RL orientation if the last letter or g_1 is the same as the first letter of g_2 (for a single overlap; similar rule specifies conditions for a double overlap and so forth). In turn if the condition is met the new string g_3 is defined as the sequence of letters of g_1 linked with g_2 removing one letter (two for a double overlap).

product feasible, but are not the main source of value. Rather, a certain specific subbundle of attributes may be the key value creator. For example, a new car design will have many elements, but there may well be a few key features bundled together that create the underlying consumer value; this view is certainly consistent with the Lancaster model. Likewise, a chemical process that fuses $a_1 - a_2$ with $a_2 - a_3$ to create the new material $a_1 - a_2 - a_3$ may use a_2 as a bridge but may have as its primary aim to join a_1 and a_3, for example creating a new material that combines these two chemical elements together for the first time. A new drug compound may be viewed from this viewpoint, with two or more thereapeutic agents bound together in a deliverable form. Innovation from this viewpoint is a combination of (i) identifying valuable bundles of attributes, and then (ii) exploring or experimenting to find a larger template of elements that can be created and that embeds the high value subbundle.

As strings get longer there are of course more and more possible combinations of subbundles and letters that are possible. It seems unrealistic that a very large number of new subbundle values will be realized, so that as elements become more complex a relatively small number of subbundles will be realized, typically a shrinking percentage of the total possible. As an example, in a very long string it may be difficult for the value contained in two letters that are placed far apart in the string to be realized - these elements are far apart in the final product and thus not so closely linked. In order to keep the number of subbundles stable I impose the condition that the only feasible new subbundles that can be co-created are those based on combinations involving parents of the main string and their parents (grandparents of the newly created string).

Figure 2a illustrates the creation of subbundles. In the figure parent strings $a_1 - a_2 - a_3$ and $a_3 - a_4 - a_5 - a_6$ are combined in a RL configuration. The first parent has been produced from what are now grandparents $a_1 - a_2$ and $a_2 - a_3$, and the second from grandparents $a_3 - a_4 - a_5$ and $a_5 - a_6$. In total there are eight potential subbundles: each parent with the opposite parent's grandparents (four), and each grandparent with the opposite parent's grandparents (four). One example is the combination of the two circled grandparents: $a_1 - a_2$ with $a_5 - a_6$. In fact the actual number of potential new subbundles may be lower for several reasons. If a subbundle is an exact duplicate of a preexisting field element or the main string it is being created with - meaning it is the same string with the same parents, created the same way – then it is not co-created since it cannot be created twice. If two subbundles are duplicates, then assuming the string defined by the two subbundles is viable just one copy is created. Lastly, if two subbundles duplicate one another but are created differently (different parents) then assuming the string they define is viable both subbundles are created but the value associated with the string they define is added to the total created value just once.[8]

An important assumption is that subbundles can be created without a conceptual overlap. Thus in Figure 2a the combining of grandparents $a_1 - a_2$ and $a_5 - a_6$ creates the new subbundle $a_1 - a_2 - a_5 - a_6$. Note that some combinations will have an overlap if they straddle the overlapping portion of the two parents, for example $a_2 - a_3$ and $a_3 - a_4 - a_5$ in the figure. Subbundles that do not share a conceptual overlap are not viewed as new strings that can be used to build further strings. But they do create value (if viable):

[8] Additional rules are (i) that two subbundles cannot be combined if one is a subset of the other; and (ii) if the concatenation already exists as an element in the field it is not re-created. Note that each subbundle has a fixed orientation in the parent string. I impose the rule that new subbundles can be created only with the orientation they occupy in the parent string, which is why there are just eight candidates.

their value is realized at the time the larger string that contains them is produced - they are incarnated or demonstrated through production of the larger string. Subbundles that do share a conceptual overlap are treated as new strings in their own right and can be used to build further strings.

The importance of allowing subbundles to have realized values is that individuals' intuitions guiding their creative development are often (and, specifically, in the model in this paper) about these smaller combinations that they then seek to find ways to realize. This is discussed in depth in 3.1.2 below.

2.3. Success Rate and Value Distributions

In this subsection I introduce model parameters governing the success probability for potential new elements and the value distribution associated with elements that are successfully created. Table 1 lists the model parameters referred to in this subsection and subsequent parts of the paper.

Not all potential new elements (that is, valid based on conceptual overlap) are viable. The fact that there is a conceptual bridge to connect two elements makes it possible that a new element can be produced, but does not ensure that this will be possible. For example, a pharmaceutical company may find a chemical bridge that in principle enables two molecules to be linked to produce a new drug, but the attempt may fail and the new drug cannot be synthesized. More generally, a firm may have developed a way to combine two elements to create a new product but that product may fail to be viable for a variety of reasons.

The probability of success of new combinations is a parameter denoted P_X, $0 \leq P_X \leq 1$. In some fields we expect the rate of success to be high, for example in more theoretical fields in which the logic of the combination is sufficient to guarrantee, at least in many cases, that the new element is viable. But in most fields, including empirical and experimental fields, P_X is likely to be closer to zero. When a new element turns out not to be viable its value is zero. In addition, no new subbundles it contains are created. For new elements that are viable, the value of the new element is drawn from a distribution. I specify two distributions, low or ordinary and high. Most elements have values drawn from the low distribution, but a fraction have values drawn from the high distribution. The probability an element has its value drawn from the high distribution is denoted P_H; the expected value for the high distribution is v_H times the value for the low distribution, the value of which is set from other constraints specified below.

The same distributions apply to subbundles. A subbundle can be created only if its parent string is viable. The event that the subbundle is viable is independent of the event that the parent is viable, and if it is viable its value and whether its value is drawn from the high distribution are independent of the value and high/low draw for the parent. In addition, the viability, values and high/low draws for any pair of subbundles are independent.

I specify the value distribution in a manner that is consistent with the empirical literature on valuation of innovations and creative output, especially patent citations and revenues and scholarly citations. It is widely accepted that the distribution of values associated with innovations is skewed with a long right tail, with a relatively small percentage of innovations generating high value. Silverberg and Verspagen (2007) in a careful analysis of several different datasets find that the main body of the value distribution in a variety of applications is well fit by a log-normal distribution but that the tail of the distribution is better fit by a Pareto distribution.[9] To capture this empirical regularity I specify a distributon with two parts that

[9] See also Scherer and Harhoff (2000), Gambardella, Harhoff and Verspagen (2008), Albarrán, Crespo, Ortuño and

are spliced together to create a single value distribution. The first, lower part of the distribution applies for ordinary creative contributions and the second upper tail applies for contributions that have unusually high value. I specify the distribution of values for the main lower part of the distribution to be log-normal consistent with the findings of Silverberg and Verspagen, specifically a truncated lognormal distribution with truncation point X_m. The mean μ and variance σ of the distribution govern how skewed the distribution is and together with X_m determine the mean and standard deviation. For contributions with unusually high value, I specify that the value is drawn from a Pareto distribution with cut-off point X_m and parameter α. Parameters are chosen to create a well-formed distribution consistent with other model parameters. In particular, μ, σ, X_m and α are chosen such that the overall cumulative probability is one, the cumulative probability associated with the Pareto portion is P_H, cumulative probability associated with the lognormal portion is equal to $1 - P_H$, the expected value for the Pareto upper tail is v_H times the expected value for the truncated lognormal, and the density function of the truncated lognormal at the point of truncation equals the density function for the Pareto at X_m, so that the two densities splice together in a continuous manner.[10] Figure 3 depicts the resulting value distribution for the base parameter values listed in Table 1.

3. Individual Creative Development

In this section I describe individual creative development and then outline the analysis used to solve for optimal choices individuals make over the course of their development, specifically about what elements to learn and which new element to try to create.

3.1. Path of Creative Development

Individuals enter the field in sequence. Each individual who enters the field engages in a process of creative development consisting of a series of four steps. First, he chooses a seed learning set from the existing elements in the field. The seed must consist of more than one element in order to generate intuitive signals as specified below; in the simulations it consists of two elements. Second, he gains intuitive signals about creative opportunities in the field, specifically bundles of attributes, from the seed elements he learns, described in more detail below. Third, based on his signals as well as the existing state of the field he chooses an additional set of elements to learn from the elements in the field. In the simulation this second learning choice also consists of two elements. Fourth, he chooses a new element to try to make from among the set of potential new elements he can make given his full learning set. This potential new element must be based on combining two elements he has learned that share a conceptual overlap. Finally, the outcome of his choice is realized. If the element is not viable there is no value created and no new elements added to the field. If the element is viable, it is added to the field and the individual realizes its value.[11] In addition, if the element is viable outcomes are realized for any subbundles co-created with the main new element: each subbundle is revealed as viable or not, and if viable it is added to the field and the individual accrues its value. However,

Ruiz-Castillo (2011).

[10] The relevant formulas are the mean and density for the Pareto and mean and density for the truncated lognormal. For formulas for the truncated lognormal see Söderlind (2012).

[11] If the string has been made before and is thus already in the field no new value is created, even if the individual makes it a new way, e.g., via a different pair of parent strings. Though it is unlikely, it is possible for an individual to choose to try to make a string that has already been made, if he believes that the subbundles that would be co-created are sufficiently valuable.

the subbundle element is available for further building only if it is itself an element formed by a conceptual overlap; otherwise it is part of the field, meaning that it is known to be viable and its value is known, but cannot be built on.

Individuals make choices with the objective of maximizing the expected value they will earn at the end of the process. In the base simulations individuals realize the value accruing from the parent string they attempt to make if it is viable, and, if the parent is viable, all viable new subbundles. I also present results, for one case, for a context in which in addition to these direct benefits an individual also earns a royalty if the individual in the field in the next period uses the element the current individual created this period as a building block to create another new element.

3.1.1 The Learning Set

Limited learning capacity is an important constraint in creative development. As a field grows it comes to contain a large number of elements, and no individual can learn everything, especially not with the degree of understanding required to build creatively with an element. The choice of what to learn is thus critical, both enabling as well as limiting opportunities to create new elements.

Limits to learning can arise and be imposed through various mechanisms. One approach is to specify a cost for each element learned, possibly linked in some way to its complexity, which in this model is measured by length; each individual then optimizes subject to a total learning budget. Rather than complicate the model to that degree, I impose the constraint that an individual selects and learns a fixed number of elements in each learning cycle. Since the length of selected elements may increase, this means that learning becomes more efficient as the field matures, in the sense that larger knowledge chunks or in this case longer strings are learned.[12]

A second related issue is which components an individual learns when she selects an element. In particular, in addition to the main element, it must be specified which if any of its component subbundles/substrings she is able to learn. I assume that an individual learns, in addition to the element itself, the two parents that were combined to form the element, but no additional elements beyond this. This assumption may not seem warranted or even needed. It might seem simpler to assume that an individual learns every component substring when she chooses to learn an element. However, as elements in the field become longer, they contain more and more substrings/subbundles - each parent has grandparents, these grandparents in turn have parents, and so on, an upward stretching tree of substrings. Thus the assumption that all substrings/subbundles are able to be learned would imply that as elements become longer (more complex) an individual could learn a very large number of elements from selecting just a single main element. Given that learning capacity is limited this is not a viable assumption – allowing an individual to learn all or many components of a single element would effectively circumvent the learning capacity constraint. This is one of several issues in which the non-stationary nature of the model of the field has important implications that the model assumptions must address. The assumption I make preserves a degree of stationarity, in that the number of elements learned remains at three regardless of how long elements become. However,

[12] The model can be interpreted as the case in which the economies of learning for each element are large so that the full string is learned for a fixed cost. Even if a more explicit learning-cost model is employed the cost of an element need not be linear in its length, which implies no economies of learning, but can be a function of length that allows for some economies, but less than the model used here.

the length of these learned elements may increase over time, reflecting greater efficiency in the way in which knowledge is packaged into blocks.[13]

Aside from learning capacity constraints the ability to deconstruct a given string and learn its component elements is undoubtedly limited. While in some cases, like a movie or simple toy, these components may be accessible, in many cases, such as new composite materials or engineered products or food they may not be directly accessible or may have been transformed in such a way that they cannot be extracted (or reverse engineered) and learned. I make a simple assumption that allows for some component learning, but limited, and, as noted above, preserves a degree of stationarity as the field develops. Parents would seem to be the most accessible components since they have been used to construct the main element (although, again, in some applications they may not be accessible). Other elements are a smaller part of the whole and more steps removed from the final element, thus less likely to be attended to or able to be learned.[14] Note that every component substring has been produced previously and therefore is part of the existing field and can be directly selected into the learning set.

3.1.2. Intuitive Signals

Based on their seed elements individuals gain signals about the value of potential new bundles of attributes. These signals in turn guide their subsequent choices about further elements to learn as well as what new element to try to make. In this formulation an individual develops intuitions about potential new bundles based on what he learns, imagining new combinations among the elements he has learned.[15]

Specifically, signals are associated with the bundles formed by concatenating an element from one seed element with an element from the other. There are three learned elements associated with each seed - the seed itself and its two parents. There are thus nine possible pairings (there can be fewer if there is duplication). Each such pair b_{1i} and b_{2j} can be concatenated in four different ways - LL, LR, RL, and RR, as described earlier for conceptual overlaps, treating bundle order as relevant, an assumption I maintain. Hence there are as many as 36 potential string concatenations that have signals associated with them for a given seed set. I denote these as *subbundles* consistent with the language in section 2. Importantly, these subbundles are not strings formed by conceptual overlap. Rather, they are bundles of elements that can generate values as subbundles of larger strings that are created.[16]

Each subbundle that has associated signals generates two signals. One signal provides information about the likelihood that this subbundle is viable, which in this context means the subbundle will generate

[13] An even more extreme assumption is that only the main element is learned. Were this assumption imposed, in order to allow individual to develop rich enough learning sets to have the potential to generate a number of intuitive signals and allow for the attempt to produce a number of new elements, it would be necessary to allow individuals to select more elements to learn. This model would allow more carefully refined choice of which elements are learned but, depending on the number of elements individuals were allowed to choose, would lead to learning sets of similar size (learning three elements in this case is comparable to learning one element in the model used).

[14] Indeed as the number of components becomes large it is very unlikely that most of the smaller and more removed substrings will be accessible just by learning the main element; reading an entire book one is unlikely to attend closely to and learn every sentence and phrase.

[15] More broadly, intuition can also be rooted not only in this kind of learning but also in experiences; however in the model in this paper there is no experience outside of what is learned.

[16] There are two kinds of concatenated strings that do not have signals associated with them as I formulate the model. An ordered bundle that has already been produced as a string with the same elements in the same order does not have any signals associated with it, since its value is already known. And signals are not associated with strings for which the two parents share a conceptual overlap on their facing ends as these two elements can potentially be produced as a new main element. It is not difficult to extend the model to allow signals to be generated for such conceptual overlap strings; I choose not to in order to keep the model slightly more streamlined.

economic value. The other signal provides information about the likelihood that the value associated with the bundle, given that it is viable, is drawn from the high distribution.

The fact that signals are associated with subbundles, not strings formed via conceptual overlap, is an important feature of the model of creative development in this paper. Due to this feature the main way signals enter into creative development is through being associated with bundles of elements/attributes that can be created *only* as a subbundle embedded in a larger string. The logic motivating this feature is that intuitions guiding creative development are in most cases not about complete products, with every detail worked out, but about possibilities that are not fully formed, but only partially imagined. Referring to the Lancaster model of attributes, one may have an intuition that a certain bundle of attributes, if combined (here in a specific arrangement) will have value. One then seeks out elements containing these attributes that fit together to create a viable full product. The product undoubtedly contains many additional elements, various "connectors" used to embed the attribute bundle one believes has value. But much of the value of the final product in fact derives from the specific combination of attributes one envisioned. Creative development, from this point-of-view, is about intuitions about simpler combinations, then searching for building blocks within which these smaller bundles are embedded such that the building blocks can fit together and embed the simpler combination.[17] Creative development starts from broader imagined possibilities and involves the search to develop this broader intuition more specifically and completely, ending in a workable final creative product, whether that be a tangible product, a theory, a model, a strategy, a song or narrative, or some other kind of cultural contribution.

Imagine if instead of the model I propose here intuitions provided information about the value of a complete new string element and the exact parent components needed to produce it via conceptual overlap. Creative development would then be focused solely on searching for these components enabling the element to be constructed exactly as envisioned. The fact is that most intuitions in the course of creative development are not this sharp, but leave room for different approaches for how an intuitively valuable combination will be realized. Thus, an engineer does not imagine every detail of a new product when he first conceives of its possibility. Rather he imagines certain critical features that will be incorporated. Then as the product is developed all details are worked out in such a way that the features fit together within a feasible product that typically includes additional elements used to combine the critical features in a single product.[18] Though it might seem simpler as a modeling strategy, it is simply not correct to reduce creative work to a by rote search for a fixed set of elements aiming at a fixed final product. One must leave room for the ongoing process of exploration and adaptation as one moves from initial conception, often just a partial vision, to a complete final product. The model in this paper allows for this, incorporating intuitions into creative development in such a way that they guide further choices, but leave room for additional learning, experimentation, and discovery.

A related implication of the model is that the intuitions guiding creative development are often at

[17] It is possible for a string that has been associated with a signal via concatenation to be produced via conceptual overlap through combining two different parents. In this case the signal has provided information about a final main element that is produced, thus the model allows for this possibility. However, because the main element is produced through a different combination, even in this case while the signal provides information about an envisioned final product it does not provide information about the mechanics of how to create it.

[18] As a second strategy that also fits the model in this paper, even in case he does envision a full set of final attributes he still may not know which components can be combined to produce this set of attributes - there may be different pairings, involving different subbundles being co-created with the main product.

a more abstract conceptual level than the specifics of an exact product. In the knowledge representation framework a short string or smaller bundle of elements refers to a more basic concept, whereas a longer string embedding this string has added more letters, refining the basic concept. For example a short string might define a simple circuit or chemical compound, and a longer string that incorporates this shorter string defines a more complex circuit or compound; the additional letters serve to refine and specify more details. China is a shorter string, hence broader topic; Beijing is refined one step, the capital city of China; Beijing population is refined another step; and Beijing population growth is refined yet another step; each step adds another qualifying attribute (the attribute may itself be either a letter or string made of more basic attributes). Thus an intuition about a relatively short string is about a relatively broad concept or conceptual combination. An example is an intuition that it will be fruitful to link a theoretical or statistical framework with a new area of application. A researcher may believe that developing a project doing this will be fruitful. His intuition however is not so fully developed that he knows exactly what dataset he will use, what the exact theoretical, statistical model will be within in the broader family that fits with the dataset he ends up using, thus what exact model he will end up estimating. Rather, he searches for a specific model and dataset that he is able to fit together and that incorporate the broader link he envisions. Different datasets coupled with specific models will generate a certain expected creative potential as he perceives it, in terms of feasibility and the distribution of possible results of the research project, and he searches for the best (highest expected value) combination. The principle that creative development is guided by broader conceptual interests and intuitions is developed and demonstrated with many examples in *The Nature of Creative Development* (Feinstein (2006), especially Chapters 2 and 3). Among many examples that are discussed there are Walt Disney's initial idea for making animation films based on animal characters (see also Mosley (1992)), Matisse's early sense for how to combine vivid, strongly contrasting colors cutting across boundaries of form (see also Spurling (1998) and Matisse (1990)); John Maynard Keynes' initial thoughts about expectations and economic fluctuations (see *The Nature of Creative Development*, Chapter 15, Skidelsky (1983), Keynes (1977, 1981)); and Benjamin Franklins's initial interest in experimentally exploring the properties of electricity (see also Cohen (1990)). In each of these cases the individual began from a conception of a broader conceptual combination, or relatively short bundle of elements, and developed creative products that incorporated this intuition, in the case of Walt Disney and Matisse art, in the case of Keynes an economic theory, and in the case of Franklin specific experiments including his famous lightning experiment. This is thus a quite general pattern for creative development.

Note finally that the null case in which an individual has no intuitive signals to guide him is also included in the model, referred to as *clean* runs. In this case the individual will consider different final elements he could make and choose the one that has associated the greatest expected value, including all potential subbundles that may be produced with it, working solely from public knowledge.

