THE UNIVERSITY OF CHICAGO

COMBINATORIAL INNOVATION IN SCIENCE, TECHNOLOGY, ORGANIZATIONS, AND SOCIETY

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE SOCIAL SCIENCES

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

DEPARTMENT OF SOCIOLOGY

BY

LIKUN CAO

CHICAGO, ILLINOIS

AUGUST 2025

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I want to express my gratitude to all those who helped bring my projects to life. This dissertation is the result of countless days and nights of brainstorming, discussion, and collaboration—sometimes with intellectual clashes, and always with bittersweet feelings. I am deeply thankful that I was not alone on this journey.

First and foremost, I want to express my heartfelt gratitude to my advisor, James Evans, with whom I have had the privilege to work closely over the years. James has offered me invaluable guidance, inspiration, and support, helping me grow into a young scholar who can engage with the latest developments in the field, exchange ideas with him—and even challenge him at times—and explore new directions together. Like any true master and apprentice, our collaboration has been both rewarding and, at times, demanding—perhaps more often the latter from James's perspective. Over time, however, I've come to deeply appreciate how much I have learned from him—not only intellectually, but also through his dedication to work, his values in life, his commitment to community, and his example of leadership. These lessons and values have become a part of who I am, and I am truly grateful for them.

Now, to John Padgett, Lis Clemens, and Dashun Wang—I feel incredibly fortunate to have some of the world's greatest scholars as my committee members. John has taught me a great deal, from theorization to finding my place in academia. His work and ideas have been profoundly enlightening—like opening a window in my mind and letting the light pour in. Lis has consistently pushed me toward intellectual rigor, offering very detailed and inspiring feedback. As both an accomplished sociologist and a deeply supportive mentor, she has encouraged me to think early and seriously about career development—advice that has shaped my path in meaningful ways. Dashun has supported me in both research and career development through his influence and insights. I have benefited not only from our conversations, but also from rich exchanges with researchers in his lab. Many parts of this dissertation would not have come into being without the help of these brilliant minds.

I owe an immense debt of gratitude to several other mentors and friends as well, both at UChicago and beyond. John Levi Martin has offered invaluable advice over the years and has always lent a helping hand whenever possible. Lingfei Wu has been like an older brother, offering thoughtful insights since the first year of my Ph.D. journey—wisdom from which I continue to benefit. Letian Zhang has provided generous guidance in both research and publication. I am also deeply grateful to many others who have guided me at various stages of my journey, both during and prior to my Ph.D. program: Lu Zheng, Hui Zheng, Jarder Luo, Xibao Li, Kai Li, Pete Aceves, Ronald Burt, Nicole Marwell, and René Flores. I sincerely appreciate the opportunities to connect, exchange ideas, and arrive at insights through thoughtful dialogue—even when it sometimes meant revisiting long-held assumptions.

Next, to my friends, collaborators, and fellow travelers: Ziwen Chen, Rui Pan, Chen Liang, Haochuan Cui, Jeffery Lockhart, Eamon Duede, Liang Cai, Yulin Yu, Yian Yin, Binglu Wang, Jie Ren, Yuanyi Zhen, Jinxian Shi, and many other brilliant folks from Knowledge Lab, UChicago sociology department, and the science of science community. You are not only sharp-minded colleagues, but also dear friends. I feel incredibly lucky to know you, and I hope this mutual support continues for many years to come.

Finally, there is my family—my grandmother Fangying Zhang, my mom Jun Rao, my uncle Hong Rao, my aunt Xiaoxia Rao, and cousins. It has taken me many years—years of struggling through countless social layers, of journeying across mountains, oceans, and distant continents—to truly understand the depth of spiritual foundation required for a family to thrive

across so many different domains, and through the shifting tides of history. You have supported me throughout this academic journey, including its darkest moments. Your life experiences have taught me that things can, and often do, work out—and that each tomorrow holds the promise of a fresh new day.

ABSTRACT

Knowledge recombination theory has been widely accepted and applied across disciplines, including sociology, management, and data science. However, its underlying promise of a generalized explanation has been challenged theoretically and lacks consistent empirical support. This project therefore explores the mechanisms through which innovation emerges, and how these mechanisms unfold in unique social contexts. Drawing on five large-scale datasets from key knowledge production systems—science, technology, business strategy, art, and online knowledge platforms—I use advanced statistical methods and computational tools to discuss the generality and specificity of the recombination theory across these domains.

The three chapters focus on different dimensions of combinatorial innovation. The first chapter closely examines invention in the context of entrepreneurship, drawing on theories of organizational emergence and strategic positioning literature. We find that startups perform best when they combine several mature technological applications within the existing economic system, rather than creating entirely new application scenarios. This study identifies risk distribution as a key predictor of recombination outcomes. Startups that mobilize "higher-order invention" —invention across subsystems rather than within them—receive disproportionate market rewards. Meanwhile, preferences for risk distribution may differ across other systems.

The second chapter uses patent data to examine combinatorial trajectories in technology. I first constructed a technological map using Poincaré embedding, a representational learning algorithm designed to capture hierarchical structures. This map effectively reflects the architecture of the knowledge tree. Based on knowledge trajectories mapped onto this space, I find that deep search—combining closely related knowledge—gains rapid attention within the community, while broad search—linking distant elements—accumulates more attention over

time by attracting a wider audience. This study highlights the availability of knowledge components as a key dimension of combinatorial innovation: the more intuitive the combination, the lower the risk, but also the lower the potential for disruptive innovation.

Building on findings from the first two chapters, the final chapter turns to the classic topic of team design. In this study, we distinguish between two dimensions of diversity in teams: background diversity—the differences in team members' knowledge backgrounds, which influences the availability and readiness of knowledge components; and perspective diversity—the differences in how team members interpret the same issue, which affects the distribution of cognitive risk. We find that teams with low background diversity and high perspective diversity consistently perform better across all five domains. This study illustrates how a generalizable social law in innovation can be identified while still accounting for contextual particularities.

This research advances theories in the sociology of knowledge, group processes, and organizational studies. For sociology of knowledge, particularly theories related to science and technology, it highlights the boundary conditions of recombination theory. For group process theories, it introduces a new concept and analytical framework for assessing knowledge complementarity. For organizational science, it proposes several future directions for exploring new dimensions of ambidexterity.

CHAPTER 1 INTRODUCTION

1.1 Engine of Economy, Core of Humanity

Innovation is widely recognized as the primary engine of economic growth and international competitiveness (Romer 1990; Schumpeter 2008). In the United States, policymakers have long regarded science and technology as pivotal to global leadership, often referring to this pursuit as the "endless frontier". This concept traces back to Vannevar Bush's influential 1945 report, inspired the establishment of the National Science Foundation in 1950, and led to the passage of major innovation bills even today (Gross and Sampat 2022). Delays in innovation, by contrast, always generate anxiety and fear. In a 2012 conversation with Fukuyama, Peter Thiel expressed his concerns about the slowdown in scientific and technological progress:

I believe that the late 1960s was not only a time when the government stopped working well and various aspects of our social contract began to fray, but also when scientific and technological progress began to advance much more slowly.

These concerns undoubtedly shaped both his investment strategies and, eventually, U.S. policy in 2025. Innovation, therefore, extends far beyond scientific inquiry; it carries profound political and historical consequences that shape the lives of millions and across centuries.

In the meantime, innovation is crucial in another sense: it can be seen as one of the defining features of humanity. In her seminal work *The Human Condition*, Hannah Arendt (1970) identifies three fundamental human activities: labor, work, and action. Labor corresponds to biological necessity, work to the creation of lasting objects, and action—the highest form of human activity, also the highest expression of freedom from Arendt's view—represents the capacity to initiate unpredictable and irreversible events. Through the act of "beginning anew," humans possess the potential to act in ways not determined by what has come before. Innovation, in this light, is not merely possible, but intrinsic to what it means to be human.

These two perspectives foreshadow the complex theoretical landscape of innovation science. While more than 10,000 papers are published annually on this topic[1], reflecting growing scholarly interest, researchers from different disciplines often hold contrasting theoretical views and value orientations. We see some pursuing innovation for efficiency, others for freedom; some focusing on discrete events, others on continuous processes; some emphasizing particularity, others seeking generalizability; and some viewing material factors as primary drivers, others emphasizing culture or social institutions. Innovation science has become an intellectual mosaic in this sense—rich in shared terms but shaped by divergent paradigms.

This is mostly true for innovation genesis, or what can be simply referred to as invention. This is a domain where theoretical tensions are most evident and, as a result, hold the greatest potential for conceptual advancement. It is the focus of this dissertation.

1.2 Theoretical Insights into the Genesis of Innovation

Conceptualized as "invention", the genesis of innovation has been well discussed in prior literature (Padgett and Powell 2012). Complexity scholars have argued for a deep similarity between technological, social and cultural innovations and the process of biological evolution, whereby sex and horizontal transfer combine genes in new, complex combinations that reproduce or die (Erwin and Krakauer 2004a). Padgett and Powell take this metaphor further and argue that new social and cultural structures are like autocatalytic chemical processes, where

[1] Based on statistics from the Web of Science.

cascading reactions form novel self-sustaining cycles (2012) in processes that reflect the biochemical origins of life (Kauffman 1996). These discussions give rise to a core theoretical assumption that has shaped the framework of empirical analysis in subsequent social science research: that disorder and disruption—whether at the systemic or individual level—are essential preconditions for the emergence of innovation. Through the process of recombination, in which elements of prior products or processes are integrated into something new, inventions "expand the adjacent possible"—the space of potential combinations one step beyond the current frontier (Kauffman 1996). Grounded in this theoretical perspective, subsequent empirical studies have identified similar social mechanisms of invention across multiple levels of analysis.

Innovation at the macro, societal scale.

The innovative capacity of civilizations rises first through destructive creation, as disruptive—even traumatic—shocks dismantle existing social, cultural, and material structures. Events such as natural disasters and war unsettle social order and redistribute resources and human capital, leading to a pace of innovation rarely seen in stable societies. A notable example comes from Paul Forman (1971)'s historical analysis of Weimar culture and quantum mechanics. The loss of WWI and the memory of horrors and chaos it unleashed on Europe, upended German society and deeply affected its intellectuals. Feeling that notions of 'causality' had inspired Germany's national identity and aggression in the war, German philosophers articulated an acausal nihilist worldview that broke down assumptions of rationality and determinism. Forman's detailed historical analysis demonstrated that physicists from the Weimar Republic not only read, but in some cases wrote philosophy in this vein. It was these and nearby physicists who imagined the possibility of quantum mechanics, an acausal physics, which would not have emerged at that time and place if hostility against science, rationality and materialism, including

Newtonian causality, hadn't broken their connection with contemporary views of physical reality. Historians drew the 'Foreman thesis' from this demonstration, that culture furnishes at least part of the imagination underlying new scientific discoveries, and the milieux from which they become accepted. This hints at our more specific thesis that wars, natural disasters and other society-scale calamities break apart established certainties into fragments that may be recombined to form new ideas and institutions.

Society-scale conflicts and disasters not only supply, but demand new forms of innovation. During such periods, military and health-related funds and associated resources are unleashed during such periods, enabling transformative science, expertise and technology (Lee 2016). In the mid-20th Century, WWII catalyzed nuclear power and explosives, large-scale computation, radar, and a number of medical techniques (Gross and Sampat 2020). For countries, desperate incentives for survival and competitive priority under novel circumstances call forth these technological breakthroughs (Hanlon and Walker Hanlon 2015), which have wide ranging impacts once war is over. Recent work demonstrates that even economic crises, which result in reduced innovation investments among most firms, stimulate an increase of innovations among the most innovative firms exploring changed product and market environments (Archibugi, Filippetti, and Frenz 2013).

In brief, by disrupting existing institutions, large-scale societal shocks facilitate creative recombination by dislodging elements of nature, culture, and society that were previously unexposed and unavailable to the forces of creative destruction.

Innovation within communities and networks.

Beneath long-term cycles of societal conflict and catastrophe, innovation emerges more routinely from "creative conflict" (Stark 2011a) among groups and individuals who bring different perspectives and lived experiences.

Social network theories highlight the importance of network position in shaping innovation capacity. Granovetter's classic work on the "strength of weak ties" (1973) discusses how weak or distant ties—connections with acquaintances or even strangers—provide access to novel information beyond the boundaries of close-knit groups. Unique, high-value information that passes along weak ties bursts self-reinforcing communication bubbles by weakening the structure and hold of local information. Expanding on this idea, Burt (Burt 2004) argues that the benefits of information flow depend not only on tie distance but also on network structure. Brokers who uniquely connect distant groups across "structural holes" in the network gain broader vision and access to non-redundant information. The resulting diversity—and even conflict—of perspectives and preferences break the unity of established understanding and community boundaries, revealing new opportunities for creative recombination.

The past several decades have produced substantial empirical evidence supporting these claims. Scientists, inventors, cultural creators, organizations, and online contributors with greater access to dissonant information—particularly across otherwise disconnected individuals and groups—consistently show stronger innovation performance (Aral and Dhillon 2023; Bell 2005; Linzhuo, Lingfei, and James 2020; Ter Wal et al. 2016; Wang 2016).

Innovation also occurs within teams,often in proportion to the level of cognitive conflict they sustain. Research on rap music (Lena and Pachucki 2013), video games (De Vaan, Stark, and Vedres 2015), political communication (Kreiss and Saffer 2017), scientific discoveries (Yang 2024) and technical inventions (Choudhury and Haas 2018) have shown that radical innovations emerge when cognitively diverse teams combine previously disconnected ideas (Shi and Evans 2019). The presence of cognitive disagreement is shaped in part by team composition, including factors such as size and hierarchy. Smaller, flatter teams tend to produce disruptive innovations by activating low-level cognitive conflict between team members holding diverse expertise (Wu, Wang, and Evans 2019; Xu, Wu, and Evans 2022a). At the same time, cognitive conflict may also arise from tensions between the team and external social structures, rather than from within the team itself. Students from minority groups, whose mindsets may differ from those of the mainstream society, have been shown to innovate more (De Dreu and West 2001; Hofstra et al. 2020; Sutton 2002). Similarly, new artists are often associated with the emergence of new musical genres, more so than collaborations among established artists (de Laat 2014). In science and technology, some of the greatest and most disruptive breakthroughs occur when "expeditions" of investigators from one area travel to another cognitively diverse region to solve their challenges in unprecedented and unfamiliar ways (Shi and Evans 2019).

The innovative potential of both networks and teams can also stem from temporal dynamics—for example, the formation of new structural holes over time (Kumar and Zaheer 2019), rotating leadership between collaborating firms (Davis and Eisenhardt 2011), and mounting pressures from external competition (Aghion et al. 2005) or internal conflict (Kidder 1981). The more strongly these dynamics disrupt the self-reinforcing structure of localized information, the more likely they are to unleash imagination and open new pathways for innovation.

Innovative conflict within individuals.