Figure 2b illustrates how a subbundle associated with signals is embedded in a larger string. In the figure seed element one contains block b_{11} and seed element two contains block b_{21}, each of which refers to a grandparent of the main string. Note that these two blocks do not share a conceptual overlap; also note that they do not form a contiguous substring in the larger string - I do not impose this restriction.[19] The subbundle

[19] As discussed above and depicted in figure 2a I restrict subbundles to combinations of parents with grandparents and grandparents with one another.

formed by concatenating b_{11} with b_{21}, string $a_1 - a_2 - a_5 - a_6$, is a possible signal-generating subbundle since its two blocks come from different seed elements. If the main string is attempted as a project and turns out to be viable then this subbundle will also be formed if it is viable. Thus if an individual receives a positive signal that this subbundle is likely to be viable, that may encourage him to try to make the main string, since if the main string turns out to be viable the subbundle is likely to be co-created, adding additional value. Indeed, following on the discussion above, it is here that intuition guides creative development: the value that may be (and is likely with a positive signal) to be created by the subbundle may be the main driving factor behind the decision to make the full string.

The individual gains signals about m oriented concatenations from his seed set (recall there are up to 36 in this pool). In the simulations m is set to two. The concatenation subbundles for which he gains signals are chosen at random from among the set of valid signal concatenations. For each such subbundle he receives two signals, a signal about the viability of the subbundle, essentially whether or not it proves to be a valuable bundle, called the X signal, and a signal about the likelihood that the value of the subbundle if it has value is drawn from the high distribution, the H signal.

The X signal is either 1, a signal that viability is likely, or 0, signaling that viability is unlikely. Signals are not perfect: The false positive rate for an X signal is s_0^X, and the true positive rate is s_1^X. Likewise the H signal is either 1, indicating it is relatively likely that the string value is drawn from the high distribution, or 0, meaning it is relatively unlikely. The false positive rate for the H signal is denoted s_0^H and the true positive rate s_1^H.

For a given concatenation string the X and H signals are independent (the H signal is however relevant only if the string is viable). Further, for any two distinct subbundle concatenations their respective signals are independent. If a given subbundle can be made via concatenation two different ways, then in any case in which signals are generated for each of two ways of making it the pair of X signals generated are conditionally independent conditional on whether or not the subbundle is viable (recall that any string including concatenated subbundles has a fixed value the same regardless of how it is made), and likewise the pair of H signals are conditionally independent conditional on whether or not the subbundle value is drawn from the high distribution. I provide formulas for learning rules covering these cases in the Appendix - see also the discussion below in Section 5.

3.1.3. Discussion

The model of creative development focuses on learning and gaining intuitive signals from what one learns. Learning is inherently selective, since in any field there is simply too much for an individual to learn everything, certainly not with the level of comprehension required for creative work. In the model used here this is presented as choices from the set of elements already created in the field.[20] In general learning occurs in stages, not all at once, so that as an individual learns and gains intuition from what he has learned, this influences his subsequent learning choices. In the model in this paper this is modeled in the simplest possible way, as a two-step learning process.[21] It would be a valuable extension to allow learning in more stages, so

[20] More generally, this can involve choices that set up learning opportunities, for examples choice about which classes to take or seminars to attend, and more broadly what experiences to engage in.

[21] It is also the case that initial learning is broader and more basic, and later learning more focused and advanced. Jones

that individuals build up larger conceptual learning sets, and explore the impact on the development of the field.

In creative work individuals must make committments about which project to pursue since projects are costly and individuals cannot pursue all projects they may imagine. Here I make the simple assumption that an individual can pursue only one project, a project being a new element the individual tries to make via conceptual overlap. It is straightforward to generalize the model to allow an individual to pursue more than one project and put out into the field the best one. This has implications for creative development since when individuals can pursue more projects they are more likely to pursue projects with high tail values but lower overall probability of success. The restriction to focusing on just one project incorpoates in a simple way the important point emphasized by Weitzman that it is costly to pursue the development of ideas, so that even though many potential combinations are available, with the number increasing dramatically as the field grows, only a relatively small number and smaller percentage over time can actually be pursued as projects.

3.2. Analysis of Creative Development

Individuals make choices so as to maximize the expected value of the total value created by their final creative project. The optimal strategy for an individual must specify: (i) which seed set he selects; (ii) for each possible signal pair that may be generated based on the seeds and each signal draw for that pair of signals: (a) the additional elements he chooses for his full learning set; and (b) the new element he chooses to try to make as his creative project. Figure 4 depicts this sequence.

3.2.1. Seed Signals and Strategy

The first decision an individual makes is the selection of seed elements. Assuming there are N_t elements in the field at the beginning of period t, and the individual entering in period t selects two, there are $N_t(N_t-1)/2$ possible seeds. For each possible seed, the individual computes the expected value associated with the optimal strategy if he chooses this seed; he then selects the seed with highest expected value.[22]

To compute the expected value associated with a given seed the individual first determines the set of valid signal concatenations between elements in the seed learning set, as described in the preceding subsection. He will receive signals from m members of this set drawn at random; I set $m = 2$ for the remainder of the discussion to make the formulas simpler. If there are m_{ts} valid concatenations in the set for seed s, there are then $m_{st}(m_{st}-1)/2$ possible signal combinations, each represented by a branch on the signal generation random event node in Figure 3. Each branch is equally likely, hence has probability $\frac{2}{m_{st}(m_{st}-1)}$.

The individual then computes the highest expected value he can gain following his optimal strategy for each one of these possible combinations. The steps involves in this calculation are outlined in the following subsections. He then averages these values to compute the expected value associated with this seed.

(2009) argues that with the growth and deepening of knowledge in a field individuals must be more specialized in what they know, and traces the implications for the development of the field including age of first invention which may be taken as linked to the learning model of this paper.

[22] Although any elements can be selected, only elements created via conceptual overlap can be used to build new strings, thus in most cases individuals will choose these kinds of elements. Because seed elements also generate signals it is possible for the optimal choice to include a non-overlapping field element; however this is very unusual.

3.2.2. Signal Probabilities and Probability Updates

Each signal concatenation has four possible signal outcomes: $X = 1$ and $H = 1$; $X = 1$ and $H = 0$; $X = 0$ and $H = 1$; and $X = 0$ and $H = 0$. Signals are generated in pairs for $m = 2$ hence there are 16 different possible branches, shown in Figure 3 for one illustrative case.

Denote the prior probability that string s is viable by $P_X(s,t)$. I use the terminology "string" because these formulas apply to both main strings and subbundles or concatenated strings. Likewise denote the prior probability that the string has its value drawn from the high distribution by $P_H(s,t)$. In this notation t is the time period. The reason this notation is used is that string probabilities are updated after each period based on what is observed about the behavior of the individual who worked in the field. The update formulas are given in Section 4.2.

Define $psig_X(s,t)$ to be the probability that the signal $X = 1$ is generated for concatenated string s and $psig_H(s,t)$ to be the probability the signal $H = 1$ is generated. These probabilities are:

$$psig_X(s,t) = s_1^X * P_X(s,t) + s_0^X * (1.0 - P_X(s,t)) \tag{1}$$

$$psig_H(s,t) = s_1^H * P_H(s,t) + s_0^H * (1.0 - P_H(s,t))$$

The calculations of the probabilities associated with each set of signals is now straightforward. When the two concatenated strings for which signals are generated are different, the signals for the first string are independent of the signals generated for the second, making the calculation an easy set of multiplications. As an example, for a pair of concatenation signal generators s_1 and s_2, the probability that the signals for the first string are $X = 1$ and $H = 1$ and the signals for the second pair are also $X = 1$ and $H = 1$ is:

$$psig_X(s_1,t) * psig_H(s_1,t) * psig_X(s_2,t) * psig_H(s_2,t) \tag{2}$$

The case in which the two strings s_1 and s_2 are the same string (made two different ways) involves somewhat more complex formulas provided in the Appendix.

After an individual receives his signals he updates probabilities for the associated concatenation strings. Updating is done using standard Bayesian formulas. If the individual receives a signal $X = 1$ for the string, his revised probability that it is viable is:

$$P_X(s,t\ |X = 1) = \frac{s_1^X * P_X(s,t)}{psig_X(s,t)} \tag{3}$$

If the individual receives the signal $X = 0$ his revised probability that the string is viable is:

$$P_X(s,t\ |X = 0) = \frac{(1.0 - s_1^X) * P_X(s,t)}{(1.0 - psig_X(s,t))}.$$

Similarly if he receives the signal $H = 1$ his revised assessment that the string's value is drawn from the high distribution is:

$$P_H(s,t\ |H = 1) = \frac{s_1^H * P_H(s,t)}{psig_H(s,t)}$$

while if he receives the signal $H = 0$ his revised assessment is:

$$P_H(s,t\ |H = 0) = \frac{(1.0 - s_1^H) * P_H(s,t)}{(1.0 - psig_H(s,t))}.$$

When the two strings for which signals are generated are the same string (made two different ways) the formulas are more complex and are again provided in the Appendix.

In general the most likely set of signal values among the 16 is the one for which all signals are zero, due to the fact that for the parameter values used in the simulations the majority of strings are not viable and the likelihood of having a value drawn from the high distribution is low.[23] In this case an individual's signals act as negative information, and may lead him not to choose a full learning set he would have chosen if he had not received any signals or not to try to make a new element that he would have tried to make if he had received no signals. Although this is the most common occurrence and therefore important for how the field develops the more interesting cases are those in which an individual receives at least one signal that is a one, in which case he may well be led to make choices such that he attempts to make a new element that has the string with which the signal is associated as a subbundle. This fits the commonsense view that individuals are guided towards elements that they believe have creative potential.[24]

In fact much of the time a signal has no import, in that the subbundle it provides information about cannot be created given the current state of the field. I call such a signal a *clean signal*. As the field grows fewer signals are clean, so that more signals have import and creative development is more likely to be influenced by signals. As a related matter, an important benchmark of the model is what I call a *clean* run, in which individuals receive no signals to guide them. It is interesting to compare how the field is projected to develop in this case versus when individuals do receive signals - I explore this comparison in the simulations.

3.2.3. Full Learning Set; New Element Set.

An individual chooses two additional elements from the field to complete his learning set. Given that two elements from the field have been chosen for the seed, he chooses from among the $N_t - 2$ remaining elements. There are thus $(N_t - 2)(N_t - 3)/2$ possible sets.

Given the choice of the full learning set, the individual determines all possible new elements he can make via conceptual overlap from among the elements in his learning set; this is called the *new element set* (note that some of these may be made strictly out of elements from the seed). There are up to 12 distinct elements in his learning set (there can be fewer if some elements are duplicates), and each pair can be combined in four different ways, thus there are up to $66X4 = 264$ potential new elements. Most of these combinations do not share conceptual overlap. Further, among those that are valid new combinations some duplicate elements already in the field. Thus the actual number of elements in the new element set is in general well below this, typically no more than a few dozen, in some cases none - but the individual will presumably not make choices for which this is the case unless the field does not provide any opportunities to make new elements, in which case a stopping point has been reached.

3.2.4. Optimal Strategy: Calculation of Expected Value.

[23] This is true initially. However as the field develops some strings may have updated prior probabilities that sharply increase the probability that they are viable or the probability that they have a value drawn from the high distribution, overturning this general tendency.

[24] Another version of the model has individuals continue to draw signals until they receive at least one positive ("one") signal.

An individual chooses one element from her new element set to attempt to make as her creative project. To make this choice she computes the total expected value associated with each potential new element, including the values of all potential subbundles, then chooses the element with highest expected value.[25] If the element turns out to be viable then it is created along with all viable subbundles, and the total value is realized and accrues to her. As noted above and discussed in section 7 this base model can be extended to incorporate royalty payments for any strings the individual creates that are used in subsequent periods.

The expected value associated with an element includes the expected value of the element itself as well as the expected values for each subbundle associated with the element that may be co-created with it. These expected value calculations are based on the probability assessments the individual makes that a given element is viable and the likelihood that its value is drawn from the high distribution. For a subbundle for which the individual has not received any intuitive signals these probabilities are her prior assessments, based on the history of the field, $P_X(s,t)$ and $P_H(s,t)$. For a string for which she has received signals the probabilities are based on the posterior probabilities given in section 3.2.2. If two subbundles duplicate one another then if one has signals associated with it but not the other, the standard signal probability formulas apply; if both copies have signals associated with them then the more complex formulas given in the Appendix apply.[26]

Assuming a string has probability $P^*_X(s,t)$ of being viable (the asterisk is used to denote that this is the posterior probability after all signal updates) and probability $P^*_H(s,t)$ of having its value drawn from the high distribution, the expected value associated with the string is:

$$P^*_X(s,t) * \exp(\mu + 0.5 * \sigma^2) * (P^*_H(s,t) * v_H + (1.0 - P_H(s,t))$$

The second term in this expression is the formula for the expected value of the basline log-normal value distribution with mean μ and standard deviation σ. The third is a weighted average of two terms: the first is the probability that the value is drawn from the baseline distribution; and the second is the probability that the value is drawn from the high distribution, multiplied times the multiplier v_H that applies in this case. This formula applies to main elements and also to any subbundles that may be co-created with a main element. Because an individual can choose only one element to try to make from the new element set, correlations in outcomes between elements in this set do not need to be evaluated.[27]

Given the calculation of the optimal new element to attempt to make and its expected value, the individual rolls back to compute the optimal full learning set given the signals he receives, then averages out over all possible signal combinations to evaluate the expected value associated with a given seed. Finally, he rolls back again to determine the optimal seed. In turn, his choices map into his optimal strategy to follow: his strategy specifies the seed he chooses, then for each possible signal combination, which elements to choose to complete his learning set and which new element to try to make.

[25] It is possible to extend the model to the case in which the individual is given the opportunity to try to make more than one element. In this case she may either produce and realize the values associated with all elements she tries to make that turn out to be viable; or may learn the total value associated with each element (including the values associated with all substrings that will be co-created) and produce the element that has highest total value.

[26] It is possible for more than two substrings to refer to the same string. However since at most two can have signals associated with them the formulas for updated probabilities are the same as given in the previous case for two substrings that refer to the same string.

[27] Correlations may arise when two main elements share substrings that are concatenated subbundles with signals in common.

4. The Development of the Field

One individual enters the field each period and lives for just the single period. The individual determines his optimal strategy as described in the preceding section. If the new element he attempts to make is viable it is added to the field along with all associated new viable subbundles. The main element that is added together with any subbundles formed through conceptual overlap become new potential building blocks to create further elements. Thus the field grows over time and the number of building blocks to create new elements also grows.

Since each element is made through combination of two preexisting elements, the field has a structure resembling a lattice. Actually it more closely resembles a semi-lattice, assuming a single unitary element that then divides into the N attributes.[28] I do not explore the possible lattice properties of the field in this paper but this is a potentially fruitful line of investigation for future work.

In general the field can grow without bounds. It is in theory possible to reach a position in which no new conceptual overlap combinations can be made, if a series of new elements that are attempted turn out not to be viable, in which case the field growth ends. However this is very unlikely especially once the field begins to grow beyond its initial conditions. Rather, in general the number of potential new elements grows over time. The field grows both in terms of depth or complexity, as new longer elements are created, as well as in breadth as new elements are added of similar lengths to elements already in the field. Overall, the nature of growth is non-stationary in that new elements become longer with no bounds to how long they can become. It is possible to work out analytic formulas for how the field develops in simplified settings. One simplification is to assume that given the current state of the field every feasible new combination that can be made, based on the rules for overlapping above, is attempted. In this case, when the viability probability is 1, so every feasible new element is viable, the formulas are especially simple and are provided in the footnote to this sentence; when the viability probability is less than one more complex combinatoric sums are involved.[29]

4.1. History Updates

The individual who enters the field in period t is assumed to be able to observe the history of the field and the choices made by previous individuals who worked in the field. I assume specifically that the individual observes the following for each prior individual: the seed set the individual chose, the full learning set he chose, and the new element he tried to make and whether it was successfully made or not.[30] I assume the individual does *not* observe the intuitive signals prior individuals received - signals are treated as private information not shared through public information sources.

[28] It is easy to see that the lattice is modular since every element has two parents.

[29] When the viability probability is one, the formulas are as follows for the case when the initial state of the field is a ring as defined above. Let N be the number of attributes as above and let l be the length of the produced new string. For l even the number of new strings that is produced is $N2^{l-2}$ and for l odd the number of new strings is $N(2^{l-1} - 2^{l-2} + 2^{(l-3)/2})$. One can also take l to be the generation number - double overlaps are subsumed in the formula. I thank Sharon Qian for working out this formula.

[30] Similarly if more than one individual works in the field in a period we can assume that for each individual the seed and learning set choices he made and the new element he tried to make are known. The formulas in this section can then be worked out in an analogous manner; however if there are multiple equilibria then issues of equilibrium selection individuals make can influence the probability updates made.

Using the information he has access to the individual forms updated probabilities, for each possible new element he may make, including subbundles, of the probabilities the element is viable and has an associated value drawn from the high distribution. Above I have denoted these updated probabilities $P_X(s,t)$ and $P_H(s,t)$ for string s. Note that they are the probabilities prior to any signals the individual receives, which prompt further updates for certain subbundles. The updating can be worked out as recursive formulas, meaning that the update probabilities computed by the individual who entered the field in period $t-1$ serve as the prior probabilities for the updating by the individual entering the field in period t. Thus I give the formulas for a single period or round of updating, in particular period t.

The seed choice made by the individual who worked in the field in $t-1$ could have been exactly predicted based on the $t-2$ period choices and outcomes since individuals entering the field do not have any private information. However, the full learning set the individual chose in $t-1$ and his selection of which new element to try to make in general depend on the signals he received. Further, more than one set of signal combinations can lead the individual to choose the given full learning set he chose and new element he chose to try to make. Hence updating is done based on pooling all signal combinations that would have led the individual to choose the full learning set and new element to make that he did in fact choose. To simplify exposition I denote this pool of signal combinations that is consistent with the observed choices of the individual in $t-1$ by *pool*. Let P_{pool} be the total probability summed over all signal combinations that fall in the *pool* (as compared with the full set of signal combinations, for which the probability sums to one):

$$P_{pool} = \sum_{i \in pool} P(sig\ combo\ i)$$

where $P(sig\ combo\ i)$ is computed based on equation (2) in 3.2.2.

Updating proceeds in two steps. In the first step probabilities are updated for all subbundles that have signals associated with them in the *pool*. For each such subbundle s, determine the set of combinations in the *pool* for which the signal received in regards s was a 1 (set 1), the set of combinations for which the signal received was a 0 (set 2), and the set of combinations for which no signal was generated for s (set 3). Let q_1 be the weighted probability of the first set relative to the overall pool probability:

$$q_1 = \frac{\sum_{i \in pool} Ind(sig\ combo\ i\ has X = 1\ for\ string\ s) P(sig\ combo\ i)}{P_{pool}}$$

In this expression the numberator uses an indicator function to identify which signal combinations fall in the set 1 and weights each such combination by its probability. Similar expressions hold for q_2 and q_3 defined for sets 2 and 3. The updated probability that s is viable is then computed as:

$$P_X(s,t\ |X=1) * q_1 + P_X(s,t\ |X=0) * q_2 + P_X(s,t) * (1.0 - q_1 - q_2)$$

The first two terms in this expression use the formulas for updated probabilities based on signals from equation (3) in 3.2.2, and the last term uses the prior probability for s since this refers to signal combinations for which no signal was generated for string s. Comparable formulas hold for updates for the probability that the value associated with s is drawn from the high distribution. When two signals in a given signal combination refer to the same string the formulas are more complex and are provided in the Appendix.

In the second updating step probabilities are updated for the new element that the individual working in the field in $t-1$ tried to make and for all associated subbundles. Here there are two cases. When the new element was successfully made, its viability probability is updated to 1.0 and since its value is now known the probability its value is drawn from the high distribution is no longer relevant and can be discarded. In this case all subbundles are revealed as either viable or not. If a subbundle is viable its value is revealed, and if not its value is set to zero. In either case the probability its value is drawn from the high distribution is no longer relevant. In the other case the new element was not successfully made. Its viability probability is thus set to 0.0; its value is no longer relevant since it is not viable and the probability its value is drawn from the high distribution is thus also no longer relevant. In this case nothing is learned about the subbundles associated with the new element: since the new element was not made, there was no opportunity for the subbundles to be co-created with it. Hence the probabilities associated with the subbundles remain as they were after the first update step.