Finally, we arrive at the most fundamental level of analysis—the individual innovator. Much of this discussion takes place within psychology, a field with a long tradition of exploring the human mind. Remarkably consistent with the broader themes discussed earlier, psychological research shows that internal conflict—within the mind and body—is associated with novel ideas, designs and action. Such internal conflicts arise from the interaction of diverse skills, knowledge and experience within an individual, as well as from the impact of traumatic or challenging life events.

Diverse experiences have been experimentally shown to enhance cognitive flexibility and the creation of novelty (Barbot 2022; Ritter et al. 2012). When individuals are embedded within complex environments and face conflicting demands, they may adopt paradoxical mental frames and become more likely to recognize and embrace contradictions, stimulating creativity (Miron-Spektor, Gino, and Argote 2011; Ritter et al. 2012). This pattern scales up: individuals with broader experience are capable of coordinating others across that breadth, yielding more collective creativity (Vries et al. 2014).

Meanwhile, creativity can also arise from traumatic or deeply challenging personal experiences (Forgeard 2013). For instance, people who hold a conflict mindset—who are led to perceive others as their adversaries—have been shown to hold broader and more inclusive cognitive categories (De Dreu and Nijstad 2008). Likewise, jarring experiences with different cultures or facing divergent stereotypes, which challenge personal biases, can stimulate creative thinking (Crisp and Turner 2011; Gocłowska, Crisp, and Labuschagne 2013). This ties in with the way diverse experiences create internal conflict and open the space for novel thinking.

1.3 The Three Fault Lines of Innovation Science

These social science theories on the origins of innovation, while offering valuable insights, reveal three critical fault lines in theorization.

The paradox of dynamics and statics.

The first paradox stems from the nature of innovation as a social process. While many studies focus on the "innovation event"—the moment when a new structure emerges from the old—innovation gains meaning only when viewed through a historical lens. Both old and new forces shape a given period, and the transition between them, though important, represents only a small part of the broader historical trajectory. The mechanisms driving this shift, its intensity, and the social landscape surrounding the transition are shaped by forces that lie beyond the immediate context of the transition.

Therefore, when we extend the time frame of observation, the first paradox becomes evident: forces that once propelled innovation can, and often do, turn into conservative forces that resist subsequent waves of change. The success of past innovations tends to enrich and empower earlier innovators and institutions, and suppresses the emergence of new ideas in the next generation. As a result, disruptive forces often arise from the peripheries of science, technology, and society. Marx offers a powerful illustration of this paradox, arguing that technological advances in capitalist society increase the rate of relative surplus-value, narrowing the accumulation of capital to the winners of industrial competition, enlarging the wealth gap and ultimately incentivizing the destruction of capitalism from the inside (Anon 2013). Supporting this argument, recent studies have also provided empirical evidence of growing inequality driven by capitalist competition (Piketty 2014; Piketty and Zucman 2014).

This paradox finds its ultimate form in recent trends in innovation productivity: while our understanding of innovations has grown, we have observed a decline in innovative activity across both society and the economy. U.S. and OECD economies have seen a marked decline in business dynamism. Market concentration has increased, oligopolies have emerged, diversity and labor productivity has dropped, knowledge diffusion has stalled, and quality-improving innovations from incumbent firms increasingly outweigh destructive innovations by entrants (Akcigit et al. 2021; Akcigit and Ates 2019, 2021; Dowd 2004; Garcia-Macia, Hsieh, and Klenow 2016). While cultural innovations like the emergence of rap have changed the popular music scene, other cultural industries, from Broadway to Orchestras have become conservative, especially for incumbents, partly to satisfy conservative tastes held by elite consumers (Kremp 2010; Phillips and Owens 2004; Thomas 2019). Science has slowed in innovation as well. The number of scientists and papers has increased dramatically over the past few decades, but individual scientist productivity has declined (Bloom et al. 2017; Jones 2009; Thomas 2019). Paradoxically, increases in research quantity have slowed scientific progress by making it harder for new ideas to gain widespread attention (Chu and Evans 2021). Citations flow to well-known researchers, making it more difficult for outsiders to challenge field leaders (Zivin, Azoulay, and Fons-Rosen 2019). Furthermore, fast-paced research gives a competitive edge to specialists, while research spanning disciplines is slower and involves greater risk (Leahey, Beckman, and Stanko 2017; Teodoridis, Bikard, and Vakili 2019).

The fracture of explanation pathways.

Within early genres of cultural criticism and social commentary, scholars and writers have consistently envisioned the social, cultural, and material consequences of emerging technologies and their externalities. In these accounts, materials were often portrayed as the primary driving force of innovation. For example, nationalism and the idea of “imagined communities” have been linked to the rise of modern print technologies (Anderson 2016). Meanwhile, the rise of the television industry, with its fragmented flow of information, is believed to have eroded the space for critical thinking and social resistance (Postman and

Postman 2005). Similar ideas have also appeared in earlier waves of social inquiry (Ogburn 1947), the natural sciences from biology to physics (Diamond 2013; Le Couteur 2003; Mukherjee 2010), and science and speculative fiction (Heinlein 1947), ranging from hard science fiction to apocalyptic and climate narratives.

In the social sciences, we encounter a different set of implicit theoretical assumptions. Economics has centered on rational human action as the engine of cumulative progress, psychology on individual motivation and disposition, and sociology on social and cultural institutions—each, in its own way, downplays the determinative role of materials and emphasizes the role of culture and society.

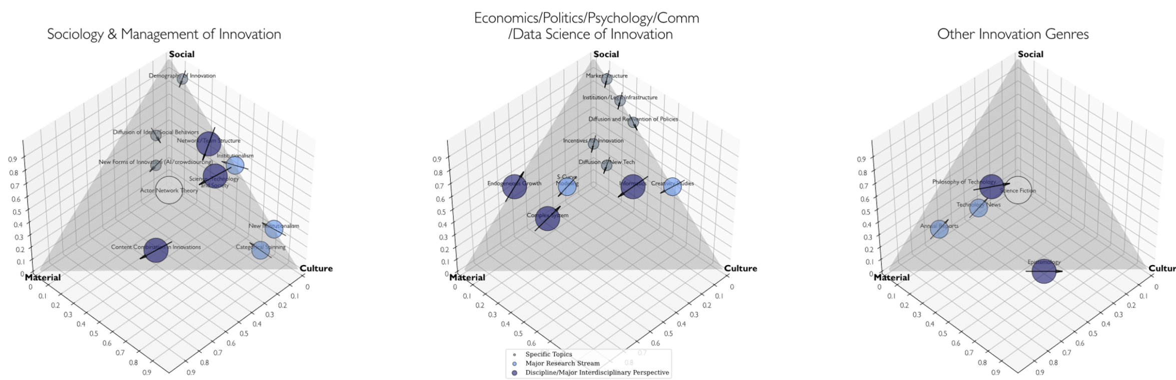

Figure 1.1: Explanation pathways across different genres.

Figure 1.1 above presents the distribution of theoretical assumptions across different genres of innovation literature. Two key insights emerge from this visualization. First, the social sciences and other genres often point in notably different—and at times opposing—directions in their theorizations. While this divergence highlights the potential of social science to offer alternative ways of understanding reality, it also reveals the limited practical influence of its theories. Second, a significant empirical gap remains at the heart of our understanding: to date, no comprehensive framework has effectively integrated the complex interplay of material, social, and cultural forces.

The co-existence of dual epistemological stances: from generalizability of social physics to particularity of social history.

The third fault line arises from an epistemological divide between a social-physical perspective—which emphasizes the parallels between human and natural systems (Erwin and Krakauer 2004b), and interpretivist social thought—which foregrounds the particularity of historical events. Despite their deep philosophical tensions, both perspectives have inspired rich and impactful empirical evidence.

The first stance is exemplified by a series of studies published in general science venues, often under the banner of the Science of Science. Here, each study typically advances a single, central claim, but supported by evidence drawn from diverse domains—such as scientific research, technological invention, cultural production, and open-source innovation on online platforms (Lin, Frey, and Wu 2023; Liu et al. 2021; Wu et al. 2019). Empirically, the central claims are supported across a wide range of domains; theoretically, and more ambitiously, they are extended to encompass all facets of social life. Thereby, this literature aspires to uncover universal laws of knowledge production—an endeavor regarded as untenable from a fundamentalist social science lens. A softer version of generalizable innovation science seeks to develop theoretical frameworks, instead of specific claims, with broad cross-domain applicability. Examples include the theory of recombination (Xiao, Makhija, and Karim 2022) and creative conflict (Stark 2011b). By operating at a higher level of theoretical abstraction, this standard becomes easier to achieve and is more readily accepted by context-sensitive social science audiences.

At the other end of the spectrum, a more classical stream of social science literature emphasizes the historical particularity of innovation. In contrast to the social-physical approach, this perspective assumes a priori incomparability—not only across different social domains, but also between individual cases within the same domain. Here, the theoretical assumption is the opposite of that in the first stream: uniqueness is presumed "unless otherwise stated". Padgett's work offers a compelling example (Padgett and Powell 2012). Padgett examines economic and institutional innovations in the Netherlands, Germany, the Soviet Union, and China separately, each situated within its distinct political and institutional traditions, making it highly unlikely to converge onto a common historical path. The collapse of an existing structure is likened to a diamond fracturing along its fault lines; the orientation and location of those lines differ from one diamond to another, and cannot be known in advance without close, case-specific analysis.

Both epistemological stances have inspired a substantial body of empirical research. Paradoxically, despite their distinct—if not opposing—philosophical foundations, studies from both perspectives often rely on similar concepts and theoretical tools, such as recombination, evolution, and creative destruction. To date, however, there are no clear empirical findings or theoretical frameworks that delineate which aspects of innovation dynamics are comparable and which are fundamentally context-specific. Answering this question requires a scale and scope of empirical data, along with advanced analytical tools, that are more common in the natural sciences than in social studies. However, with the rapid expansion of digitized social records, such a mission is becoming increasingly feasible over the years (Salganik 2017).

1.4 The Big Question and The Central Argument

This study is therefore guided by a central question: Is it possible to develop and apply a unified theoretical framework—and a common set of concepts—to all forms of innovation? And if so, how far can we reasonably go in that direction?

Two theoretical considerations motivate this inquiry. The first is a common concern in the social sciences: every theory operates within a defined scope and has a set of boundary conditions. As the level of abstraction increases, a theory can more readily encompass a wider range of phenomena; yet as theories become more specific, offering broadly applicable explanations becomes both more difficult and more meaningful. Rooted in Aristotle's epistemology, this tension is inherent to all social theory and is particularly relevant in the field of innovation studies. In this domain, most practical problems are highly specific, yet the dominant theoretical frameworks have thus far remained at a high level of abstraction.

The second consideration is more immediate and directly relevant to our current context. Innovations are not isolated events; they are deeply interconnected across both space and time. Horizontally, different forms of innovation often unfold simultaneously—scientific advances, for instance, may compete with cultural developments for limited resources or public attention. This competition plays out not only in the material world but also in more abstract domains, such as society's collective cognitive bandwidth. Vertically, innovations evolve over time. What is seen as groundbreaking in one era may become conventional—or even a barrier to further progress—in the next. Yet within the continuous flow of time, the tipping point between these stages is rarely clear or easily defined.

Therefore, theoretical and empirical approaches for innovation must account not only for diverse phenomena, but also for those that emerge, evolve, and decline across time and space—

often in complex, interconnected, and competitive ways. To address this, I propose a systematic perspective that complements existing frameworks in innovation science—an approach developed through the three empirical studies presented in this dissertation.

1.5 Three Empirical Studies as Cases

Each of the three studies explores the social mechanisms underlying the emergence of novelty within a distinct social domain. Grounded in different academic traditions, each study also engages with theoretical debates relevant to scholars who examine innovation from diverse disciplines.

The first chapter, *Modularity, Higher-Order Recombination, and New Venture Success*, engages with innovation research in organization science. It examines how modularity in complex systems takes shape in the context of new venture creation and how it influences startup performance. Theoretically, this chapter contributes to discussions on the architecture of recombination in complex systems. Herbert Simon (Simon 1991a) was among the first to conceptualize recombination architecture in the study of emergence. The idea gained broader attention when Henderson and Clark (Henderson and Clark 1990) introduced the concept of "architectural innovation" as distinct from "component innovation." Since then, numerous studies have explored how individual elements are integrated into larger systems. Building on this tradition, the chapter extends the theoretical conversation to the empirical setting of entrepreneurship, where technologies and their applications serve as the building blocks of successful business models.

The second paper, *Deep versus Broad Technology Search and the Timing of Innovation Impact*, follows the intellectual tradition of computational social science and examines the social consequences of different search strategies. In contrast to the first research, which is situated in a highly competitive and dynamic setting, the second work explores how innovation occurs in a cooperative social system—technology, where inventions are typically based on the prior work of pioneers. We base our new measures for *recombination depth and breadth* on the notion of information flow across the hierarchical "tree" of knowledge, a conceptual representation of human knowledge accumulation, modeled continuously by the Poincaré embedding algorithm on negatively curved hyperbolic space. The findings reveal that deep search (and, consequently, deep recombination) garners immediate recognition within local communities, while broad search (and wide recombination) fosters wider diffusion over time, leading to greater long-term impact and higher potential for disruption. This paper illustrates how computational tools can be used to advance theory—demonstrating that new methods can capture previously unmeasurable dimensions of core concepts and generate fresh theoretical insights.

The first two chapters are grounded in social science, driven by theoretical motivations and supported by computational methods to reveal subtle but important social mechanisms within specific contexts. The third chapter, *Subjective Perspectives within Learned Representations Predict Innovation*, by contrast, adopts the paradigm of general science and seeks to uncover generalizable social laws. Drawing on five large-scale datasets—covering scientific publications, technological inventions, business strategies, movie screenplays, and Wikipedia articles—we investigate which forms of team composition most effectively lead to high-quality and impactful innovations. We apply a consistent set of team composition measures across all five domains and estimate models using an identical specification. Through this analysis, we introduce a new dimension of diversity—diversity in subjective perceptions—and offer a novel contribution to social theories of creativity. Our empirical findings reveal both the comparable and incomparable effects of team composition across five domains.

CHAPTER 2 MODULARITY, HIGHER-ORDER INVENTION AND NEW VENTURE SUCCESS

2.1 Introduction

The emergence of complex systems, whether organisms in nature, technologies in the marketplace, or institutions in society including entrepreneurial ventures, rely on modularity to hedge against the risks of exploratory failure (Sharma et al. 2023; Simon 1991b). Complex organisms, technologies and social institutions do not coalesce all at once. Neither do they typically build up through the addition of disaggregated components surviving from the past. Instead, components coalesce into stable modules that themselves become objects of macro-recombination. Complex biological organisms are composed of organic molecules, but their evolutionary emergence came to involve natural selection on much larger scales. Cells, tissues, organs, and macro-properties like size and shape were selected through hierarchical objects of inheritance, regulatory genes. The invention of complex technologies similarly involves an evolutionary process of recombination. The invention of concave lenses to treat myopia and the printing press vastly entrenched both the supply and demand for glasses in the 15th Century. It took another 150 years (1608) before the combination of convex and concave lenses formed the telescope and solved problems unimagined—the discovery of distant ships, troops and celestial bodies. And the combination of a complete telescope with a gun for targeted siting did not occur for more than two centuries (1830s), long after long-range marksmanship had become a routine component of military action. Each new invention involves the recombination of increasingly complex components or modules into a new system. The resulting system—a scoped rifle—would never have emerged if all subcomponents needed to be assembled simultaneously.