5. Model Solution - Simulation

I analyze the model through extensive computer simulations employing the analytic framework and expressions given in the preceding sections. The analysis is performed by a suite of FORTRAN programs, and much of it has been done using the Yale High Performance Computing System which enables access to as many 700 processors at a time. In this section I outline the approach. Table 2 provides a flowchart overview of the simulation procedure.

There are a few general features of the analysis. Simulations are done based on N letters. The cases $N = 3, 4, 8$ are simulated. For each N case three simulation runs are conducted in order to provide a sense of diversity of possible patterns of development of the field. In addition three companion *clean* runs are conducted for which there are no signals. For $N = 3$, I run three additional simulations for the case of a high viability probability and three further simulations for the case of a royalty payment. In all versions of the model a run begins with the initial state of the field and runs for T time periods. All paths that might occur (based on signal realizations and optimal choices) are computed thus determining the full distribution of possible paths of development of the field given initial conditions and parameter values. For the results in this paper T is set at 5 the largest number of time periods that is practical to evaluate with available computational resources, which are in fact quite substantial. For the base model and high viability probability sensitivity analysis individuals earn just the value that accrues at the time they create an element (including all co-created subbundles), thus the model can be stopped at any time point without having to extrapolate into the future. The royalty model, for which individuals must forecast the likelihood a given element will be employed in the future, is considerably more computationaly intensive and is discussed in section 7.

For each simulation run I generate a *masterlist* which contains fixed values for all elements including all subbundles that may potentially be created in the field, extending through the last period of the simulation, including information for each element about whether it is viable and whether its value is drawn from the high distribution. The masterlist is generated separately from the analysis of individual strategies, as described below, and thus holds all actual values fixed across all paths of development of the field for

the given simulation as well as the companion clean run.[32] Values are taken to be objective in that they are fixed, everyone agrees they are fixed and, once an element is created, everyone agrees on its value. Of course in many creative fields value may be more subjective: fixing values enables exploration of how a field develops without introducing subjective valuations.

Define a *node* to be a given state of the field, defined by the set of elements that have been created and the order in which they were created as well as the seed and full learning set choices made in each preceding period, thus encompassing the public history of the field. As described in earlier sections each *node* typically branches to a number of nodes next period since different signal combinations lead the individual working in the field to make different full learning set choices and different choices about what new element to try to make. Thus the number of *paths* in the field grows with each generation; for $T = 5$ there are typically several thousand paths.

Initial State

The initial state of the field consists of N_0 elements. To facilitate comparison across cases with different numbers of letters N I specify a "standard" initial condition which is a simple ring consisting of N elements. Each element in the ring is of length 2 and links two adjacent letters together with the final element in the ring linking the last letter to the first. Thus for $N = 4$ the initial ring consists of $a_1 - a_2, a_2 - a_3, a_3 - a_4, a_4 - a_1$. This ring structure ensures that all letters can be combined with all other letters, which may not be the case for other initial states. Note that since these elements consist of only two elements they cannot be created via conceptual overlap; I assume they have been created in a neighboring field, perhaps as subbundle byproducts. The one exception to the ring structure is $N = 3$. In this case with just the three ring elements there are no signals that can be generated in the first period, for any seed, that provide information about a subbundle that can actually be produced in the first period, thus all signals are *clean*, a degenerate case. To ensure there are non-clean signals in period 1, I add one additional element to the initial state for $N = 3$, $a_1 - a_2 - a_3$, so that $N_0 = 4$ and the initial state consists of: $a_1 - a_2, a_2 - a_3, a_3 - a_1, a_1 - a_2 - a_3$.

Steps of Analysis

The analysis consists of five basic steps shown in Table 2. The first step specifies the initial state, as described above, and a masterlist with values for the initial elements and all elements that can be created using the initial elements.

The next three steps are recursive, performed for each time period. Step 2 identifies all nodes in the field for the current time period. Except for time period 1, each node has a parent node associated with it. Each node also has a history. The history is based on the history of the parent, updated based on what happened at the parent node leading to this node, as described in the preceding section. Note that typically there are multiple nodes that share a common parent node but differ in the new element that was attempted and possibly in their seed and full learning set choices. The list of field elements associated with the node is then built: this is based on the parent node list and any new elements that have been created including the main element that was attempted if it is viable and, if it is, any associated viable subbundles. Finally,

[32] In theory it is possible for the clean run to generate elements not covered by the signal simulation, but in practice this does not happen as the clean run involves many fewer paths and elements.

the *masterlist* is extended based on the set of new field lists. This is done by going through each current period node in turn, and based on the field list associated with the node determining all new strings and subbundles that can be created utilizing pairs of list elements and parents of list elements, and adding all such new elements that are not already on the masterlist (as updated through the preceding node) to the masterlist. For each new element that is created values are drawn, based on model parameters, to determine whether or not it is viable, whether or not its value is drawn from the high distribution, and its value if it is viable.[31] When an element is created via conceptual overlap all subbundles that may be co-created with the element are listed as part of the entry for the main element in the masterlist, with a link to their address in the masterlist. Note that for a given element the probabilities it is viable and has a value drawn from the high distribution depend on the history of the field and therefore may be different for different nodes; these probabilities are thus generated locally for each node based on the path leading to the node and are stored locally with each node, whereas the masterlist itself is global and unified across all nodes, its listed probabilities are strictly the baseline parameter values.

Steps 3 and 4 calculate optimal learning and projects for each node in the current time period. Step 3 computes the optimal seed. This follows the logic outlined above. All possible seeds are evaluated. For each seed the program identifies all valid elements that can be used to build new elements (the seed elements themselves and their parents), then all combinations of these that have signals associated with them following the protocol in 3.1.2. For each valid signal combination, for all possible signal draws (there are 16 in general as described above in 3.2.1) the optimal full learning set and optimal new element to try to make are identified and the expected value is computed. *Clean* signals as defined in 3.2.2 all share the same choice of optimal full learning set and optimal new element to try to make, which can significantly reduce computational burden. Rolling back, the expected value associated with the seed is computed and the seed with highest expected value is identified as the optimal choice at this node; if more than one seed has the same highest value a mixed strategy is assumed so that each seed is chosen with equal probability. Note that this calculation is entirely analytic. Step 4 identifies the optimal strategy and outcomes for the optimal seed identfied in step 3. For each valid signal combination associated with this seed the optimal full learning set and new element to try to make are identified. The *masterlist* is then used to identify whether the new element is viable, and, if so, which associated subbundles are viable and are co-created with it. Next the full learning set and new element choices jointly are partitioned into sets - each set contains a unique combination of full learning set and attempted new element. These partition sets in turn partition the signal combinations into corresponding *pools*, and the history of the field is upated for each pool, as described in 4.2. Finally, starting from the current node as parent, each seed (one, more than one if a tie) defines a branch; and for each seed each partition set with its corresponding signal pool defines a node in the subsequent period. We then return to step 2.

The most computationally time-consuming step is step 3 since all possible seeds must be evaluated for each node and each seed is evaluated by forming an expectation over all possible signal combinations. As the field grows the field typically gets larger so the number of elements it contains grows. The number

[31] Each node has its associated value drawn from the basline log-normal distribution. If the actual value is drawn from the high distribution the baseline value is multiplied by v_H. If a new element is a duplicate string to an element already on the masterlist, but created in a new way, the values associated with the string (viability, whether it has a value drawn from the high distribution, and value) are assigned based on the element it duplicates.

of possible seeds grows roughly with the square of the number of elements in the field, and the number of possible full learning sets for each seed also grows at this rate. Thus the computation cost rises fairly rapidly. While the stationarity assumption that only elements and their parents can generate signals keeps the maximum number of signals from growing, in practice as the field grows there are fewer duplicate signals and fewer signals that are *clean* and thus do not need full evaluation. Further the number of nodes also grows with time periods, typically exponentially. In practice for the parameter values explored here the number of nodes in a period is equal to anywhere from 2 to 12 times the number of nodes in the preceding period.

In addition to the main run for each set of initial conditions and parameter values I also simulate a *clean* run for which there are no signals generated, as a comparison for how the field is predicted to develop when individuals do not gain any intuitive signals. The basic structure of the simulation is similar, but because there are no signals strategies are simpler and computational time is substantially lower.

The final step 5 consists of the analysis of the outcomes and paths generated by simulating the field development through T time periods. There are three main kinds of analyses. One is identifying the set of paths and field structures generated over all paths. Two is the distribution of output and new elements over all paths. Third is identifying and exploring patterns in output and new element creation, including the correlation between the creation of a new element or failure to create a new element and the level of output generated in the next period, as well as how the choices made by an individual in a period influences the development of the field going forward. In addition the results from the full model are compared with results from the *clean* runs.

6. Simulation Results

Table 1 lists parameter values used in the simulations. The base scenario parameter values have been chosen to focus on exploring the way intuitive signals guide create development and influence the development of a field. Thus, as presented in table 1, the intuitive signals for the base case are of high quality with low false positive and high true positive rates. For the random variable governing whether or not a given concatenation bundle is viable the false positive rate is set at .05 and the true positive rate at .9. For the random variable governing whether or not a concatenation has its value drawn from the high distribution the false positive rate is .1 and the true positive rate is .9. The *clean* simulations provide a comparison world in which there are no signals guiding creative development.

For convenience I present detailed results here for one simulation run for $N = 3$, then present and discuss the pattern of results across all simulations.[33] Figure 5 depicts a tree showing paths of development of the field for this run. The development through the first two generations is shown in some detail; generation 3 is shown more schematically. For the root node there is a single optimal seed (no ties). There are 7 paths emanating from the root node, which differ in terms of which new element the individual tries to make; these different choices in turn are driven by different possible signal pairings that may be generated given the optimal seed (recall that every signal pairing is simulated, so that every possible path of development of

[33] I note that because the initial state for $N = 3$ includes one additional element beyond the simple ring, there are none or very few ties in generation 1 or other generations, and as a result there are fewer total paths. There are also a relatively large number of distinct lattices, so that the ratio of number of distinct lattices to number of paths is relatively high compared to the other simulations.

the field is identified). For this run, based on the *masterlist*, 3 of the main elements turn out to be viable, paths 2, 5 and 7, and one of these spawns an additional new subbundle. Thus in period 2 there are 7 nodes, with 3 associated with a field that has grown beyond its initial state and 4 associated with a field consisting of just its initial elements. Even for these latter 4 nodes the histories associated with each node are different however, and thus the further development of the field may be different; most obviously which element was attempted in period 1 is different, and no individual along a path will attempt again an element that has already been attempted and shown not to be viable. The probabilities are quite high for the first two paths, the first associated with what I have called clean signals (signals that pertain to subbundles that cannot be created given the current state of the field), the second associated with most of the other signal draws. The remaining 5 paths have far lower probabilities; each of these paths is associated with specific positive signals that are drawn with low probability. In fact this is a typical pattern at most nodes for most simulations: 1 or 2 paths of relatively high probability, the remainder with quite low probability.

In generation 2 there are a total of 35 paths; 12 lead to the creation of one or more viable new elements. For three of these latter 12 paths the main new element that is created is brand new in the sense that it was not created along any path in generation 1; all of these are paths for which a new element had been created along the path in period 1, opening up new possibilities in period 2. The most paths emerge from node 2, for which a new element was created in period 1, reflecting the greater richness and options in the field as elements are added. However, the fewest paths are associated with node 5 for which a new element was also created. This can happen when the new element is part of the optimal seed at the current node, and for this seed, given the current configuration of the field, there are just two distinct learning sets and new elements to attempt that are optimal choices - high convergence across signals. Overall, through just two generations clear differences emerge in the how the field is developing depending on intuitive signals the individuals working in the field receive and the choices they make based on these signals.

In period 3 there are 199 paths in total. I show this more schematically in the Figure, with text boxes listing the number of paths from each period 2 node, the number of these for which new elements are created, and the number for which a brand new element is created. On average there are more paths emanating from nodes for which more viable elements have been created in the preceding generations. This is because when there are more elements in the field there are more possible combinations that can be formed, hence more signals for a given seed that are not clean but rather provide relevant information about different possible subbundles that can be made. In turn this triggers more variety of choices of full learning sets and new elements to try to make. Among these nodes, there are never more than 9 paths emanating from a given node. In fact it is generally the case that it is rare to have more than 10 paths emanating from a node. This is so even though the number of distinct signal combinations of not *clean* signals can be considerably higher; it turns out that in many cases subsets of signal pairs lead to the same choice of full learning set and new element to attempt and therefore are pooled for the *history* of the field as discussed in Section 4. Of the new elements created in period 3, there are 6 main new elements and 2 subbundles that are brand new in the sense that they were not created along any path through period two; of these 1 is created multiple (5) times and the others each just along a single path. Many of these are created on paths that link back to node 2 in period 2, which spawns the richest set of paths going forward. Thus there are paths that are unique in terms of novel elements being created, but there are a relatively small number of such paths, due in part

to the relatively low baseline viability probability of 0.3. The number of new elements created in period 3 varies across the parent nodes. In particular, for the last node (7), 4 of the 6 nodes have no new elements created along any paths. This contrasts with nodes 1 and 3 for which there are no nodes for which no new elements are created. Thus field productivity not only varies quite locally but also more globally depending on which elements have first been created earlier along a given set of paths. I do not exhibit generations 4 and 5 on the tree due to the large number of paths.

Through generation 5 the number of paths in total is more than 6000 for this simulation. The number of distinct field structures that is created is in general far smaller than the total number of paths of development of the field, because many paths lead to the same structure in terms of which elements have been created through period 5. For this particular simulation run there are 131 distinct field structures through period 5 compared with the more than 6000 paths. For other simulations there can be a considerably smaller number of distinct structures, as described later in this section.

Figure 6 depicts the development of the field for one illustrative path. In period 1 a new element is created and a subbundle is co-created. In period 2 the new element that is attempted is not viable. In each of periods 3, 4 and 5 a new element is added. In period 3 the new element is created using a pair of the initial field elements. In period 4 the new element combines an initial element with the element created in period 3, showing how individuals build on previous work in the field. Lastly, in period 5 the new elements created in periods 3 and 4 are combined, creating a new element that builds on the work of the preceding two individuals. The final structure has 5 elements and a height of 5 (height is measured as the longest chain, beginning from an attribute; note that the attributes are not shown in the figure except for attribute 1, shown by an open circle, utilized in the creation of the subbundle). It uses all 4 of the initial field elements and contains as the most complex (longest) element a string of length 8, created in period 5. In fact multiple paths generate this same final structure, differing in the signals received by individuals working in the field and the order in which elements are attempted to be made, as well as and including different elements that are attempted but turn out not to be viable. Note thus that two paths that generate the same field structure nonetheless represent different *states* in terms of the probabilitities of viability and of having a value drawn from the high distribution for strings that have not yet attempted, thus may lead to different choices in subsequent periods. Nonetheless the structure of actual elements that have been created provides a basic empirical description of the field, hence it is sensible to consolidate paths into structures for purposes of summarizing field development.

Figure 7 depicts examples of 15 other field structures out of the total of 131 distinct structures for this simulation. For each structure its cumlative probability summed over all paths that lead to it is listed. The first structure shown is associated with the path of *clean* signals every period. This structure is also created along some non-clean signal paths, and the probability shown is the cumulative probability along all paths that lead to the structure. Note that the bottom element is composed of initial element 4 combined with itself (in reverse order, so 1-2-3-2-1). That is possible because a new element added earlier has been made using initial element 4, thus two copies of element 4 are available. This is not unusual in the development of the field and satisfies the rules for constructing new elements laid out earlier. The second and third structures show the structures that have the greatest cumulative probabilities, 0.34 and 0.18. These are structures where just a few elements that are attempted turn out to be viable - thus many different paths,

with different histories of failed attempts to make other elements, pool into these structures. The remaining 12 structures provide examples drawn from the 67 structures out of the total 131 that are uncovered in that they are not contained in any other structure in the set. These structures thus typically contain more elements, though I have selected examples that cover the full range of number of elements, from 3 to 10. Note that some elements are made different ways for different structures, for example the combination of elements 3 and 4 can generate a main element (1-3-2-1 in one ordering) as well as a subbundle that links attribute 1 with initial element 1 (1 with 2-1 creating subbundle 1-2-1), but in other cases the 3-4 combination 1-3-2-1 is made itself as a subbundle and thus does not generate the second subbundle 1-2-1. Most of these structures are rare: they are generated along a small number of low probability paths typically tied to positive signals for particular subbundles that in turn lead to distinctive choices for which new element to attempt to make; however, many of these structures subsume structures with fewer elements that have substantially higher cumulative probability.

What is noteworthy about Figure 7 is the diversity of forms. This diversity is true not only for the overall structures, but also in terms of statistics, such as height. This finding of significant diversity highlights the wide range of possible paths of development of the field, an important point of this paper. In some cases one initial element is pivotal. For example, structure 5 (first structure, second row) shows a pattern of development in which element 4 plays a central role, being the parent for all but one of the subsequent elements created; similarly, in structure 7 element 1 plays a central role. Structures on the bottom row show more complex elements playing a central role in the further development of the field, stretching the structure vertically. Every path is generated through optimizing behavior on the part of the individuals working in the field, governed by the specific starting node and intuitive signals these individuals receive. Thus this diversity is not due to non-optimizing behavior, but rather to differing intuitive signals about valuable subbundles, which in turn guide the creative process. It is also noteworthy that because individuals build on previous work - elements created in preceding periods – these unique signals do not in any sense "cancel out" over the periods. Rather, the diversity is maintained and indeed grows over time at least through period 5 - the number of distinct field structures is more each period. Thus forecasting how a field like this will develop requires being sensitive to considering the wide range of possible paths of its development.

Figures 8, 9, 10 and 11 present statistics for the field structures generated in this simulation. Figure 8 shows the distribution of the number of new elements created. The top figure shows the distribution by cumulative probability over all paths that generate a given number of elements; the bottom figure shows the distribution by number of paths. The mean number of elements created (based on cumulative probabilities) is 4.3. The modal number of new elements is 3, associated with structures having a cumulative probability of .36. There is also substantial probabilities associated with 4, 5 and 6 elements being created, low but significant probabilities associated with 7 and 8 elements, and very small but nonzero probabilities for 9 and 10 elements being created - these are paths for which several subbundles have been co-created with main elements. Figure 9 shows the distribution of output, again first for cumulative probability (Figure 9a) and then for number of paths (Figure 9b). Mean output is 2.1 and the standard deviation is 1.4. Maximum output is 8.0; thus there is considerable variation, which is not surprising given the underlying value distribution (figure 3) and its long right tail associated with very high values. The distribution is actually multi-modal,

at least in terms of cumulative probability of paths, due to the fact that high value elements are created on a few high probability paths. The highest probability bin is a relatively low output of between approximately 0.48 and .98 with an associated probability of 0.4. Most of the paths in this bin are paths generated based on *clean* signals, which tend to have a higher overall probability since they all lead to the same choice of which element to try to make, and lower average output since there are no positive signals guiding choices. High probabilities are also associated with the bin covering the range 2.90 to 3.38 and the bin covering 3.86 to 4.35, substantially higher output levels. Consistent with the skewed element output distribution we expect and see a tail to the output distribution at the right, with non-negligible probability, 2.0%, of output above 6.2. This cumulative probability is associated with 140 distinct paths - indeed the skewness is clearer for the number of paths shown in Figure 9b. Overall, the output distribution fits our intuition about innovation and creative products, but here expressed at the field level: a relatively small percentage of paths of development have high output, while many paths for the field have relatively low ouput through 5 periods.