Alongside organizational theorists (Levinthal 2021; Padgett and Powell 2012) and evolutionary economics (Nelson and Winter 1977), we argue that recombinant innovation occurs not only in technological invention, but also in organizational creation, combining components from technology, culture, society and economy. In this paper, we explore recombination and organizational emergence in the context of new business enterprises, which move from birth to maturity or death within the public eye. Like other emerging organizations, new ventures emerge as complex systems within the overarching innovation order of society. Innovation systems with societies encompass scientific discovery, technological invention, and industrial transformation (Freeman 1995; Lundvall 2010; Nelson and Rosenberg 1993). Entrepreneurial ventures draw together novel configurations of people, discoveries, technologies, problems and opportunities into processes that compete to open and satisfy markets for goods and services. We theorize new venture success and failure in terms of the strategic combination of business components (e.g., technology, product, market). Among the range of existing and possible combinations, we identify the locus of novelty most consistently associated with venture success.

Entrepreneurial firms assemble diverse technologies, and their associated solutions, to construct and stably fill new market niches (Stuart and Podolny 2007). Consider the phrase "this company is the 'Uber' of X", repeated on nearly 200,000 websites and across thousands of start-up descriptions and commentaries, where X represents everything from private jets and houses to mobile gyms, gasoline and garbage (Barns 2020; Cornelissen and Cholakova 2021). This phrase suggests a combination of mobile technology, dynamic cloud-based scheduling, and distributed assets that propelled Uber's success in order to decentralize distribution and create a "sharing economy" across many domains.

Here we build a theory of organizational emergence within entrepreneurship as higher-order invention, whereby successful new ventures construct novel combinations of component modules. Ventures that combine successful, applied technologies limit their risk of failure by assembling these modules into an immediately value-generating system, rather than risking failure in lower-order discovery and application. Like other complex, adaptive systems in nature, we posit that modularity is a critical feature in the evolutionary emergence of new firms robust to exploratory failure.

Figure 2.1 highlights the nested relationship between lower and higher-order invention within new business. In that diagram, venture (1) develops a single component (e.g., a novel technology out of a university lab) and firm (2) develops three related components that trace an existing module (e.g., it applies a technology to two related application areas), both acts of critical but risky lower-order invention. Start-up (3) combines three modules (e.g., successfully applied technologies) into a value-generating system, an act of higher-order invention. Venture (4) attempts to do it all; to generate new components and combine existing modules within a broader system, simultaneously undertaking lower and higher-order invention. We theorize that the action underlying firm (3) is most likely to lead to outsized new venture success, as moderated by the distance between the successful modules it combines (see Figure 2.2).

We evaluate these propositions with new ventures documented in Crunchbase (*N*=197,978) and VentureXpert (*N*=63,018) over 45 years (1976-2020) in relation to predictable outcomes of market success and failure. Specifically, we decompose detailed descriptions of U.S. venture-funded entrepreneurial start-ups over the last half-century to identify their underlying components and modular structure.

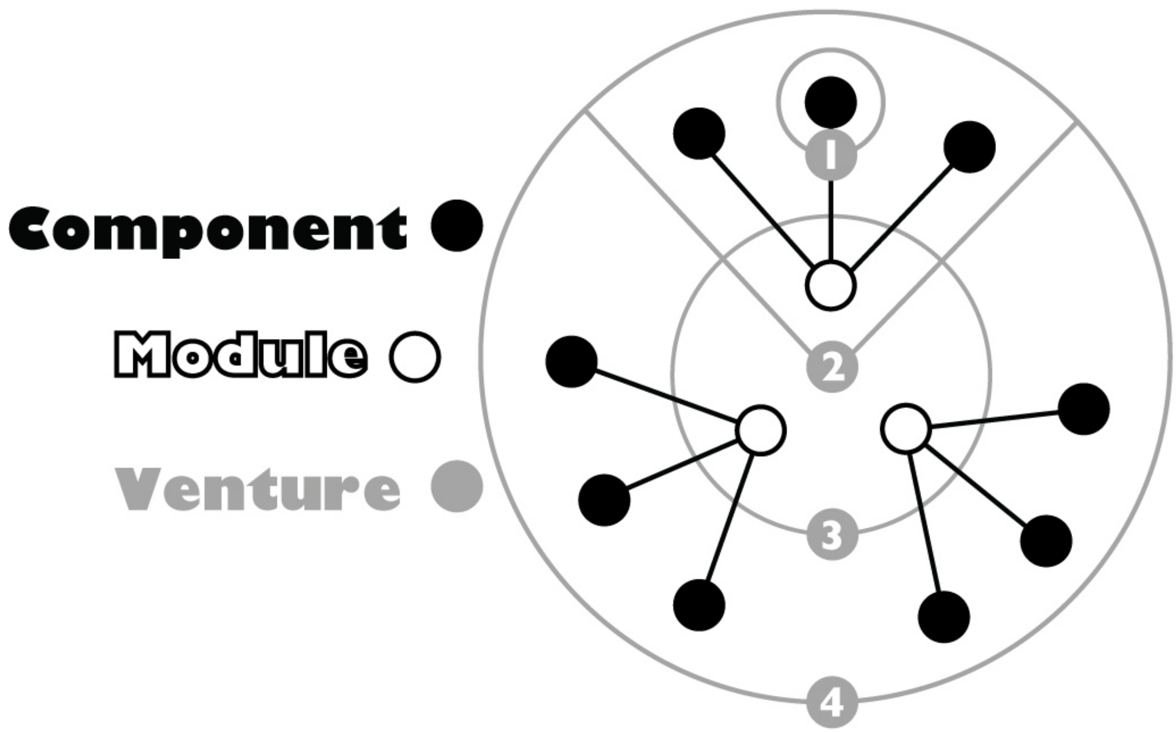

Figure 2.1: Schematic Diagram of the relationship between technologies, application components and their combination into new modules and ventures.

This theory requires a precise assessment of the evolution of common business understanding, which we assess with dynamic word embedding models built from a large corpus of business and technical documents. We then develop measurements to identify and distinguish lower- from higher-order entrepreneurial inventions. Next, we link these identified entrepreneurial innovations with detailed firm-level outcomes, ranging from closure to acquisition to a successful initial public offering (IPO) and establish the relationship between them. Finally, we link these with intermediate entrepreneurial processes to identify mechanisms through which modularity enables new venture success. We conclude by drawing out the implications of our for social theories of innovation, such as actor-network theory.