Figure 10 presents statistics on the distributions for height as well as element of greatest length, a natural measure of complexity of created elements. Height is measured as the longest chain from an initial element, each link in a chain linking a parent with a child created from that parent and a second parent. As shown in Figure 10a nearly all structures cluster in the 4 to 6 height region, with a handful having height 7 and 2 having heights below 4. Figure 10b shows the distribution of element of greatest complexity measured by longest length. This distribution is considerably more spread out. This is due to the fact that elements are created by overlapping parents of various lengths, and thus while the longest element created is in general the element at the bottom of a chain (and in most cases this is one of the longest chains in the structure), elements at the bottom of these longest chains vary in length because their parents vary in length. The modal longest length is 5, associated in particular with *clean* signal paths. There are also significant probabilities of structures associated with lengths 6, 9, 12 and even 15, with over 5% of structures have longest element of length 12. Considering that the field begins with initial elements of heights 2 and 3 this shows a very substantial growth in complexity over 5 periods. Figure 11 presents the distribution of number of initial elements used to create new elements in the field through the first 5 periods. Far and away the most common case is for 3 of the 4 initial elements to have been used, with the structures associated with this outcome having probability of 0.92 There are 15 structures out of the 131 that use fewer elements, including one case in which no new elements are created. There are 16 structures that use all 4 elements, however having quite low cumulative probability: these are associated with paths having unusual positive signals that lead to unusual choices for new elements to attempt to make. This is another indication of how important the intuitive signals are for generating diversity in the field, in the case in leading to every initial element to be used at least once in the creation of a new element, which does not occur for the *clean* paths.

Output Dynamics

The simulation results provide information about not only overall output but also output dynamics. One important calculation is the evaluation of expected output generated on path segments for which a new element (or more than one) was successfully created in the preceding period, compared with segments for which no new element was created. Essentially this is a time series property of output, exploring the correlation in output from one period to the next. There are in fact two offsetting effects that enter into this

relationship in the model. When a new element is created it opens up new frontiers for new seed choices, new signals that can be generated, and new elements that it is now possible to try to create which it was not possible to try to create previously. These factors tend to increase expected output, but by how much varies depending on the state of the field and the new element(s) made. Offsetting this, when a new element is created it is no longer available to attempt - it is removed from the pool of potential new elements to attempt to make. This effect tends to reduce expected output going forward. The reason this holds is that the pool of potential elements and in particular which elements in the pool are viable is fixed, prespecified by the *masterlist*. Thus considering two paths that start from the same node and supposing that a viable element is created along one of these paths but not the other, then the proportion of viable elements in the remaining pools (each reduced by a single member) is lower along the path for which the viable element was created. To clarify, suppose for the given node the pool of potential elements to attempt to make includes m elements that are viable and n that are not. Suppose further that on the first path from the node an element is attempted that is not viable and on the second path an element is attempted that is viable. If in the next period an element is attempted from the remaining elements in the pool (which will be the case for path 1, does not have to be the case for path 2 but may be), then the probability the element will be viable is, other things equal, $n/(n+m-1)$ on the first path and $(n-1)/(n+m-1)$ on the second path, thus greater on path 1.

The first row of Table 3 provides information about this tradeoff averaged over all simulation path segments. It shows that conditional on a new element being created in the preceding period expected output is 0.363, whereas conditional on no new element having been created expected output is .399, ten percent higher. Thus the second effect on average is actually more important, which may seem surprising. However, a further cut of the results reveals an additional factor. It stands to reason that the second factor, the fact that a new element has been created, will be more important when that element was attempted and created based on the individual receiving *clean* signals - for this element will be the most promising to try whenever the signals are clean, which happens relatively often. In contrast, when we restrict to paths for which the individual in the preceding segment received at least one active signal, the second factor will typically be less important, since he may well attempt to make an element that would not commonly be attempted, while the first factor will be quite important when he does in fact succeed in making a relatively more unusual (less commonly attempted) element. Based on this insight, the second part of the table shows results for expected output conditional on whether or not a new element was created in the preceding period, split out by whether the individual in the preceding period followed a *clean* path or did not.[34] We see that conditional on a clean path, expected output is 0.608 when no new element was created in the preceding period, and 0.33 when a new element was created. This large difference highlights the importance of the second factor for clean path segments. In contrast, the second line shows that conditional on a path that is not clean expected output is 0.357 when no new element was created in the preceding period and 0.372 when a new element was created. Thus for the non-clean path segments the effect reverses, the first factor dominates the second, though not by a large amount, and expected output is greater when a new element was created in the preceding period. These results carry interesting empirical predictions about the development of

[34] This path is the one adopted by individuals who receive signals that are both clean. It may also be followed for other signal draws.

fields: In general it is not correct that having a success immediately prior to the next attempt in the field is associated with higher output, but that is the case if the preceding attempt was based on active signals and thus entailed attempting to make a more unusual new element.

Path Dependence A second facet of output dynamics is how the field develops along distinct paths. As discussed above in the context of Figure 7 there is a large amount of diversity in terms of the number of distinct field structures. One important aspect of this diversity is the possibility of a high degree of path dependence. Once two paths diverge there is a high likelihood they will remain distinct and indeed grow apart. This is most evident when along one path, due to an unusual signal, an element is attempted and created and added to the field that is not typically attempted. For once this element is added, it provides a springboard for different seed choices and different intuitive signals being generated, which in turn leads to different elements that can be and at least in some cases are attempted and created, tending to drive the field even further away from other paths.

Table 4 provides evidence on path dependence for the simulation run being discussed. It shows the overlap and diversity between pairs of sets of elements created through period 5 for the 7 period 1 nodes, a total of 21 comparisons. Node 1 is associated with the *clean* signal path for period 1, as well as a few additional signal paths. It has 29 new elements created over all paths emanating from it. Recall from the discussion around Figure 5a that 3 of the 7 nodes – not node 1 – had a new element successfully created, and one, node 2, had more paths and more distinct elements in the subsequent period 3. The numbers in the table corroborate that node 2 is associated with a large, distinct set of created elements all the way out through period 5. In particular there are a total of 50 new elements created along all paths emanating from this node. As an example of overlap and diversity, nodes 1 and 2 share 24 elements in common, while node 1 has 5 new elements associated with its paths that are not created by any path emanating from node 2, and node 2 has 26 elements associated with its paths that are not created along any path emanating from node 1. Interestingly, no pair among the 21 different pairings of period 1 nodes has two paths that are identical in terms of the final set of elements created. Further, in all but two cases each node set contains elements not contained in the other node set, so that the development of the field in each case produces elements not created along the other path. Thus there is a high degree of path dependence beginning from the first period.

Learning Sets

The simulation results also provide information about the learning sets individuals choose. In fact learning sets are an important area of empirical prediction of the model and have policy implications for education and what kinds of learning paths are supported. One important aspect of learning sets is the issue of what a central administrator knows and how such an administrator might structure learning, as opposed to the learning sets individuals will freely choose. For considering policy around this issue it makes sense to assume that central administrators have all public knowledge about the state of the field at a given node, including the full history of the field, but do not have any information about the intuitive signals individuals working in the field receive. Based on this assumption there is a crucial distinction between seed and full learning sets. For a given node the optimal seed learning set is based only on public information. Thus a central administrator should be able to determine the optimal seed learning set. However, conditional on

the seed set, the optimal full learning set depends on the intuitive signals an individual receives, which are not known to the administrator. Therefore the administrator will make a less well-informed decision about the full learning set.

It follows that a useful issue the simulations can shed light on is how much variability there is in full learning sets at different nodes. This is because a central administrator, acting only from public information, will specify a single full learning set at a given node, typically the one compatible with *clean* signals or no signals. When there is more than one optimal full learning set the administrator will therefore miss opportunities, for some signal pairs, to guide the optimal choice. As it turn out there is significant variation in full learning sets. Intuitively, we also expect this variation to grow with the size of the field. For the simulation run focused on here are descriptive statistics about learning sets. Note first that at any node for which the field includes only the initial elements there is no choice regarding the full learning set since all 4 elements will be included. Thus for period 1 there is only one possible full learning set. For period 2 there are 3 nodes (out of 7) for which one or more elements were created in period 1 and therefore there is a choice regarding the full learning set. For 2 of these 3 nodes there is more than one optimal full learning set, thus the optimal set does depend on intuitive signals; though in each case only one out of several paths (6 for one node, 7 for the other) is different from the other paths and has relatively small probability. In period 3 there are a total of 23 nodes at which the field includes additional elements. For 13 of these there is more than one optimal full learning set depending on the signals. For 4 of these nodes there are 3 distinct full learning sets, for the other 9 nodes there are 2. For period 4, out of 161 nodes for which the field includes at least one additional element beyond the initial elements, there are 79 nodes for which there is a single full learning set common to all paths, representing 49% of the nodes, 32 for which there are 2 distinct full learning sets, 48 for which there are 3, and 2 for which there are 4. Lastly, for period 5, out of 1047 nodes for which the field includes at least one additional element, there are 471 nodes for which there is a single full learning set, representing 45% of the total number of nodes, 298 nodes for which there are 2 distinct full learning sets, 247 for which there are 3, 29 for which there are 4, and 2 for which there are 5. Thus as the field grows in complexity there is in general more diversity in full learning sets and an administrator specifying a single learning set will err for more than one-half of the paths and typically somewhat less than one-half of the path probabilities. Overall, variability in full learning sets is important and as expected becomes more important as the field grows and we can expect that a central administrator would not be able to assign optimal learning sets in many cases.

Clean Runs

I have alluded above to clean paths associated with *clean* signals that provide information about subbundles that are not able to be created given the present configuration of the field. I now provide a more precise way to calibrate and appreciate the role of signals in the development of the field by presenting a set of comparison results from simulations starting from the same initial conditions for the case in which individuals gain no intuitive signals. In this case there are typically many fewer paths. Indeed for many nodes there is just a single path forward based on the evaluation of the optimal strategy given the current state of the field. In some cases more than one seed has the same expected value and in those cases I assume each seed is chosen with equal probability, essentially a mixed strategy, thus generating more than one path.

For this particular simulation there are 21 clean paths, but they all lead to the same final structure through period five. The field structure generated through this path matches the first field structure in Figure 7 for the main simulation. Its longest element is length 5, its height is 3, and it uses 3 of the 4 initial elements. The output from this path is 1.2, well below the 2.1 mean value for the signal case, showing the value of information associated with the signals. The number of elements created is 4, slightly below the mean of 4.3 for the signal case. It is intuitive that the value of the signals is expressed more fully in output, since that is what individuals aim to maximize, rather than number of elements, though that is also increased. In particular, in some cases a signal of a potentially high value may lead an individual to attempt to make an element for which the probability of viability of the main element or associated subbundles is actually somewhat lower, but more than offset by the higher expected output if the element (and subbundle(s)) turn out to be viable. Overall the clean run shows how much richer the potential development of the field is with the intuitive signals, both in terms of the range of possible paths as well as the number of elements, output, and complexity of created elements and structure. Indeed an important finding is that it is the signals that generate the very great range of possible paths of development and structures for the field.

Full Set of Simulations

Table 5 and Figures 12 and 13 summarize results for the full set of 9 simulations for the base case. 3 simulations were performed for each of 3 initial conditions: $N_0 = 3$, $N_0 = 4$ and $N_0 = 8$. The ring initial condition was used for all simulations for $N_0 = 4$ and $N_0 = 8$ while the initial condition shown in Figure 6 was used for $N_0 = 3$. For each N_0 value the difference across runs is the $master file$ generated for that run that details which new elements are viable and which are not, and the values associated with viable new elements. Note that because the $N_0 = 3$ initial condition includes one additional element (recall this is done so that there are valuable signal for the first period), results for this case are slightly less comparable with the other two N_0 values, whereas the $N_0 = 4$ and $N_0 = 8$ cases represent the natural scaling up of the initial ring (from 4 to 8 elements) as N_0 increases, and thus are quite directly comparable. The table shows that the number of paths is substantially higher for the $N_0 = 8$ simulations. Extrapolating specifically from the $N_0 = 4$ results we see that the number of possible paths of development of the field increases with N_0, which is not surprising since there are more basic elements hence more possible distinct combinations. However, the number of distinct field structures through five periods is not significantly higher for the $N_0 = 8$ case than for $N_0 = 4$ or for that matter for $N_0 = 3$. Thus there are more duplicate paths as N_0 increases.

It is noteworthy that the number of distinct structures varies very widely, from a low of only 5 to a high of 143. Indeed the distribution has two central regions, one containing runs with quite low numbers of distinct structures, 20 or below, and a second region containing runs for which there is a far larger number of distinct structures, above 90; only 1 out of 9 simulations does not fall in these two regions. Thus the degree of variability in set of possible paths of development of a field is itself highly variable.

A consistent regularity across all simulations is that expected output is higher for the signal model than for the clean model, reflecting the value of information in the signals. The differential between the two varies substantially, reflecting the high variability in output, especially due to the right tail of the output distribution. While the expected number of elements would not necessarily be greater for the signal model, since as noted above it is expected output that is being maximized not expected number of new elements

formed, nonetheless it is also consistently higher.

Conditional expected output is higher following a successful new element being made in 4 out of the 9 simulations, and higher following no new element being made for the remaining 5 simulations. Differentiating between cases when a clean path has been followed, versus a not clean signal path, conditional expected output is higher following a new element being made in 3 out of 9 simulations for clean paths, and in 5 out of 9 simulations for not clean paths. Further, the difference in expected output between when a new element is formed and when it is not is greater for the not clean paths in 7 out of the 9 simulations. Thus the intuition discussed above is largely born out, that making an existing element is a more important negative factor, relative to the benefit of opening up new possibilities with the new element, for clean paths than not clean paths.

Figures 12 and 13 show the distribution of output and number of new elements for each of the 9 simulations, based on the cumulative probability for each outcome or bin. Every simulation shows substantial variation in both the number of new elements created and output. This again highlights the very substantial variability in how the field may develop for a given set of parameters and intitial conditions. Although path dependence results similar to those shown in Table 4 for the first simulation are not shown due to space, there is also very substantial path dependence exhibited in every simulation.

Alternative Scenario: High Viability Probability

I ran an alternative scenario for $N_0 = 3$, altering the probability a new element is viable from 0.3 to 0.5. We expect more new elements to be produced, and the simulations of this scenario show how this translates into differences in the way the field may develop. Figure 14 and Tables 6 and 7 present results for 3 simulation runs for this scenario.

The most striking difference between these results and those for the baseline probability are that there are many more distinct structures. This is related to the fact that there is a greater diversity of created elements. There are two reasons for these results. One is that more elements that are attempted are made, which leads to a larger set of made new elements each period.[35] The other, more subtle, is the fact that, with a higher baseline probability of viability, individuals are free to respond more to the probability a given element will have a value drawn from the high distribution in making their choice of which element to attempt to make. Being more sensitive to these signals, whereas previously the viability probability and signals of viability were more dominant, leads to a greater variety of choices of which elements to attempt along different paths. Table 7 provides additional evidence on the diversity of sets of created elements across paths for the first simulation. This table, similar to Table 4, shows the overlap and diversity between pairs of sets of elements created starting from each period 1 node. Just as for the baseline simulation results shown in Table 4, there are 7 period 1 nodes. Compared with the figures in Table 4 the number of created elements for each node is substantially higher. The diversity across nodes remains very high.

Related to the results about diversity of elements, a second finding is that there is now a somewhat larger, though still small, chance of a very high output draw. An illustration of this is the output distribution for the first simulation. Expected output is not unusually high, but there is a very small probability, associated with just 2 paths, of a very high output above 100.0 - shown in the figure as a very small bar

[35] Note that this does not extend to the number of paths being larger. Paths are generated even when an element is not viable, since the history update is distinct.

on the far right. There are several other paths also with high output, also shown as small bars. A similar pattern holds for simulation 3.

Finally, a third interesting result is that expected output and number of elements is relatively higher for the clean paths, compared with the baseline case. This is because the value of information associated with the viability signals is greater when the baseline probability is lower. Reflecting the lower value of information of signals, in the third simulation the expected number of elements and output are actually higher for the clean run than the full model, a result which does not obtain in any of the 9 baseline simulations.[36]

Overall fields with higher baseline probability of viability can be expected to exhibit greater variety of created elements and a small but non-neglible probability of a very high output path.

7. Royalty Simulation

Important policy questions center on the use of intellectual property and financial incentives to guide individual creative development and shape the development of a field. As an illustration of how these issues may be investigated in the context of the model in this paper I have explored how the payment of a royalty for use of a created element influences the development of the field. The specific policy experiment is that an individual is paid a fixed dollar royalty if an element he creates is used as a building block to create a new element in the subsequent period. The royalty rate is set at 0.3 or 30% of the mean value of a created element. The royalty is only paid if the element that is attempted is in fact viable and thus made. Since the probability of viability is set at 0.3 in the simulations, the expected payment is .09.

Simulation of the model in this case is significantly more complicated than for the main model. Individuals must now forecast, for each element they may choose to attempt to make, the likelihood that if created the element will be used in the subsequent period as a building block. This requires forecasting what seed the individual next period will choose, the distribution of signals he will receive, and the elements he will choose to attempt to make for each signal draw. Since this individual is in turn forecasting what the individual in the following period will do, the model must be projected out, here to the last period 5, taking into account all potential paths of development, and then rolled back. To make this exercise clean I assume no royalty is paid for elements created in period 5 - this can be taken as the final period of development of the field. A further complication is that an individual does not know, at the time he decides whether or not to attempt to make a given element, whether the element will be viable. To calculate the expected royalty he must forecast the field assuming the element is viable. Thus many counter-factual paths are generated. Ultimately, when the full rollback is done, the optimal period one seed, learning set, and creative project element are determined, feeding forward into period two and so on.[37]

Results of this simulation demonstrate that a royalty payment can influence the creative development of a field, in expected ways. The simulation was performed for the third simulation for the $N_0 = 3$ case, for which the number of paths is intermediate for this N_0. The *masterlist* for the base scenario was used, so all elements that overlap between the two cases are the same in terms of whether they are viable and their

[36] This is possible when there are some elements attempted along a clean run that are not attempted for the signals; whether or not these elements are viable and their output is drawn at random and thus a high proportion may by chance turn out to be viable and have high value.

[37] This simulation was very computationally burdensome, requiring several months of dedicated computer time on the Yale HPC.

value if so. New elements are added as needed to the *masterlist* and indeed there are many new elements that are added since the individual must consider the possibility an element will be viable (even if it turns out not to be) and then used to build further hypothetical elements.

Figure 15 presents results comparing the royalty simulation to the base case. I focus on periods 1 and 2, since these are well before the termination of the model and thus the cleanest for exploring the impact of the royalty. In period 1 the set of paths (attempted elements) is in fact identical, 7 elements or paths. Since the *masterlists* are the same, the same three elments are made, one having a viable subbundle. However, in period 2 there is significant divergence betweeen the two cases. The 4 nodes for which no element was created in period 1 are identical. However, the 3 nodes for which an element was created are all different. In two of the three cases there are two additional paths for the royalty case (10 versus 8), while in the third case the number of paths is the same (two) but one of the paths is different; the 5 new paths are shown in red in the figure and for clarity probabilities are shown only for these paths. The logic behind the shift to attempt to make these elements instead of the previous elements for the base case is that for some signals the individual shifts from another element, that has higher expected value but is less likely to be used as a building block if it turns out to be viable, to a second element that has lower expected value but is more likely to be useful as a building block if viable. The average number of new elements created along paths emanating from these 5 nodes is also relatively high, on average approximately 0.5 elements higher than the average for the base case. The probabilities associated with these 5 new paths however are modest. For the first node, which itself is a highly probability period 1 path, each of the two new paths has approximate probability of .01. For the second node, also quite high probability from period 1, one new path has probability approximately .044 and the other approximately 0.09. For the last node, the new path has high probability of 0.70, but this path itself is a relatively low probability in period 1. Thus the overall impact of the royalty is modest, but in the expected direction of spurring the development of elements that in turn are useful as building blocks for further development. Since the royalty encourages individuals to attempt to make elements that are useful as building blocks for future work in the field we expect the royalty to increase output. It does so, but the effect is modest, in line with the modest overall impact in terms of new paths with associated probabilities. Output is increased by approximately 1% over the base case, from 1.77 to 1.79.

For the clean run the royalty results are identical to the base run. Despite this, there are many nodes for which new elements are made that are in fact built on in the following period. Thus the royalty is paid out, but does not alter how the field develops. In fact, it is easy to understand why a royalty set at 0.3 has no effect in this case. This value is exactly the same as the expected value of a new element when the probability of viability for the new element is set at the baseline value of P_X and the probability its value is drawn from the high distribution is at the baseline of P_H. Since there are no signals every potential new element in fact has its probabilities at these baseline values. For the clean case, for which these probabilities are all the same, among a set of potential new elements that might be attempted at a node the highest ranked are those with the most potential new subbundles that may be co-created. Below this are elements with one less subbundle; and these elements have expected value exactly 0.09 below the top group, hence are not preferred even if they were to earn a royalty (would be equivalent with a probability 1.0 of earning a royalty which in fact does not occur in the clean run as at every node there is randomization over two or more new elements). Note that this argument does not apply to the signal model since probabilities are

updated based on signals and therefore the probability of viability can be greater or smaller, if that element can be represented as a signal subbundle, which some new main elements can be.

Overall, a royalty value of 0.3 though a reasonable percentage of expected output in economic terms has only a modest effect on the development of the field.

7. Conclusion

In this paper I have presented a model of the creative development of a field. The model is explicitly based in individual learning and creativity. The model depicts the structure of the field as it grows resembling a lattice.

The model is used to explore a range of issues, including the diversity of possible paths of development of the field, how individuals build on the work of their predecessors, and the role of intuitive signals in the development of the field. The model also has implications about output, including average output, the variance of output, and serial correlation in output. Main results include the very substantial diversity of generated structures for the field's development, the importance of intuitive signals both in guiding creative development and generating the diversity of outcomes, and a high degree of path dependence.

The approach taken in this paper opens the way to modeling knowledge creation in a more structured way than has been done previously in formal decision-based models of creativity and innovation. This forges an important bridge between knowledge representation and economics, centering on creativity and innovation and the production of new knowledge. This is a link truly in its infancy with a great deal more to be done if we are to represent the incredible rich knowledge sets individuals form, the distinctiveness of knowledge sets across individuals, and the importance of these sets for creativity and innovation in all its forms, including strategy, technology, policy, and many other realms. Further and more broadly there is much to be done in representing the rich structure of human knowledge in a way that connects with economic and decision theory models. The structure of knowledge is integral to the human condition, including the modern economy and human culture. Only by representing this structure can we hope to be able to understand how we as individuals build on the work of their predecessors, in specific conceptual ways and following defined rational decision processes, and how humans in general respond to circumstances, including shocks to their environment, in innovative ways.

Appendix: Additional Formulas

Signal Probabilities: Case in which both signals are associated with the same string. In this case the probabilities for the pair of X signals are worked out jointly, and likewise the probabilities for the pair of H signals are worked out jointly. Thus the probability that both signals have $X = 1$ is:

$$s_1^X * s_1^X * P_X(s,t) + s_0^X * s_0^X * (1.0 - P_X(s,t)$$

Note that the prior probability is the same since a given string s has a given prior probability $P_X(s,t)$ associated with it, regardless of how it is made. The probability one signal has $X = 1$ and the other has $X = 0$ is:

$$s_1^X * (1.0 - s_1^X) * P_X(s,t) + s_0^X * (1.0 - s_0^X) * (1.0 - P_X(s,t))$$

The probability both signals have $X = 0$ is:

$$(1.0 - s_1^X) * (1.0 - s_1^X) * P_X(s,t) + (1.0 - s_0^X) * (1.0 - s_0^X) * (1.0 - P_X(s,t))$$

Comparable formulas hold for the H signals. Note that the X and H signals continue to be independent.

Probability Updates Based on Signals: Case in which both signals are associated with the same string. The X and H signals remain independent and thus update formulas are worked out seperately for these two random variables. I provide the formulas for the X update here; the H formulas are similar.

There are 3 possible cases for the X signals: both are 1; one is 1 and the other is 0; or both are 0. When both signals are 1 the updated probability that the string is viable is:

$$P_X(s,t \mid X_1 = 1, X_2 = 1) = \frac{s_1^X * s_1^X * P_X(s,t)}{s_1^X * s_1^X * P_X(s,t) + s_0^X * s_0^X * (1.0 - P_X(s,t)}$$

where X_1 refers to the X signal associated with the first version of the string and X_2 refers to the signal associated with the second. When one signal is 1 and the other is 0 the upated probability the string is viable is (without loss of generality assume $X_1 = 1$ and $X_2 = 0$):

$$P_X(s,t \mid X_1 = 1, X_2 = 0) = \frac{s_1^X * (1.0 - s_1^X) * P_X(s,t)}{s_1^X * (1.0 - s_1^X) * P_X(s,t) + (1.0 - s_0^X) * s_0^X * (1.0 - P_X(s,t))}$$

Finally the updated probability for the case in which both signals are 0 is:

$$P_X(s,t \mid X_1 = 0, X_2 = 0) = \frac{(1.0 - s_1^X) * (1.0 - s_1^X) * P_X(s,t)}{(1.0 - s_1^X) * (1.0 - s_1^X) * P_X(s,t) + (1.0 - s_0^X) * (1.0 - s_0^X) * (1.0 - P_X(s,t))}$$

Update Formulas When There are Two Signals Associated With The Same String. These formulas apply when two signals in the set of signals from which signal combinations are selected refer to the same string. In this case there are six distinct subsets within the *pool*: (i) the set of combinations in the pool for which one of the signals refers to string s and the other does not, and the signal associated with s is 1; the set of combinations for which one of the signals refers to s and the other does not and the signal associated with s is 0; (iii) the set of combinations for which both signals refer to s and both signals are 1; (iv) the set of

combinations for which both signals refer to s and both signals are 0; (v) the set of combinations for which both signals refer to s and one signal is 1 and the other 0; and (vi) the set of combinations for which no signals refer to s. Note that any of these six subsets can be null, since the *pool* is a selected group of signal combinations. For example cases in which both signals refer to s and both signals are 1 might lead to the individual in $t-1$ choosing to try to make a different element; in that case his observed choices reveal that he did not receive two 1 signals for s. Let q_1, q_2, q_3, q_4 and q_5 denote the weighted probabilities associated with the first five subsets. Then q_1 is given by:

$$q_1 = \frac{\sum_{i \in pool} Ind(one\ sig\ in\ combo\ i\ has X = 1\ for\ string\ s) P(sig\ combo\ i)}{P_{pool}}$$

A comparable expression holds for q_2. For q_3 the formula is:

$$q_3 = \frac{\sum_{i \in pool} Ind(both\ sigs\ in\ combo\ i\ have\ X = 1\ and\ refer\ to\ string\ s) P(sig\ combo\ i)}{P_{pool}}$$

Comparable expressions hold for q_4 and q_5. The updated probability that s is viable is then given by:

$$P_X(s,t\ |X=1) * q_1 + P_X(s,t\ |X=0) * q_2 + P_X(s,t \mid X_1 = 1, X_2 = 1) * q_3 + P_X(s,t \mid X_1 = 0, X_2 = 0) * q_4 +$$

$$[P_X(s,t \mid X_1 = 1, X_2 = 0) + P_X(s,t \mid X_1 = 0, X_2 = 1)] * q_5 + P_X(s,t) * (1.0 - q_1 - q_2 - q_3 - q_4 - q_5)$$

Comparable formulas hold for the updated probability that the value associated with s is drawn from the high distribution.

References

Abramovitz, M. (1956): "Resource and output trends in the United States since 1870," National Bureau of Economic Research paper.

Acemoglu, D., S. Johnson and J.A. Robinson (2005): "Institutions as a fundamental cause of long-run growth," in *Handbook of Economic Growth*, Vol. 1, Part A, ed. P. Aghion and S.N. Durlauf. Philadelphia, PA: Elsevier, pp. 385-472.

Adam, C., C. Ofria, T.C. Collier (2000): "Evolution of biological complexity," *Proceedings of the National Academy of Science*, 97, 9, pp. 4463-8.

Aghion, P. and P. Howitt (1992): "A model of growth through creative destruction," *Econometrica*, 60, 2, pp. 323-51.

Aghion, P., C. Harris, P. Howitt, J. Vickers (2001): "Competition, imitation and growth with step-by-step innovation," *Review of Economic Studies*, 68, 3, pp. 467-92.

Albarrán, P., J.A. Crespo, I. Ortuño, J. Ruiz-Castillo (2011): "The skewness of science in 219 sub-fields and a number of aggregates," Working Paper, Departmento de Economía, Economic Series 11-09, Universidad Carlos III de Madrid.

Bramoullé, Y. and G. Saint-Paul (2010): "Research cycles," *Journal of Economic Theory*, 145, pp. 1890-1920.

Campbell, A. (2013): "Word of mouth and percolation in social networks," *American Economic Review*, forthcoming.

Campbell, A. and J.S. Feinstein (2013): "What is taught in core micro courses at top US doctoral programs: overlap and diversity," working paper.

Campbell, D. (1960): "Blind variation and selective retention in creative thought as in other knowledge processes," *Psychological Review*, 67, pp. 380-400.

Cohen, I.B. (1990): *Benjamin Franklin's Science*. Cambridge: Harvard University Press.

Cohen, L.M. (2009): "Linear and network trajectories in creative lives: A case study of Walter and Roberto Burle Marx," *Psychology of Aesthetics, Creativity, and the Arts*, 3, 4, pp. 238-48.

DiPaola, S. and L. Gabora (2009): "Incorporating characteristics of human creativity into an evolutionary art algorithm," *Genetic Programming and Evolvable Machines*, 10, pp. 97-110.

Dorf, R.C. and J.A. Svoboda (2006): *Introduction to Electric Circuits.* John Wiley & Sons, 7th Edition.

Evenson, R.E. and Y. Kislev (1976): "A stochastic model of applied research," *Journal of Political Economy*, 84, 2, pp. 265-82.

Fauconnier, G. and M. Turner (2002): *The way we think: conceptual blending and the mind's hidden complexities*. New York: Basic Books.

Fauconnier, G. and M. Turner (1998): "Conceptual integration networks," *Cognitive Science*, 22, 2, pp. 133-87.

Feinberg, Y. (2005): "Subjective reasoning - dynamic games," *Games and Economic Behavior*, 52, 1, pp. 54-93.

Feinberg, Y. (2008): "Meaningful talk," in *New perspectives on games and interaction*, ed. by K.R. Apt and R. van Rooij, Amsterdam: Amsterdam University Press, pp. 105-19.

Feinstein, J.S. (2006): *The nature of creative development*. Stanford, CA: Stanford University Press.

Feinstein, J.S. (2011): "Optimal learning patterns for creativity generation in a field," *American Economic Review Papers and Proceedings*, 101, 3, pp. 227-32.

Fleming, L. and O. Sorenson (2004): "Science as a map in technological search," *Strategic Management Journal*, 25, pp. 909-28.

Gabora, L. (2005): "Creative thought as a non-Darwinian evolutionary process," *Journal of Creative Behavior*, 39, pp. 262-83.

Gabora, L., B. O'Connor, A. Ranjan (2012): "The recognizability of individual creative styles within and across domains," *Psychology of Aesthetics, Creativity and the Arts*, 6, pp. 351-60.

Gambardella, A., D. Harhoff and B. Verspagen (2008): "The value of European patents," *European Management Review*, 5, pp. 69-84. Helbig, H. (2006): *Knowledge representation and the semantics of natural language*. Berlin: Springer.

Ganter, B. and R. Wille (1997): *Formal concept analysis: mathematical foundations*. New York: Springer-Verlag.

Ganter, B., G. Stummer and R. Wille (2005): *Formal concept analysis: foundations and applications.* Berlin: Springer.

Garicano, L. (2000): "Hierarchies and the organization of knowledge in production," *Journal of Political Economy*, 108, 5, pp. 874-904.

Gentner, D. (1983): "Structure-mapping: a theoretical framework for analogy," *Cognitive Science*, 7, pp. 155-70.

Ghiglino, C. (2012): "Random walk to innovation: why productivity follows a power law," *Journal of Economic Theory*, 147, pp. 713-37.

Goyal, S. (2007): *Connections: An introduction to the economics of networks*. Princeton, NJ: Princeton University Press.

Griliches, Z. (1992): "The search for R&D spillovers," *The Scandinavian Journal of Economics*, 94, Sup., pp. S29-47.

Gruber, H. (1974): *Darwin on man: A psychological study of scientific creativity*. New York: E.P. Dutton.

Harhoff, D., F. M. Scherer and K. Vopel (2005): "Exploring the tail of patented invention value distributions," in D. Harhoff, F.M. Scherer, and K. Vopel, eds.: *Patents: Economics, policy, and measurement*. Cheltenham, U.K.: Elgar, pp. 251-81.

Hayek, F.A. (1960): *The constitution of liberty*. Chicago: University of Chicago Press.

Helbig,H., (2006): *Knowledge representation and the semantics of natural language*. Berlin, Heidelberg : Springer-Verlag.

Heylighen, F. (1996): "The growth of structural and functional complexity during evolution," working paper, Center "Leo Apostel", Free University of Brussels.

Jaffe, A.B. and M. Trajtenberg (2002): *Patents, citations, and innovations: A window on the knowledge economy*. Cambridge MA: MIT Press.

Jaffe, A.B., M. Trajtenberg and M.S. Fogarty (2000): "Knowledge spillovers and patent citations: Evidence from a survey of inventors," *American Economic Review*, 90, 2, pp. 215-8.

Jaffe, A.B., M. Trajtenberg and R. Henderson (1993): "Geographic localization of knowledge spillovers as evidenced by patent citations," *Quarterly Journal of Economics*, 108, 3, pp. 577-98.

Jackson, M. O. (2008): *Social and economic networks*. Princeton, NJ: Princeton University Press.

Jones, B.F. (2009): "The burden of knowledge and the "Death of the Renaissance man": Is innovation getting harder?," *Review of Economic Studies*, 76, 1, pp. 283-317.

Kaplan, S. and K. Vakili (2013): "Novelty vs. usefulness in innovative breakthroughs: A test using topic modeling of nanotechnology patents," working paper.

Kauffman, S.A. (1993): *The origins of order: Self-organization and selection in evolution*. Oxford: Oxford University Press.

Keynes, J.M. (1977): *The Collected Writings of John Maynard Keynes. Volume XVII; Activities, 1920-1922; Treaty Revision and Reconstruction*. Edited by Elizabeth Johnson. London: Macmillan and Cambridge University Press.

Keynes, J.M. (1981): *The Collected Writings of John Maynard Keynes. Volume XIX; Activities, 1922-1929; The Return to Gold and Industrial Policy*. Two Volumes. Edited by Donald Moggridge. London: Macmillan and Cambridge University Press.

Koestler, A. (1964): *The act of creation*. New York: Macmillan.

Kortum, S.S. (1997): "Research, patenting, and technological change," *Econometrica*, 65, 6, pp. 1389-1419.

Lancaster, K.J. (1966): "A new approach to consumer theory," *Journal of Political Economy*, 74, 2, pp. 132-57.

Li, M., and P. Vitanyi (1997): *An Introduction to Kolmogorov complexity and its applications*. New York: Springer.

Marshall, A. (1919): *Industry and trade*. London: Macmillan.

Mas-Colell, A. M. Whinston and J. Green (1995): *Microeconomic Theory*. New York: Oxford University Press.

Matisse, H. (1990): *Matisse on Art*. Edited by Jack D. Flam. New York: Phaidon Press.

Mead, C. and L. Conway (1980): *Introduction to VLSI Systems*. Addison-Wesley.

Mednick, S.A. (1962): "The associative basis of the creative process," *Psychological Review*, 69, pp. 220-32.

Mill, J.S. (1978; originally published 1859): *On liberty*. Indianapolis: Hackett Publishing.

Miller, G.A., et. al. (1990): "WordNet: An online lexical database," *International Journal of Lexicography*, 3, 4, pp. 235-44.

Mokyr, J. (2002): *The gifts of Athena: historical origins of the knowledge economy*. Princeton: Princeton University Press.

Mosley, L. (1992): *Disney's World*. Chelsea, MI: Scarborough House.

Murray, F., P. Aghion, M. Dewatripont, J. Kolev, S. Stern (2009): "Of mice and academics: examining the effect of openness on innovation," NBER Working Paper No. 14819.

Pakes, A. (1986): "Patents as options: some estimates of the value of holding European patent stocks," *Econometrica*, 54, 4, pp. 755-84.

Poincaré, H. (1908, 1952): "Mathematical creation," in *The creative process: a symposium*, ed. B. Ghiselin, Berkeley: University of California Press, pp. 33-42.

Princeton University (2013): *WordNet*. Address: wordnet.princeton.edu.

Romer, P.M. (1990): "Endogenous technological change," *Journal of Political Economy*, 98, 5, Part 2, pp. S71-S102.

Scherer, F.A. and D. Harhoff (2000): "Technology policy for a world of skew-distributed outcomes," *Research Policy*, 29, pp. 559-66.

Shannon, C., and W. Weaver (1971): *The Mathematical theory of communication*. Urbana, IL: University of Illinois Press.

Simonton, D. K. (1999): *Origins of genius: Darwinian perspectives on creativity*. New York: Oxford University Press.

Simonton, D.K. (2003): "Scientific creativity as constrained stochastic behavior: the integration of product, process, and person perspectives," *Psychological Bulletin*, 129, pp. 475-94.

Silverberg, G. and B. Verspagen (2007): "The size distribution of innovations revisited: an application of extreme value statistics to citation and value measures of patent significance," *Journal of Econometrics*, 139, pp. 318-39.

Skidelsky, R. (1983): *John Maynard Keynes. Volume Two; The Economist as Saviour*. London: MacMillan.

Söderlind, P. (2012): *Lecture notes - econometrics: some statistics*. University of St. Gallen.

Solow, R.M. (1957): "Technical change and the aggregate production function," *Review of Economics and Statistics*, 39, 3, pp. 312-20.

Sowa, J.F. (1984): *Conceptual structures: information processing in mind and machine*. Reading, MA: Addison-Wesley.

Sowa, J.F. (2000): *Knowledge representation: logical, philosophical, and computational foundations*. Pacific Grove, CA: Brooks/Cole.

Spurling, H. (1998): *The Unknown Matisse*. Berkeley: University of California Press.

Ward, T.B., S.M. Smith and J. Vaid, editors (1997): *Creative thought*. Washington, DC: American Psychological Association.

Weitzman, M.L. (1998): "Recombinant growth," *Quarterly Journal of Economics*, 113, 2, pp. 331-60.

Wille, R. (1992): "Concept lattices and conceptual knowledge systems," *Computers and Mathematics with Applications*, 23, pp. 493-515.

Table 1: Model Parameters

Parameter	Definition	Value (base case)
A. Knowledge Structure and Value Distributions		
N	Number of attributes.	3, 4, 8
N_0	Number of initial field elements.	4, 4, 8
p_X	Probability that string is viable.	0.3
p_H	Probability that string value is drawn from tail of distribution.	0.1
v_H	Tail distribution expected value multiplier.	10.0
μ	Mean of log-normal for value distribution.	1.0
σ	Standard deviation of log-normal distribution.	0.25
b	Truncation point for lognormal (calculated).	
α	Pareto distribution parameter (calculated).	
B. Learning Sets and Signals		
M_{Seed}	Number of elements chosen for seed learning set.	2
M_{Full}	Number of elements in full learning set.	4
m	Number of signals individual receives.	2
s^x_0	X signal false positive rate.	0.05
s^x_1	X signal true positive rate.	0.9
s^H_0	H signal false positive rate.	0.1
s^H_1	H signal true positive rate.	0.9

Table 2: Simulation Protocol

Step 1: Generate initial field state. N_0 elements each of length 2 generated as a ring or (N_0 = 3) a ring with one additional element.

Recursive Block

Step 2: Identify all nodes for current period. Build list of elements in field for each node. Extend *masterlist* - loop through all nodes, identify all new elements can make given list, add the element to the *masterlist* and generate values for the element.

Step 3: For each current period node compute optimal seed choice and identify optimal strategy for that seed choice. Analytic calculation.