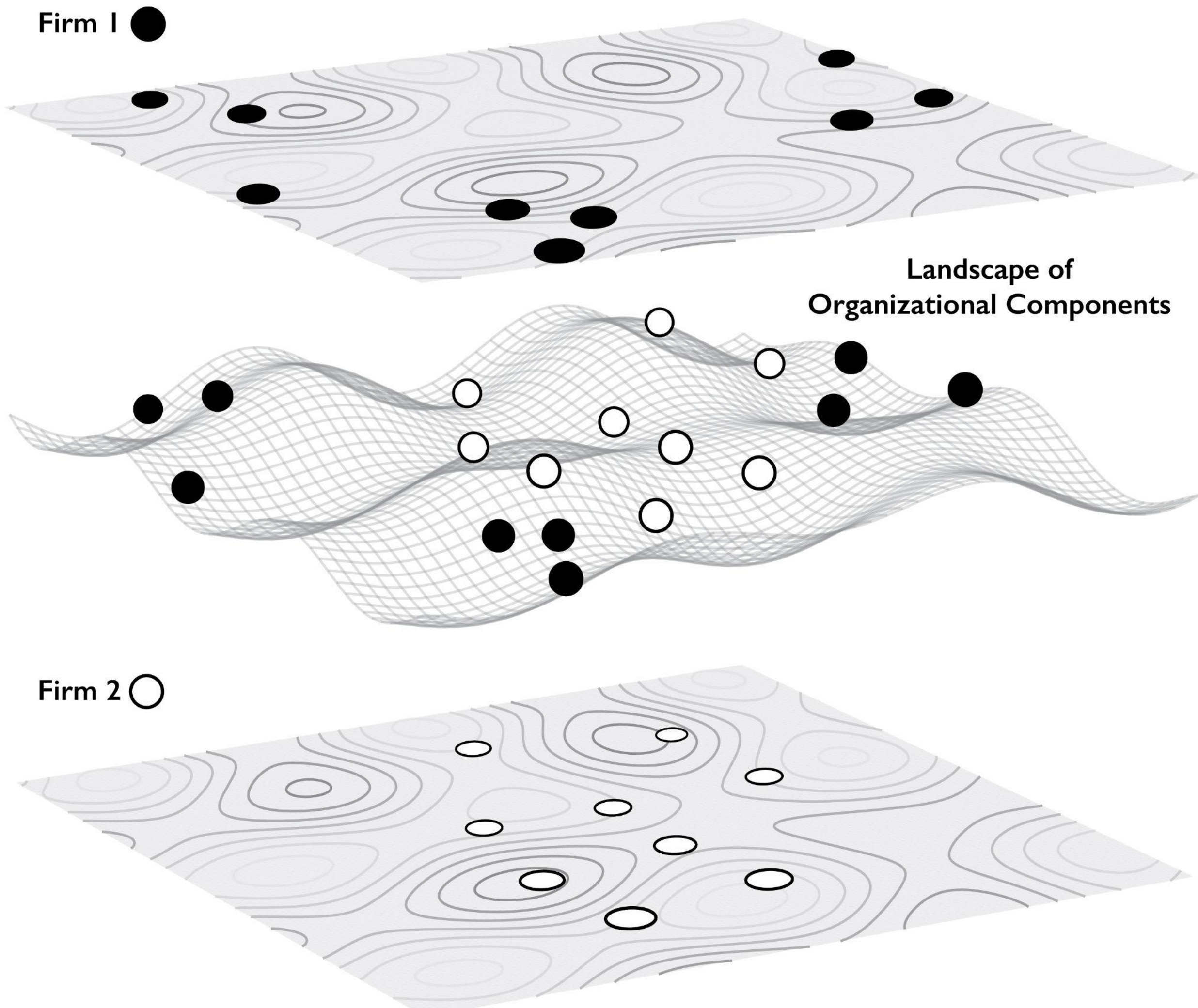

Figure 2.2: Research strategy for measuring the relationship between technologies, application domains and new ventures.

Figure 2.2 illustrates our research strategy for measuring the relationship between existing socio-economic modules and their higher-order combination into new ventures. The middle layer examines the landscape of business components (e.g., technologies, applications, markets) within a complex semantic space conditioned by business discourse, with the components of two hypothetical firms highlighted. The bottom layer projection contains the components of an incoherent firm our theory proposes as most likely to fail, containing neither modules nor diversity. By contrast, the top layer projection contains distant modules of components for the firm our theory proposes will be more likely to succeed.

2.2 Recombination in Socio-Economic Innovation

The notion of recombination has been highly influential in the study of biological evolution and physical structure with the scientific paradigm of complex, adaptive systems (Holland 1992). Conceiving of something as a combine-able component within a complex, adaptive system focuses attention away from its essential properties or characteristics, and toward their interaction and the emergence of complex collective outcomes that result. In chemistry, recombination occurs through the combination of elements and isotopes within larger molecules. In animal biology, recombination occurs through the shuffling of genetic material during sexual reproduction, resulting in offspring with combinations of genetic traits distinct from their parents to solve the problem of evolutionary adaptation to changing environments. The combinatorial notion underlying evolution through genetic recombination has attracted attention from physical scientists, inspiring new explanations for the emergence of material complexity in the universe (Sharma et al. 2023). The value of a combinatorial perspective has been demonstrated for intellectual and material inventions within human culture. Like offspring in the biological world or materials in the physical world, new ideas, tools, and economic capabilities are typically generated by combinatorially shuffling the components of prior ones. For this reason, the theoretical framework of recombination has become central to understanding innovation in the context of scientific discoveries (Shi, Foster, and Evans 2015; Uzzi et al. 2013) and technological inventions (Brian Arthur 2009; Fleming 2001; Fleming and Giudicati 2018; Fleming and Sorenson 2004).

Social scientists have broadened the application of the combinatorial perspective to understand the emergence of new social and cultural institutions (Levinthal 2021; Nelson and Winter 1977, 2002; Schumpeter 2006 [1911]) including teams (de Vaan, Vedres, and Stark 2015), social movements (Wang and Soule 2016), organizations (Stark 2011b; Weitzman 1996; Williamson and Winter 1993), and markets and states (Padgett and Powell 2012). For these complex social processes, combined components can be diverse, including professional expertise (de Vaan, Vedres, et al. 2015), social tactics (Wang and Soule 2016), network resources (Powell et al. 2005) and social-political traditions (Padgett and Powell 2012).

Studies of recombinant social innovation, however, have to date explored only a modest range of theoretical possibilities suggested by complexity theory. Most studies involving social recombination rely on a focused, well-defined space of components. Less attention has been paid to scenarios in which there are different types of components, or components are not clearly differentiated, as in a nascent business.[2] In such contexts, components can be of fundamentally different types (e.g., team, market, technology), which recombine with one another in distinctive ways (Pontikes 2018). Qualitative and conceptual studies of organizations suggest that cross-category combinations are more likely to yield innovation (Cohen, March, and Olsen 1972; Olsson and Frey 2002), but they remain unmeasured by most quantitative empirical approaches. As a result, current studies measure innovation in a variety of ways ([self-citation placeholder]), including more or less novel feature sets (Murdock, Allen, and DeDeo 2017; Pennington, Socher, and Manning 2014a) and more or less novel patterns of recombination (Fleming 2001; Fleming and Sorenson 2001), but the effects of cross-domain recombinations remain largely

[2] A notable exception is the work of Elizabeth Pontikes, which shows that software firms are more successful if they assemble surprising combinations of technology classes and markets, but not technology classes alone (Pontikes et al. 2018).

unexamined. Incorporating more types of components requires not only new tools for systematically assessing them, but also a broadened theory regarding what types of combinations matter (Cao, Chen, and Evans 2022).

Furthermore, the complexity perspective theorizes the importance of multiple levels of structure and interaction. Sociology and organization scholars of innovation have focused on singular fields of interacting components, like ideas traced in cited papers or technologies underlying patent categories. These approaches imply a “mean field” assumption that untried component combinations are equally likely to be attempted, and have an equal likelihood of influence on emergent outcomes that result. Complex systems theory, by contrast, has argued that in the evolution of increasingly complex biology or technology, the units of recombination are rarely individual lower-order components, but rather higher-level *modules* of components.

In Herbert Simon’s “Architecture of Complexity”, he articulates how modularity reduces the risk of catastrophic setbacks in biological and technological evolution. If an evolving architecture collapses completely everytime a new combination fails, then a successful, complex object would be vanishingly improbable. Instead, successful combinations of components consolidate into modules that themselves become objects of higher-order recombination and evolution. This approach was recently formulated as the basis of Assembly Theory, a new branch of complex systems analysis that highlights the critical importance of modular recombination in all forms of biological and technological invention, as measured by the copies of modules within a system, and the shortest path of their combination into higher-order structures (Sharma et al. 2023). In what follows, we build a theory of scientific, technological and business innovation based on these principles.

2.3 Organizational Emergence as Higher-Order Recombination

We explore the role of modularity in the evolution of successful social forms within the context of entrepreneurial emergence for two reasons. First, many have viewed markets (Arthur 1995; Battiston et al. 2016; Erwin and Krakauer 2004a; Farmer and Geanakoplos 2009; Kauffman 1996) and organizations (Padgett and Powell 2012) as complex systems, in part from how they aggregate seemingly simple exchange dynamics into highly complex and self-regulatory systems. More specifically, organizations have been recognized as a complex combination of knowledge (Hargadon and Sutton 1997; Kogut and Zander 1992; Zander and Kogut 1995), routines (Argote and Guo 2016; Becker and Zirpoli 2008; Pentland and Rueter 1994; Stańczyk-Hugiet 2014), experiences (McDonald and Eisenhardt 2020; Omezzine and Bodas Freitas 2022; Singh et al. 2016; Stark 2011b; de Vaan, Stark, and Vedres 2015) and resources (Eshima and Anderson 2017; Galunic, Charles Galunic, and Rodan 1998; Helfat and Eisenhardt 2004; Hess and Rothaermel 2011; Lippman and Rumelt 2003).

In the organization emergence context of entrepreneurship, better knowledge about relevant technologies, markets, customers, and other aspects of the new venture and its environment, all enable them to dynamically hedge against the risk of failure and identify profitable opportunities. These business components interact in nonlinear ways, such that a few combinations (e.g., a technology applied to a customer demand) will yield outsized success and most to unambiguous failure. Hence, entrepreneurship provides a strong case within which to empirically disentangle innovation across different types of components and explore whether modularity adds to combinatorial success, as it does in chemical discovery and biological evolution as characterized by the complex systems literature (Sharma et al. 2023).

Second, the existence of new venture databases allow us to explore questions of modularity with empirical data for statistical inference. This allows us to go beyond simulation analyses, a method historically relied on by complexity scholars[3] and for which complexity studies have been criticized by many domain scholars as insufficiently realistic. For example, mature entrepreneurial datasets such as Crunchbase and VentureXpert contain not only founder and investor information, but also detailed company descriptions. Company descriptions are central to processes of new venture fundraising, which are written both by the new ventures themselves and also by independent news that services markets for investment. The direct financial incentive for entrepreneurs guarantees the accuracy of the information. Moreover, the availability of well-tracked startup outcomes, such as initial public offering (IPO), high and low-priced acquisitions, or bankruptcy and closure, which have long been used to measure commercial success or failure, enable us to empirically evaluate the influence of modular recombination on organizational outcomes.

To study entrepreneurial emergence through the lens of modular innovation, it is essential to distinguish between lower- and higher-order combinations of business components. Lower-order combinations refer to the combination of components that have not been fully commercialized, such as technologies, novel market niches, and unrecognized expertise. When a lower-order combination is novel, we refer to it as a *lower-order invention.* In entrepreneurship, *lower-order inventions* are those new ventures whose businesses involve novel combinations of business components. This might include the invention of new technologies from diverse technical components, their novel application to solve old problems in radically new ways, or

[3]These include N-K models pioneered by Kaufmann, hypercycle models (Padgett and Powell 2012), cellular automata, system simulation (Ethiraj and Levinthal 2004), discursive theory (Anderson 1999), and agent-based models (Levinthal 1997). A notable exception is (Fleming and Sorenson 2001).

their construction of entirely new opportunities from diverse market niches. Lower-order inventions are mostly sponsored through public research funds and public institutions, and sometimes sponsored and incubated by industry, then protected through intellectual property and industrial secrecy (Cohen, Nelson, and Walsh 2000). The greatest risk associated with attempts at lower-order invention is failure. However, if *lower-order inventions* perform successfully, they could become reusable *modules* that may be later selected into novel higher-order combinations or *higher-order inventions*.

Higher-order combinations provide another recipe for entrepreneurial emergence. New ventures may seek to connect existing modules that have already demonstrated success, such as new successful technologies or existing technologies with proven application to business problems. Successful entrepreneurs assemble these working modules into complex systems that produce emergent values that did not exist within any one of the underlying modules alone. These include interconnected technological systems, like a factory, an energy plant, an airline, or a fleet of digitally interconnected cars and phone apps (e.g., Uber). These assembled systems stimulate further technological invention and business application, but may also lock-in technologies through their complex system-level interdependencies. We deem these *higher-order inventions* because they generate values that emerge from the novel interaction of underlying component modules. The risks of assemblage also include failure. Working technologies combined together may fail to complement one another. Invented systems may fail to sustainably produce the targeted good or service. They may even become sufficiently complex that they cannot be easily iterated upon and improved (Christensen et al. 2018). Higher-order inventions are always constructed by complex organizations, whether in government or business, because they require sustained investments of capital and coordination for invention and maintenance.

We note that higher and lower-order inventions do not reflect higher- and lower-value. Both are essential to the knowledge economy. Successful higher-order inventions rely upon the successful lower-order inventions they combine. Lower-order inventions are often funded as public goods through university or government research because of their perceived contributions to subsequent higher-order inventions that build upon them.

Lower- and higher-order invention each involve separate risks to the emergent organization, such that if a single venture seeks to accomplish them both—discover a new technology, apply it in a novel way to solve a human problem, and assemble it into a value-generating system that can be sustained through markets—the chance of success are vanishingly small. This is the crux of Simon's theory regarding the importance of modularity for the emergence of successful complex systems (Simon 1991b) and it has been demonstrated in natural (Sharma et al. 2023) and technical systems (Brian Arthur 2009). For the new venture, the risk of failure is the product of risks associated with more or less improbable lower and higher-order invention. In the same way that successful complexity emerges from robust modules in evolution, we argue that new ventures which assemble diverse modules into novel, synergistic systems are most likely to succeed. By investing their risk in the synergy of successful lower-order inventions, which represent robust modules within the industrial economy, new ventures hedge against failure and increase their speed and the ultimate likelihood of commercial success. These dynamics result in the following hypotheses, which differ from one another insofar as we separately measure lower and higher invention within entrepreneurial ventures:

Hypothesis 1a: Entrepreneurial firms that engage in lower-order invention by combining distant components within a module, will be less likely to (a) attract new venture capital, (b) obtain successful IPOs or high-priced acquisitions; and more likely to (c) close.

Hypothesis 1b: Entrepreneurial firms that engage in higher-order invention, combining distant modules, will be more likely to (a) attract new venture capital, (b) obtain successful IPOs and high-priced acquisitions; and less likely to (c) close.