Step 4: For each current period node, based on optimal seed and strategy identified in step 3, determine set of new elements to try to make and use *masterlist* to determine outcomes, including all subbundles that are co-created with main element (if main element is viable); divide outcomes for the node into *pools* based on full learning set and new elements chosen to try to make, and update field history based on *pools*.

RETURN to Step 2. Exit after period T.

Step 5: Analysis of results.

Figure 1: Combining Strings, Creating New Elements

Figure 1a: Simple RL Combination

$$a_1\text{-}a_2\text{-}a_3 \quad + \quad a_3\text{-}a_4\text{-}a_5$$

$$\Downarrow$$

$$a_1\text{-}a_2\text{-}a_3\text{-}a_4\text{-}a_5$$

Figure 1b: Double Overlap

$$a_1\text{-}a_2\text{-}a_3 \quad + \quad a_2\text{-}a_1\text{-}a_4$$

$$\Downarrow$$

$$\begin{array}{r} a_1\text{-}a_2\text{-}a_3 \\ a_4\text{-}a_1\text{-}a_2 \end{array}$$

$$\Downarrow$$

$$a_4\text{-}a_1\text{-}a_2\text{-}a_3$$

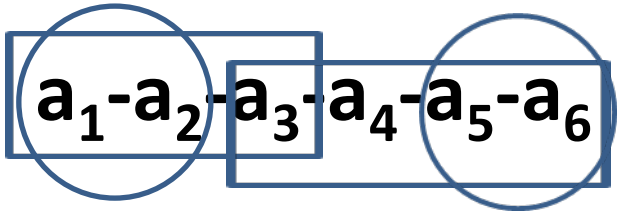
Figure 2: Sub-bundles

Figure 2a: Set of Possible Sub-bundles

Parents:

a_1-a_2-a_3

a_3-a_4-a_5-a_6



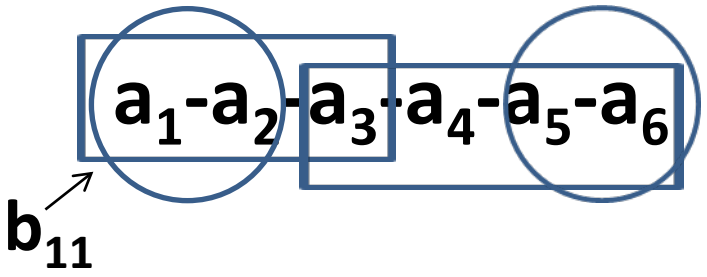
Grandparents:

a_1-a_2, a_2-a_3,

a_3-a_4-a_5, a_5-a_6.

Figure 2b. Signal-Generating Seed Element Blocks as Sub-bundles

Seed 1 contains b_{11}.



Seed 2 contains b_{21}.

Figure 3: Value Distribution

Body is Lognormal, Tail is Pareto

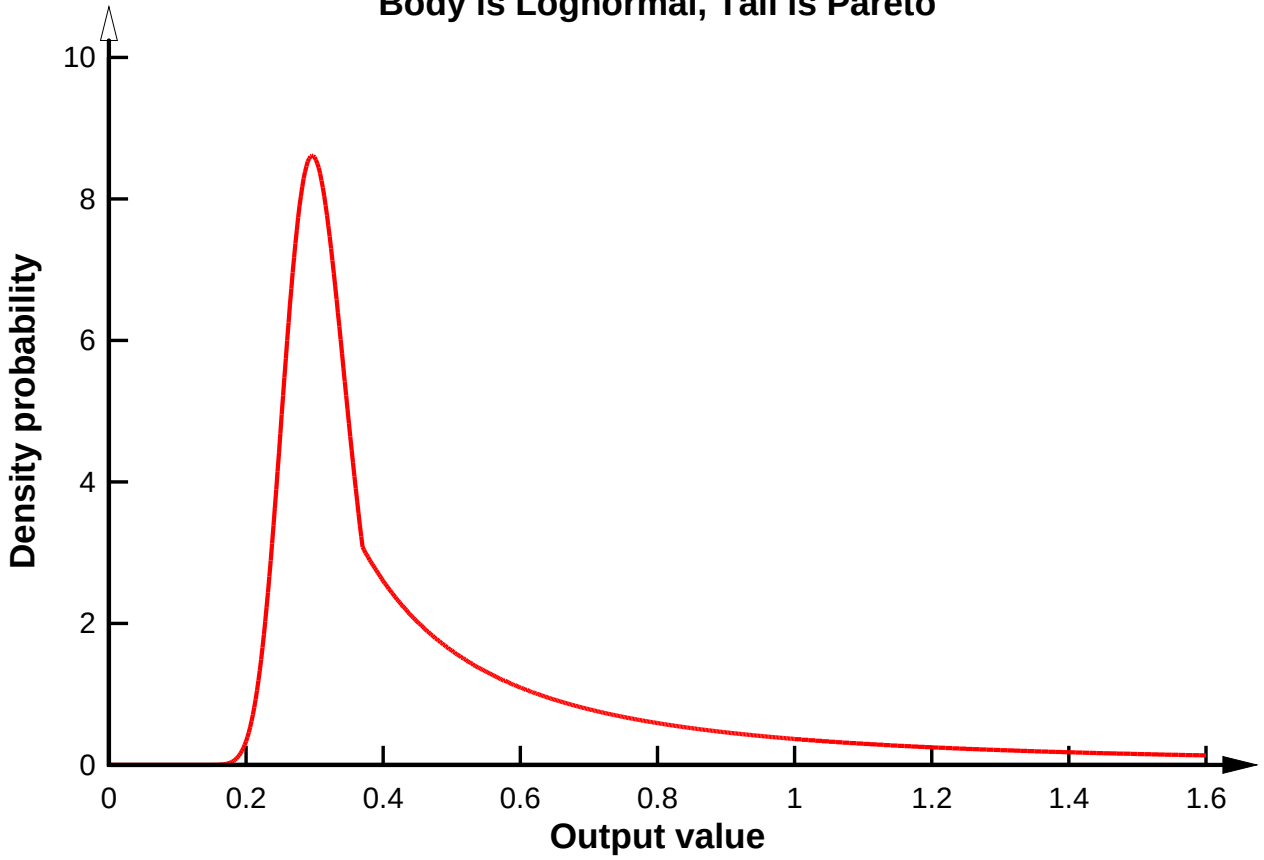


Figure 4: Decision Tree for Creative Development Strategy

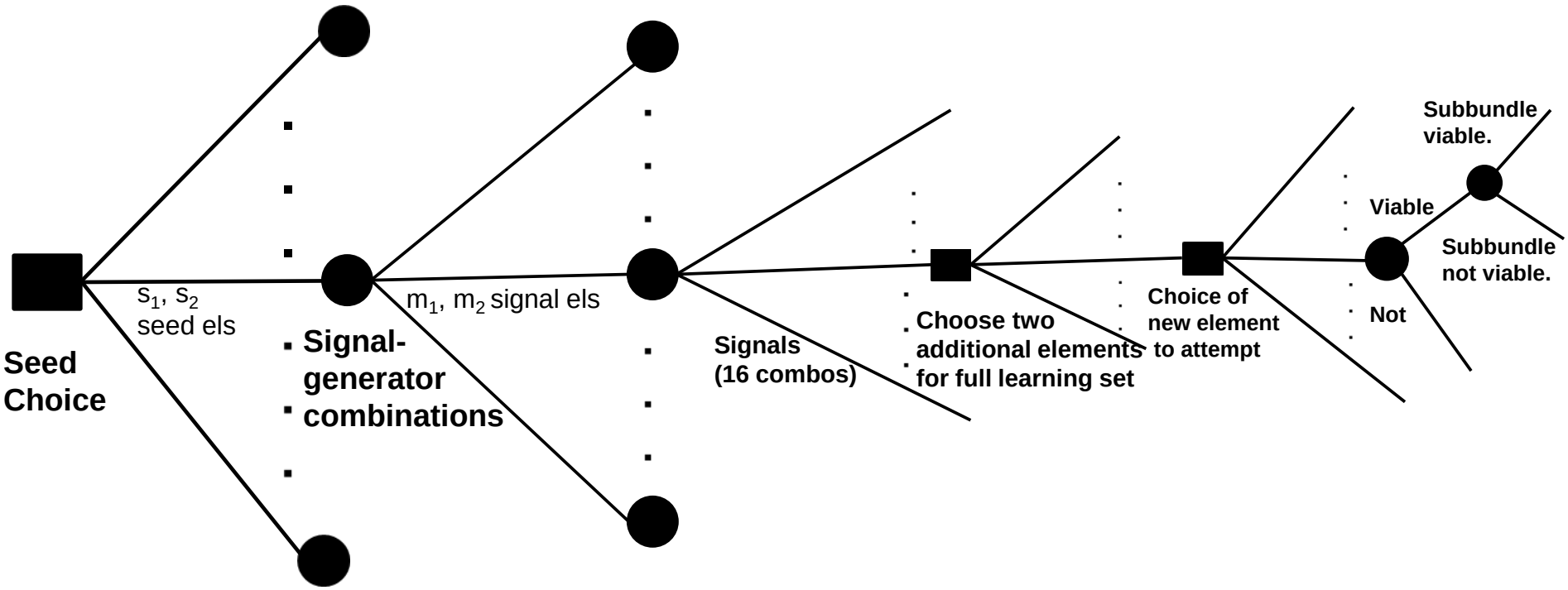


Diagram is schematic. There are $\binom{N}{2}$ possible seeds. Number of signals varies depending on seed elements: for m signals there are $\binom{m}{2}$ signal-generator combinations. Each pair of signal-generating elements has 16 different possible signal combinations, 4 for each signal generator (2 possible signals for viability, 1/0, and 2 for high value distribution). There are $\binom{N-2}{2}$ additional element choices for the full learning set. The cardinality of the set of new elements that can be attempted varies. If new element is viable, any subbundles that can be made are either viable and created or not viable. One example subbundle with outcomes is shown.

Figure 5: Tree Showing Paths of Development of the Field

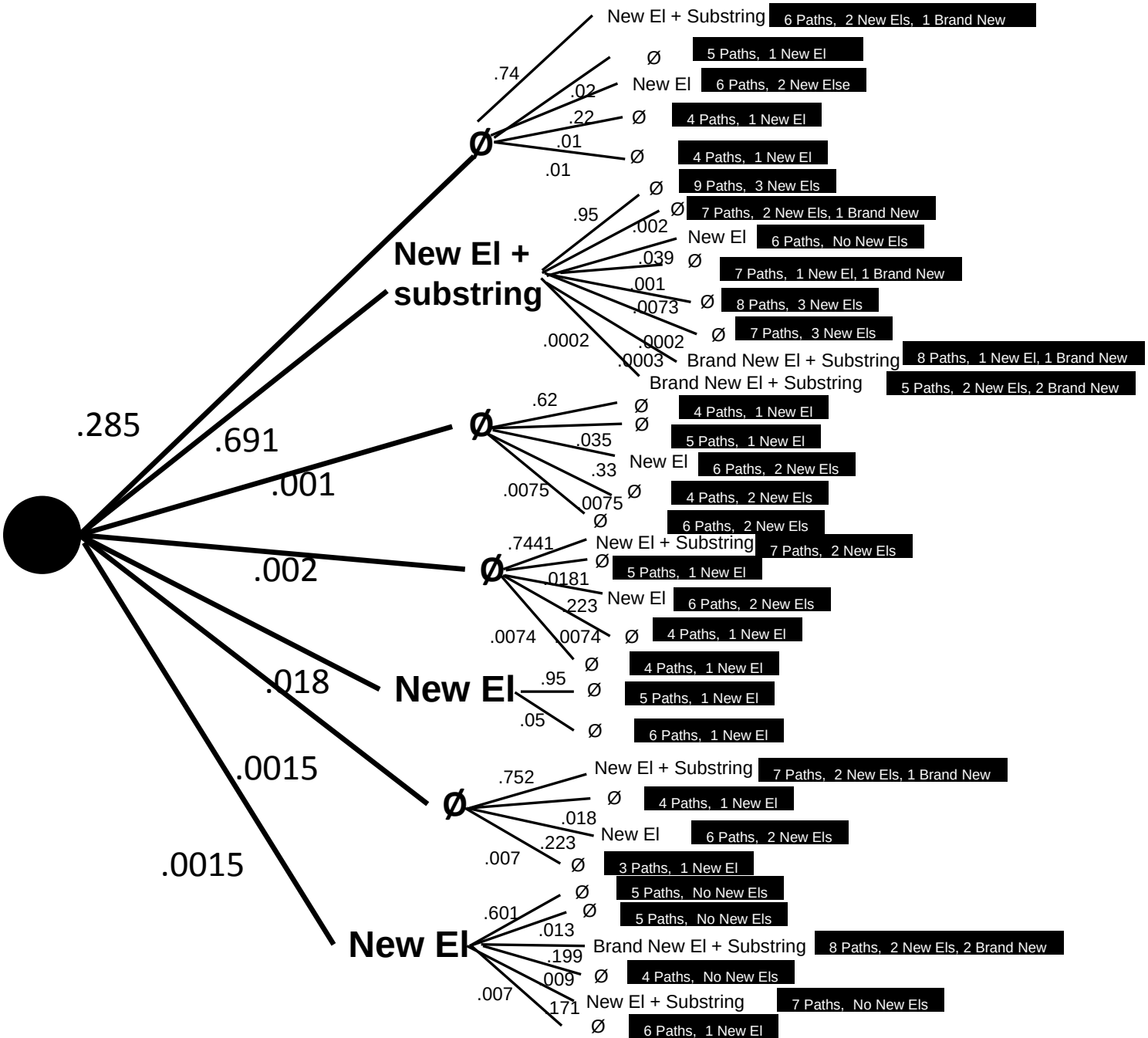
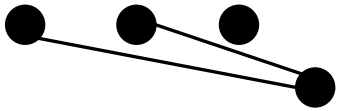
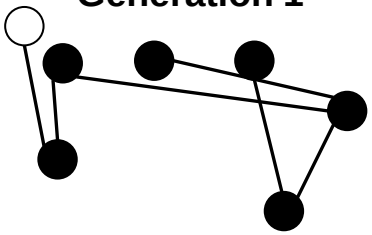


Figure 6: Field Structure Development: Example

Initial State

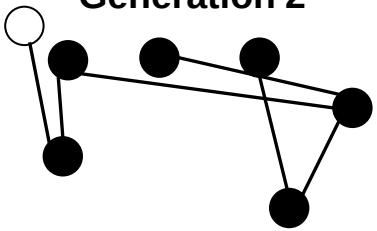


Generation 1



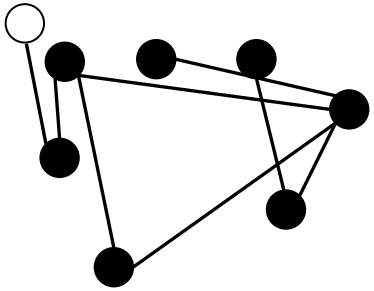
New El + Substring.

Generation 2



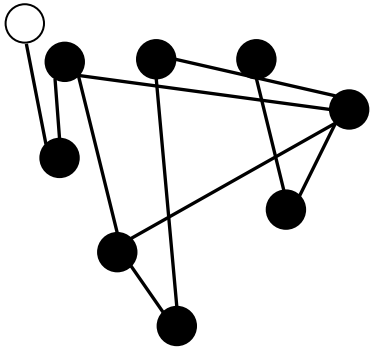
No New El Added.

Generation 3



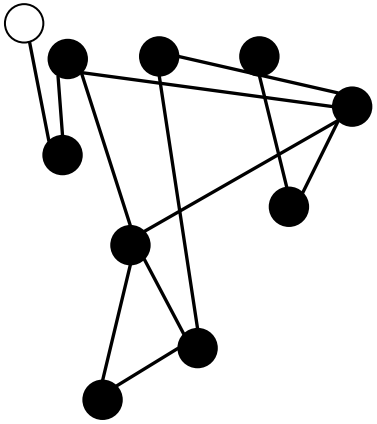
New El.

Generation 4



New El.

Generation 5

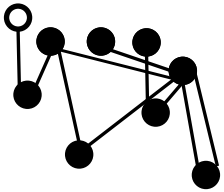


New El.

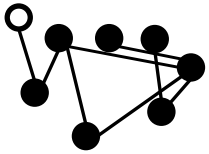
SUMMARY

Num Els Created: 5
Longest: 8
Height: 5 (Top Level attributes not shown except No. 1 shown as open circle.)

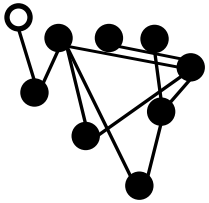
Figure 7: Field Structures: Examples



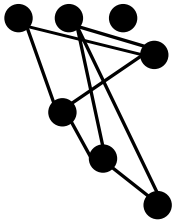
Clean path structure.
Probability: 3.7E-2



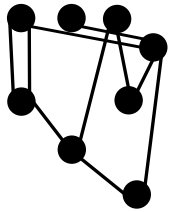
Highest Probability.
Probability: 0.34



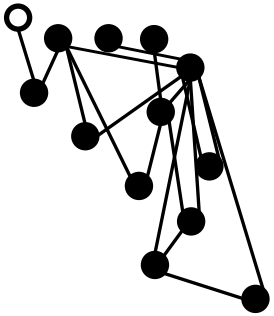
Second Highest Probability.
Probability: 0.18



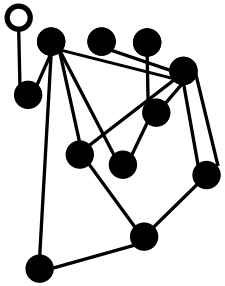
Probability: 1.0E-8



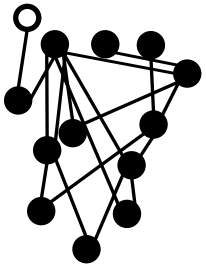
Probability: 6.6E-3



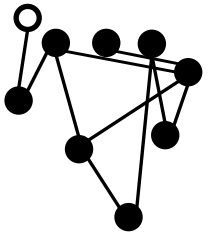
Probability: 8.0E-5



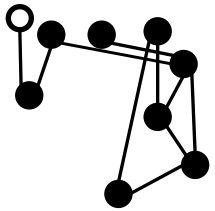
Probability: 4.8E-11



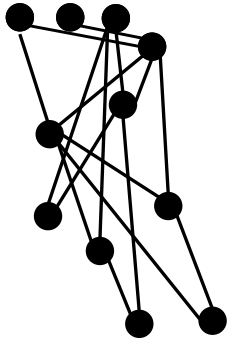
Probability: 3.0E-5



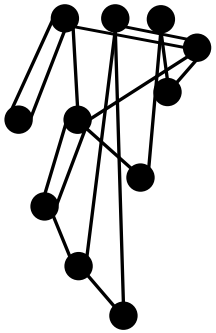
Probability: 2.1E-3



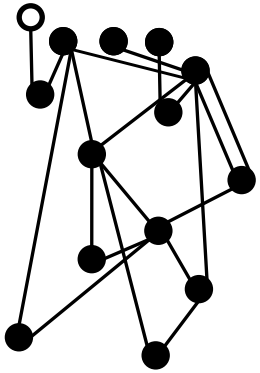
Probability: 8.3E-8



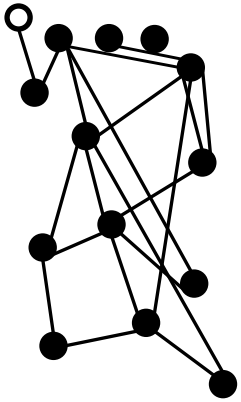
Probability: 2.5E-5



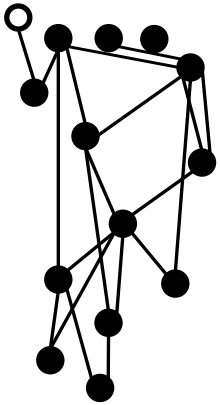
Probability: 1.2E-6



Probability: 3.0E-8



Probability: 1.0E-9



Probability: 1.7E-8

Figure 8: Number of Elements Created

Figure 8a: Distribution by Cumulative Probability of Paths Generating Given Number of Elements.

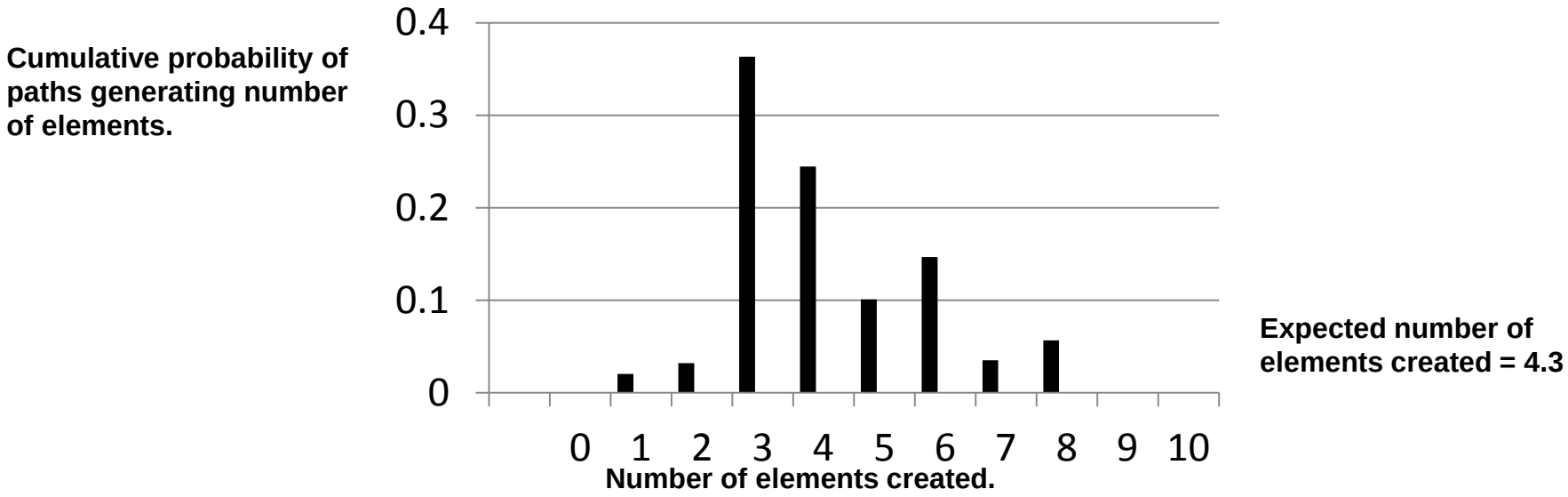


Figure 8b: Distribution by Number of Paths Generating Given Number of Elements.

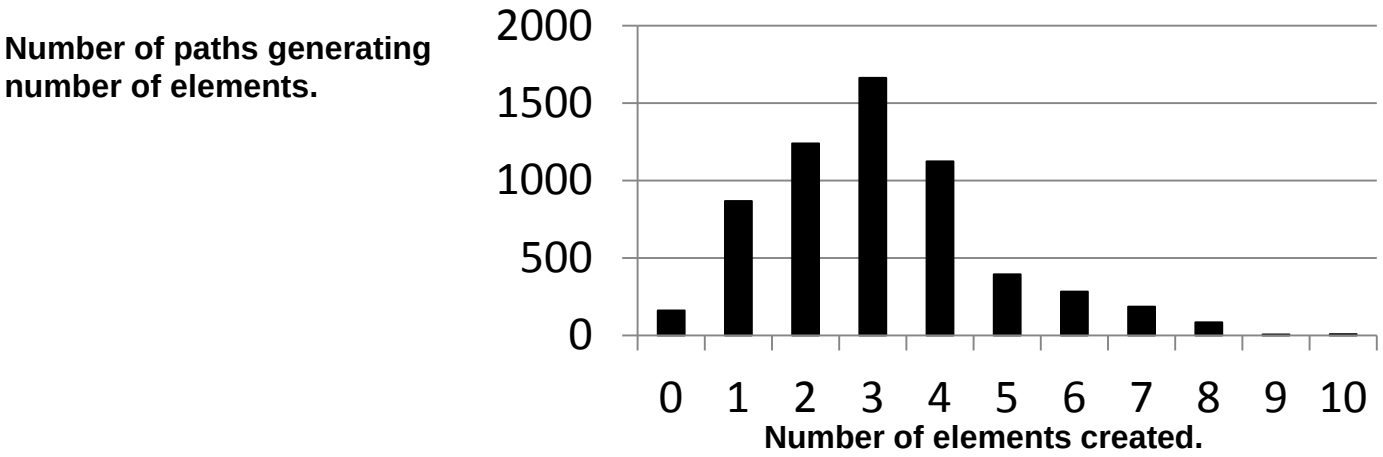
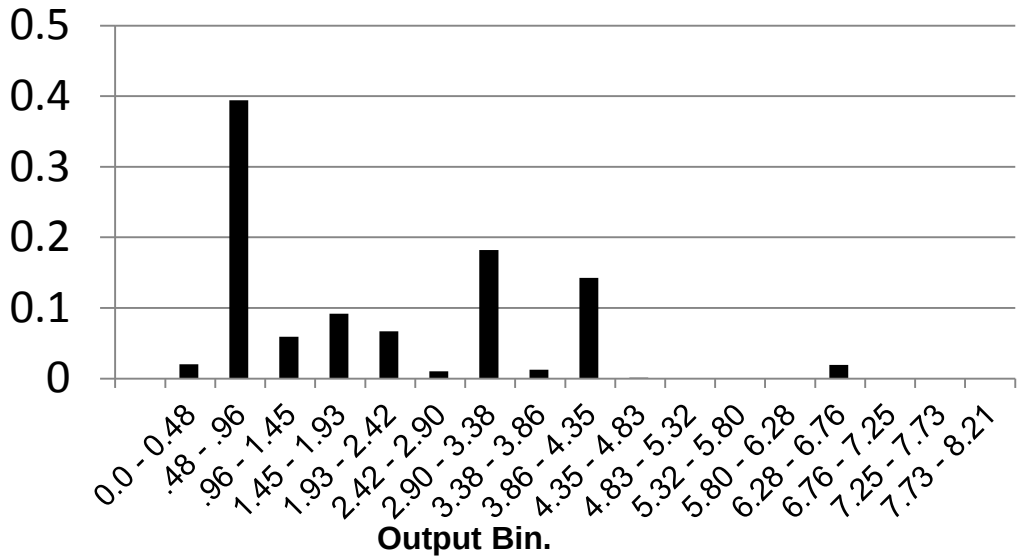


Figure 9: Output Distributions

Figure 9a: Distribution by Cumulative Probability of Paths Generating Output Level.

Cumulative probability of paths generating output.



Expected output = 2.1
Standard Dev. = 1.4

Figure 9b: Distribution by Number of Paths Generating Output Level.

Number of paths generating output level.

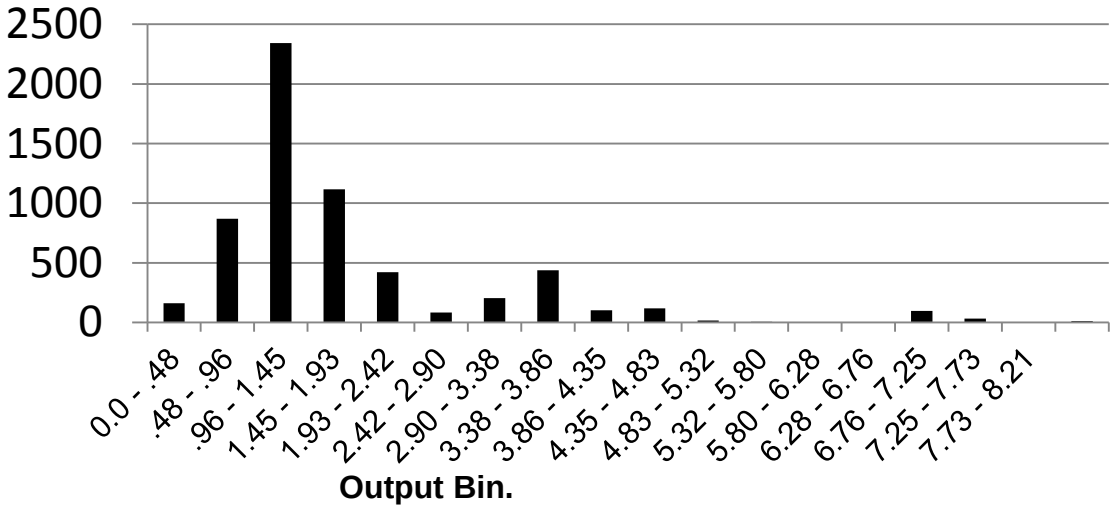
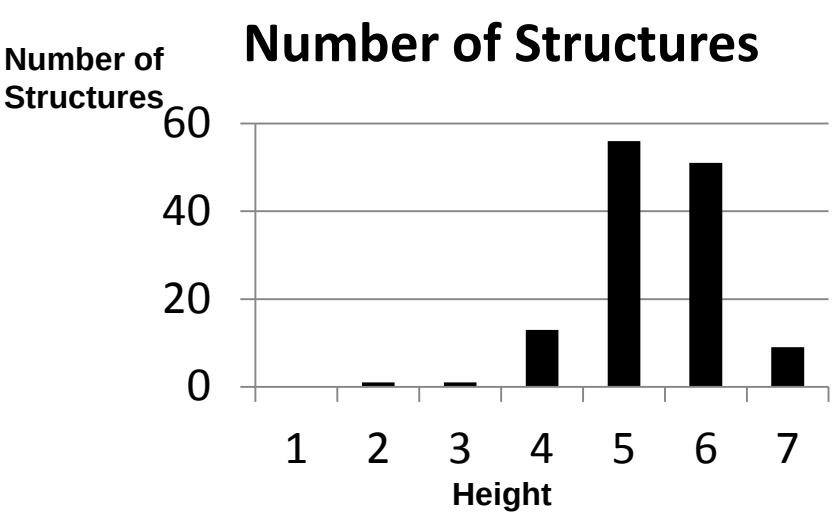


Figure 10: Height & Greatest Length Distributions

Figure 10a: Height.



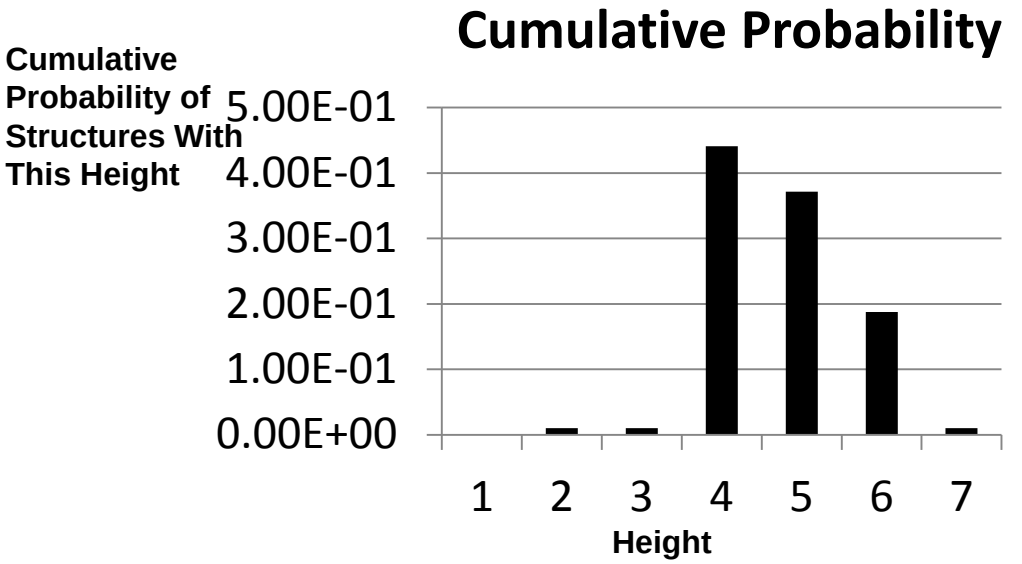
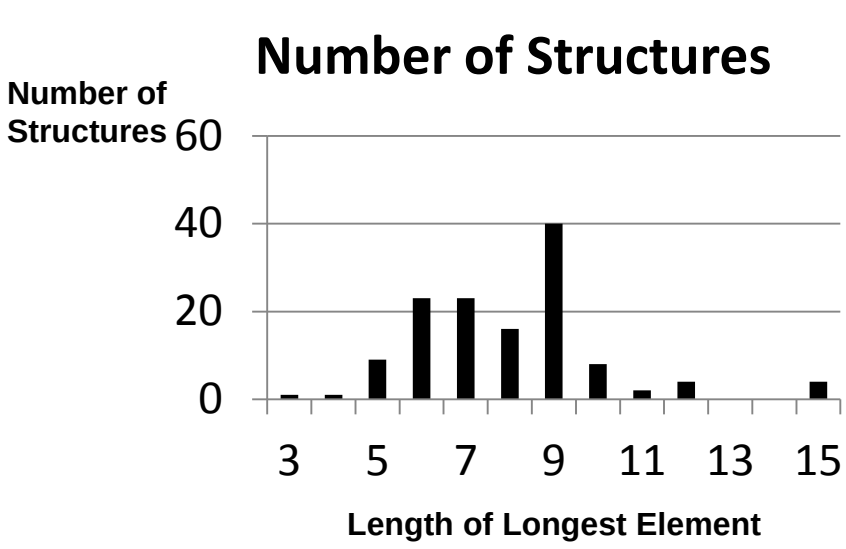


Figure 10b: Length / Complexity of longest element.



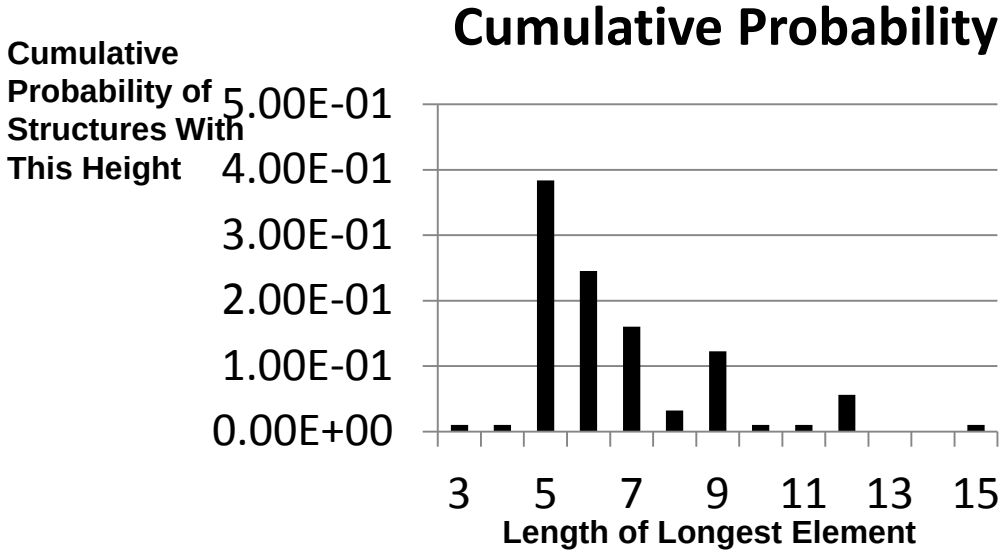


Figure 11: Number of Initial Elements Utilized

Figure 11a: Distribution by Cumulative Probability of Paths.

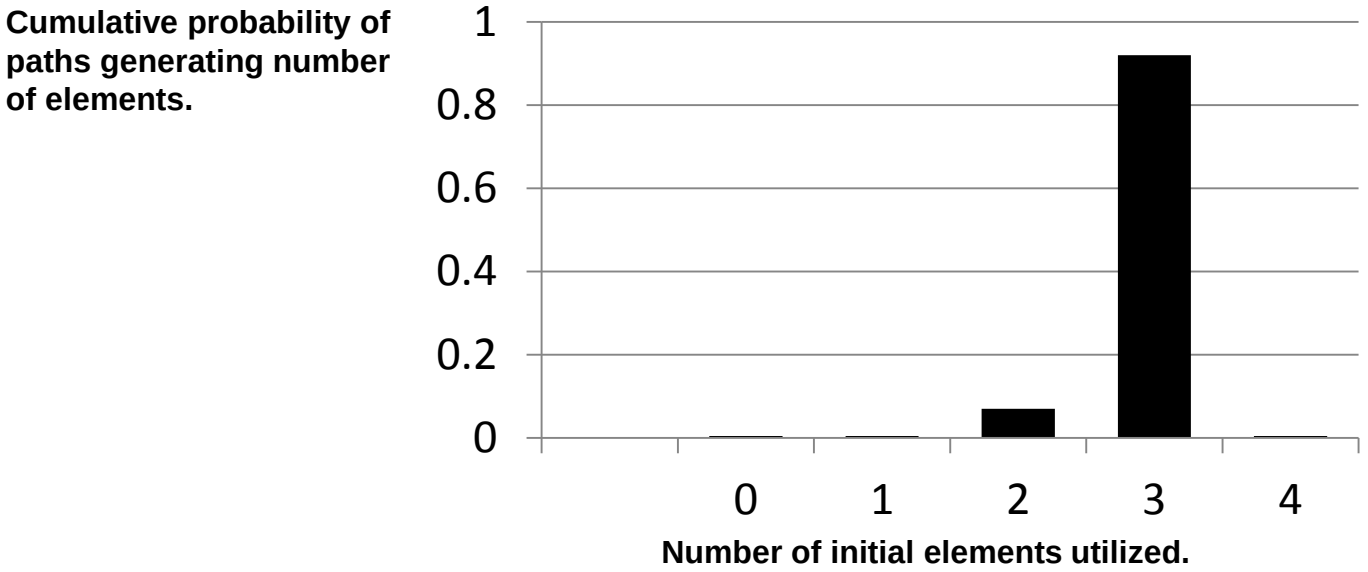
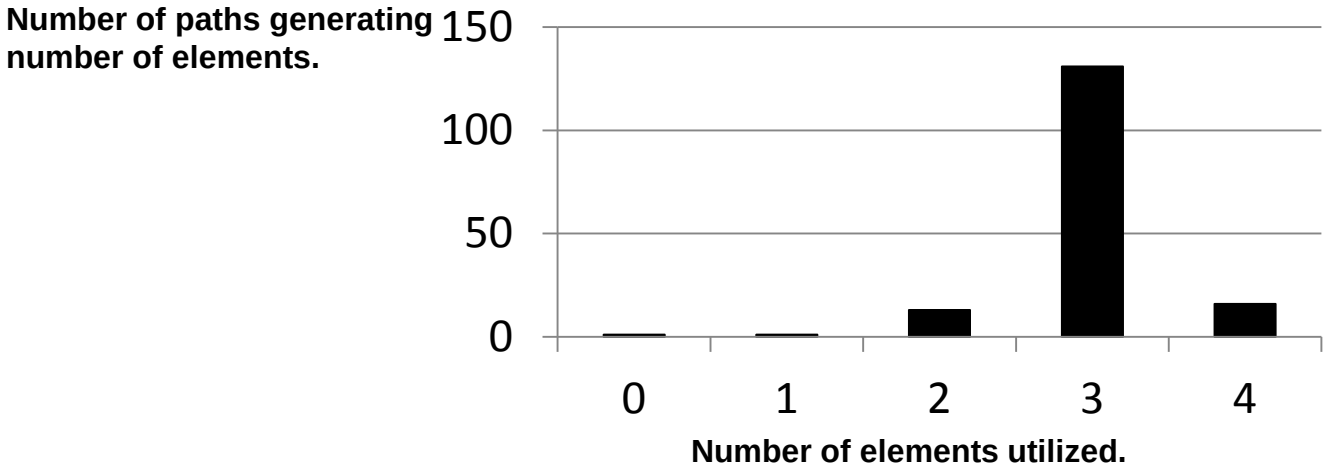


Figure 11b: Distribution by Number of Paths.



Note: “1” initial element being used occurs when one of elements 1 2 or 3 is combined with itself as parent of initial element 4.

Table 3: Conditional Output Based on Previous Period

	No Element Created Last Period.		Element Created Last Period.	
	Expected Output	Number of Paths	Expected Output	Number of Paths
All Paths	.399	4133	.363	1882
Clean Paths	.608	789	.330	426
Other Paths	.357	3344	.372	1456

Table 4: Path Dependence: Pairwise Comparison of Generation 1 Node Element Sets

Period 1 Node	Period 1 Node 2	3	4	5	6	7
1 29 Els	55 24 5 26	35 19 10 6	34 29 5 0	36 10 19 7	33 29 0 4	41 18 11 12
2 50 Els		57 18 32 7	56 28 22 6	55 12 38 5	56 27 23 6	61 19 31 11
3 25 Els			36 23 2 11	30 12 13 5	35 23 2 10	33 22 3 8
4 34 Els				39 12 22 5	35 32 2 1	42 22 12 8
5 17 Els					39 11 6 22	32 15 2 15
6 33 Els						41 22 11 8
7 30 Els						

Notes: Number below each node is number of elements created along all paths from that node.

In each cell 4 numbers are shown: (i) Number of elements in the union; (ii) number in the intersection; (iii) number in the row node set and not the column node set; and (iv) number in the column node set and not the row node set.

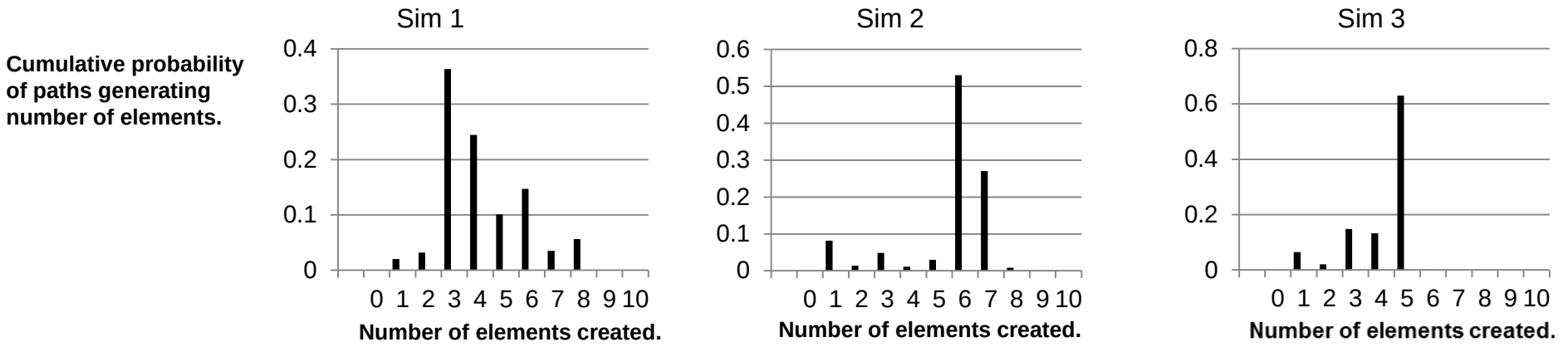
Table 5: Full Set of Base Scenario Simulations: Summary Statistics

Scenario

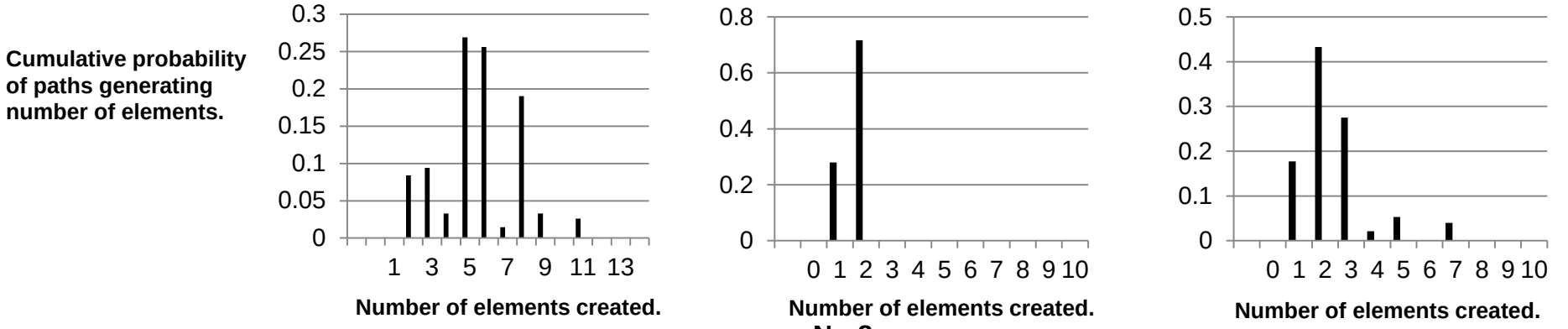
	N_0=3			N_0=4			N_0=8		
	Sim 1	Sim 2	Sim 3	Sim 1	Sim 2	Sim 3	Sim 1	Sim 2	Sim 3
Number of Paths	6015	5958	5511	3694	1246	2725	20,400	14,660	17,478
Number of Distinct Structures	131	20	111	136	5	48	92	9	143
Expected Height	4.75	5.43	4.68	4.84	3.00	3.98	4.29	3.05	3.90
Expected Max Length	6.55	7.50	8.96	6.97	4.00	5.06	5.59	4.07	5.00
Expected Number of Elements	4.24	4.25	5.63	5.68	1.72	2.50	5.27	2.05	3.59
CLEAN Expected Number of Elements	4.00	1.25	3.00	3.58	1.00	1.60	4.81	1.55	2.34
Expected Output	2.10	1.72	1.77	1.66	.505	.732	1.69	.627	4.46
CLEAN Expected Output	1.20	.345	.875	1.36	.323	.490	1.65	.475	.786
Conditional Output: All Paths: 0	.399	.371	.286	.264	.089	.143	.413	.139	.689
1	.363	.495	.328	.210	0.0	.196	.400	.044	.926
Conditional output: CLEAN Paths: 0	.608	.122	.240	.316	.102	.201	.355	1.69	.639
1	.330	.213	.300	.141	0.0	.140	.411	.012	.191
Conditional Output: Not Clean Paths: 0	.357	.402	.325	.250	.082	.117	.439	.120	.742
1	.372	.683	.365	.238	0.0	.211	.389	.081	1.23

Figure 12: Full Set of Simulations
Distributions of Number of Elements Created

a. N_0=3



b. N_0=4



c. N_0=8

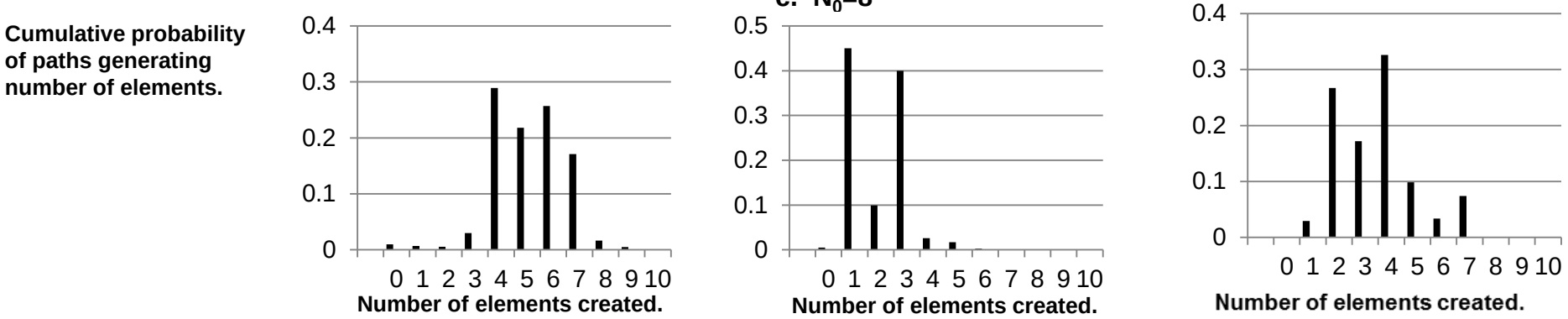
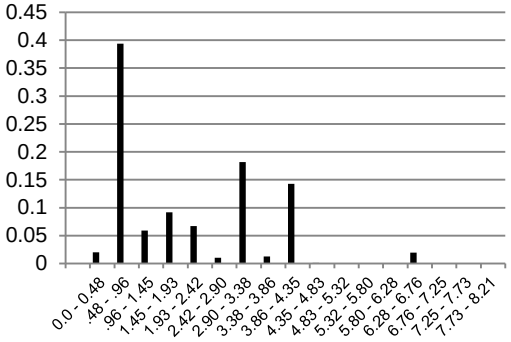


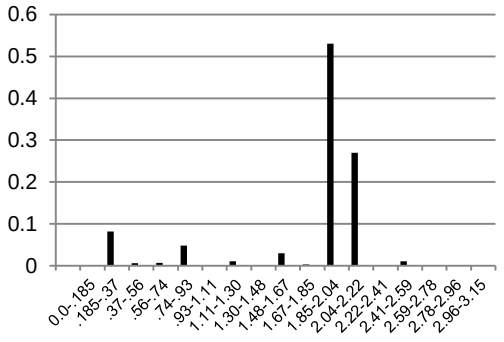
Figure 13: Full Set of Simulations: Output Distributions

a. N_0=3

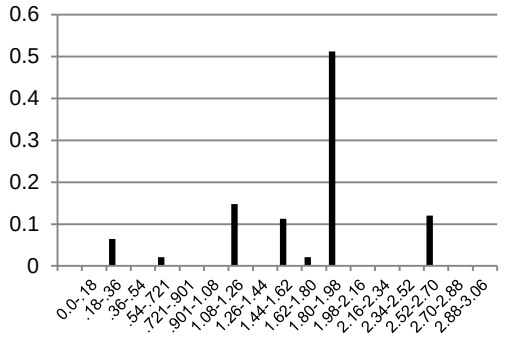
Sim 1



Sim 2



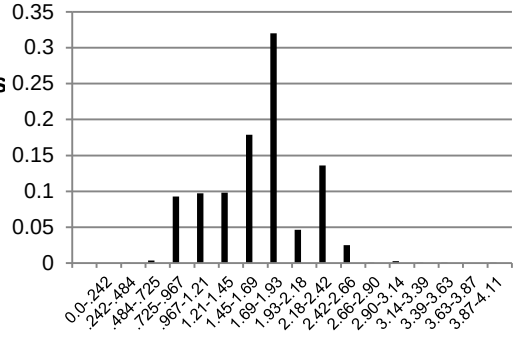
Sim 3

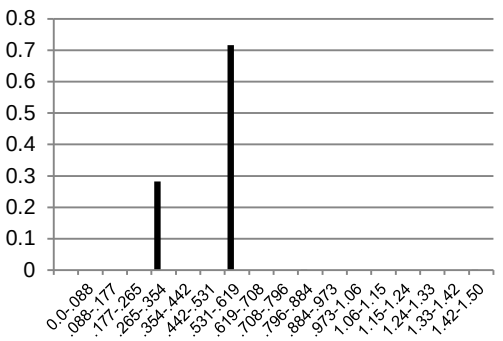


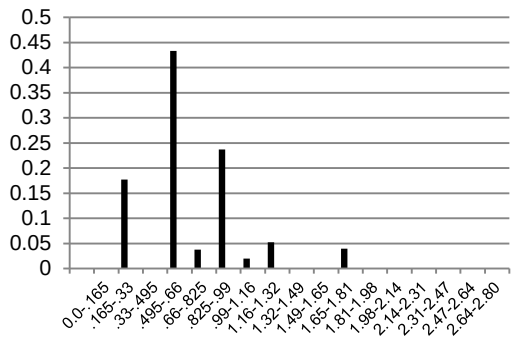
Cumulative probability of paths generating output.

Output Bin.

b. N_0=4



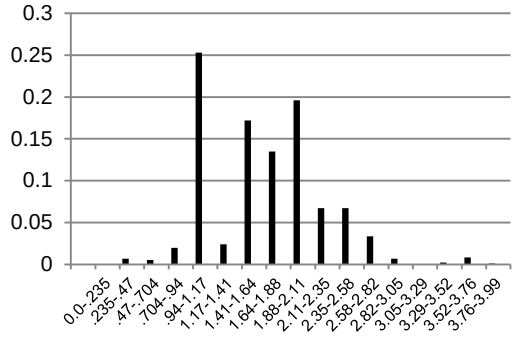


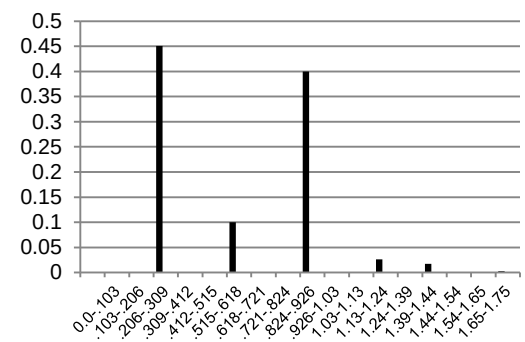


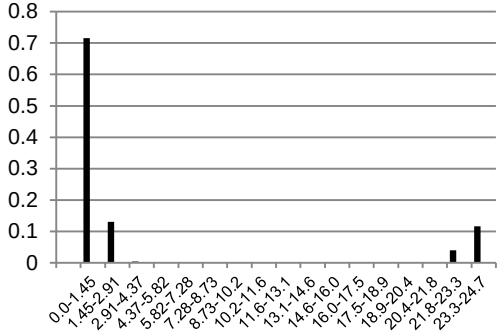
Cumulative probability of paths generating output.

Output Bin.

c. N_0=8







Cumulative probability of paths generating output.

Output Bin.

Output Bin.

Table 6: High Viability Probability: Summary Statistics

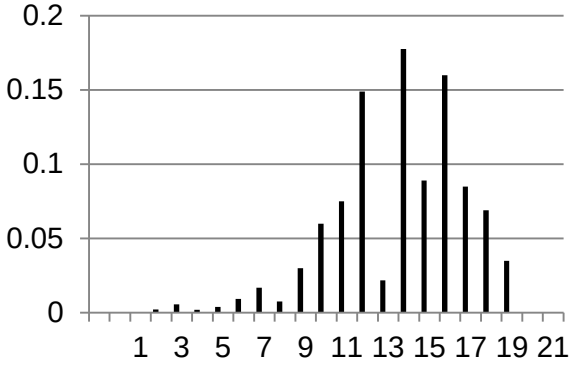
	Sim 1	Sim 2	Sim 3
Number of Paths	10,548	5531	2679
Number of Distinct Structures	1408	90	409
Expected Height	6.21	4.22	4.54
Expected Max Length	10.7	5.27	6.64
Expected Number of Elements	13.8	1.70	5.71
CLEAN Expected Number of Elements	9.86	1.25	6.33
Expected Output	4.15	.574	1.76
CLEAN Expected Output	3.52	.417	1.83
Conditional Output: All Paths: 0	1.00	.131	.516
1	1.25	.311	.515
Conditional output: CLEAN Paths: 0	.926	.171	.583
1	1.20	.237	.540
Conditional Output: Not Clean Paths: 0	1.01	.123	.488
1	1.27	.328	.510

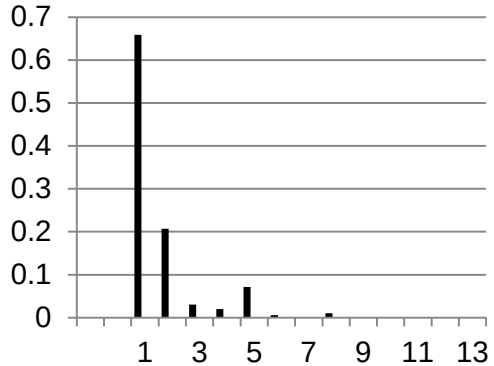
Figure 14: High Viability Probability: Distributions For Number of Elements and Output

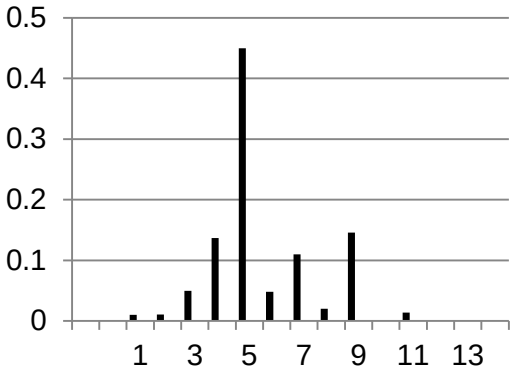
Sim 1 **Sim 2** **Sim 3**

a. Distributions of Number of New Elements

Cumulative probability of paths generating number of elements.

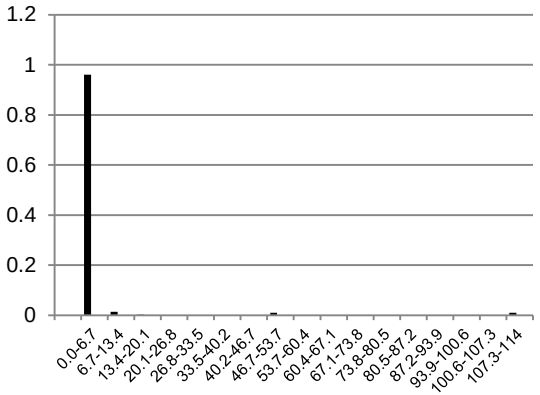


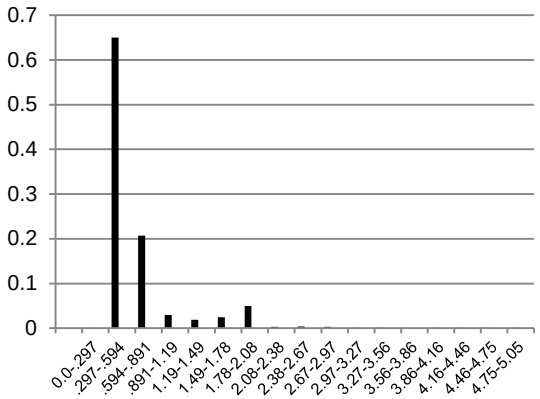




b. Distributions of Output

Cumulative probability of paths generating output.





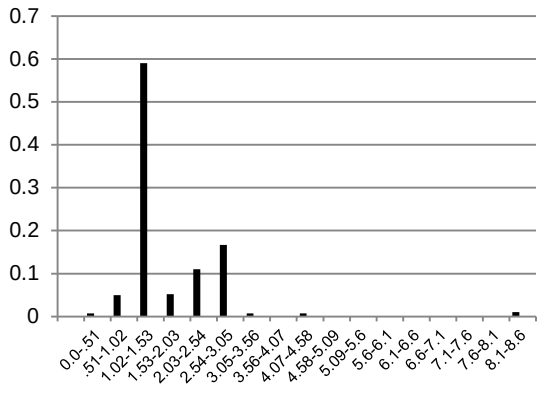


Table 7: High Viability Probability Path Dependence: Pairwise Comparison of Generation 1 Node Element Sets

Period 1 Node	2	3	4	5	6	7
1 182 Els	251 85 97 69	247 126 56 65	223 35 147 41	192 61 121 10	268 112 70 86	312 121 61 130
2 154 Els		262 83 71 108	190 40 114 36	170 55 99 16	262 90 64 108	320 85 69 166
3 191 Els			231 36 155 40	199 63 128 8	276 113 78 85	349 93 98 158
4 76 Els				121 26 50 45	235 39 37 159	292 35 41 216
5 71 Els					221 48 23 150	261 61 10 190
6 198 Els						382 67 131 184
7 251 Els						

Period 1 Node

Notes: Number below each node is number of elements created along all paths from that node.

In each cell 4 numbers are shown: (i) Number of elements in the union; (ii) number in the intersection; (iii) number in the row node set and not the column node set; and (iv) number in the column node set and not the row node set.

Figure 15: Royalty Comparison With Base Scenario