2.4 Mechanisms underlying Higher-Order Inventive Success

A key logic underlying the success of higher-order invention is that the assemblage of previously successful business modules reduces the risk of success for new ventures. As described above, one risk new ventures face is that their combination will fail to generate value for consumers. A second involves whether the new venture will succeed in appropriating the benefits of their value-generating system. Higher-order invention hedges against these risks by increasing the speed of assemblage. Like cooking celebrity Julia Childs when she rapidly assembled a recipe on air with premixed ingredients, new ventures that assemble successful modules invented by others can prototype their system rapidly. This increases the number of iterations companies can make in evolving a value-generating system before early stage funds run out.

The accelerating effects of modularity were first articulated by Simon (1991b [1968]), and examined by later simulation and empirical analysis (Arthur and Polak 2006; Fink and Teimouri 2019). Simon illustrates this effect by contrasting two hypothetical watchmakers.[4] One always began construction of a new watch with its elemental components. Whenever this watchmaker lost concentration or made a mistake, she would need to begin again from scratch. The second watchmaker instead marked her progress by robustly encasing functional modules composed of successfully combined lower-level components. Whenever distraction or error

[4] Simon's argument was used to justify natural evolution and engineering possibility by reinterpreting William Paley's classic parable for the existence of God in his 1802 book Natural Theology.

resulted in the failure of this watch's construction, the watchmaker would restart by building from her most recent modules. The first watchmaker would take eons to finish her watch. The second, like an object-oriented programmer, would not lose progress when distracted. Assembly Theory builds on this insight to demonstrate how the availability of functional modules within chemical and biological systems are critical to the evolution of environmental complexity and life (Sharma et al. 2023).

Prior organizational studies demonstrate that mature skill complementarities (Müller 2010) and founders' previous experiences (Capelleras and Greene 2008; Kotha, Shin, and Fisher 2022) also accelerate the growth of emerging organizations. Assembling low-level business components in novel ways will slow new venture creation. For example, new technologies need time to learn and gain legitimacy in a new application area (Qin, Wright, and Gao 2017).

The strategic advantages from acceleration have been well established (Eisenhardt and Martin 2000). The most direct benefit involves improved competitiveness. When organizations act fast, rivals cannot overtake the firms' products or services, but only establish themselves through differentiation (Porter 1980). Lower-order invention like the creation of new technologies or crafting new applications requires time-consuming trial and error. Higher-order inventions involve the assemblage of functional lower-order inventions, building on the shoulders of prior success, and so enables rapid market experimentation. Management researchers have also hypothesized that faster innovation may relate directly to lower product cost and higher quality through the intensive interaction and organizational learning it catalyzes (Kessler and Chakrabarti 1996). Together, this leads us to the following hypothesis.

Hypothesis 2: Conditional on receipt of initial funding, entrepreneurial firms that engage in higher-order Invention—combining distant modules—will move more quickly from seed round *funding to early stage venture funding; and this speed will account for significant variation in their likelihood of achieving successful IPO, high-priced acquisition, or closure.*

Another factor underlying the success of higher-order invention is that the assemblage of diverse modules may enable a system (e.g., new venture) to draw together and flourish within multiple environmental niches simultaneously. In the context of business innovation, this will broaden the appeal and increase the perceived novelty for the emerging organization's audiences. Specifically, this will widen the pool of potential customers and diverse potential investors with experience in different parts of the economy. We test this hypothesis with data on investor backgrounds.

New ventures can obtain investment from many sources including angel investors, venture capitalists (VCs), and corporate venture capitalists (CVCs). These financial relationships involve much more than the simple transfer of funds. In the U.S., investors often work closely with the founding team, providing guidance, and influencing future business decisions (Drover et al. 2017). Studies demonstrate that investor reputations (Nahata 2008), social networks (Ter Wal et al. 2016) and institutional logics (Pahnke, Katila, and Eisenhardt 2015) influence their portfolio companies' performance and innovation. The scope and quality of the investor pool, therefore, represents an important factor in new venture success.

When a startup is engaged in higher-order invention that crosses multiple fields, it will likely attract more investors as it provides multidimensional value. A wider pool of investors, in turn, diversifies the expertise that new ventures can draw upon in appealing to more diverse customers and markets. Moreover, a diversified pool of investors also increases a new venture's robustness to market risk, and the likelihood investors will be available to support the new venture through multiple stages of its progress toward profitability and independence. This leads to the following hypothesis:

Hypothesis 3: Entrepreneurial firms that engage in higher-order invention by assembling lower-order modules will attract funding from venture capital firms with diverse expertise and backgrounds; which will account for significant variation in their likelihood of achieving successful IPO, high-priced acquisition, or closure.

We explore these hypotheses about organizational emergence in the context of 20th and 21st Century U.S. entrepreneurship, drawing components and modules from business descriptions positioned within the landscape of market discourse.

2.5 Methods

Research Design.

Emerging businesses describe themselves and their strategies on websites and online databases for the purposes of raising funds and marketing products. In order to track market opportunity and value, the same companies are actively followed and described by business analysts and the business press. These descriptions provide measurable data about the strategy of entrepreneurial firms. In this way, firm components like technologies or product areas that are spoken and written about together, such as the use of mobile cloud scheduling for personal transportation (e.g., Uber), reflect the widespread application of mobile technology.

Our hypotheses relate to new venture success in Initial Private Offerings, high- and low-priced acquisitions, further rounds of venture funding, and closure as a function of whether they combine diverse modules of business components. We first identify new venture components as the nontrivial words in their descriptions, then partition them based on their proximity within an embedding space constructed from a broad sample of market discourse. Being in the same embedding region suggests that the business and inventing public have frequently discussed the components together before, and view them as a functional module (Vicinanza, Goldberg, and Srivastava 2023). Then we calculate the distance between modules, with higher distance suggesting that modules have rarely if ever been used (and described) together. We perform our analyses on multiple data sets and with multiple, conceptually similar but technically distinct measures in order to validate their robust confirmation of our conclusions.

Dynamic Word Embeddings to Estimate the Business Landscape.

Many innovation studies base their measurements of novel emergence on new combinations of classes from existing classification systems, such as patent or paper sub-classes (Fleming 2001; Taylor and Greve 2006). These methods rely on the assumption that the classification and citation systems are accurate and stable across domains and time. This assumption naturally breaks down in the dynamic market scenario, where new elements emerge every year and industrial patterns can radically transform across a single decade. In order to compensate for this dynamism, recent studies take advantage of contemporary natural language processing (NLP) techniques to extract information from technical and business corpora and generate more refined, dynamic measurements (Kaplan and Vakili 2015; Shiller and Brown 2019). As the economic and technological environment plays a fundamental role in start-up innovation and decision making (Marx, Gans, and Hsu 2014), which hedge their bets, here we first construct a landscape of business concepts and their associations, which provides the context for entrepreneurial opportunities and innovations.

We represent the shifting landscape of business associations with word embeddings. Word embeddings produce dense, continuous models that represent words as high-dimensional vectors to enable the measurement of precise semantic distances between them (Joulin et al. 2016; Mikolov, Chen, Corrado, and Dean 2013; Pennington et al. 2014a).[5] Distance between word vectors in embedding space serves as a proxy for semantic relationships underlying the referenced concepts. Substantial research has demonstrated the ability of embedding models to replicate even subtle semantic relationships, including analogies and metaphor (Mikolov, Chen, Corrado, Dean, et al. 2013), which correspond to subjective understandings held by human agents (Caliskan, Bryson, and Narayanan 2017; Garg et al. 2017; Kozlowski, Taddy, and Evans 2019), but also vertical shifts in society and culture (Garg et al. 2017; Kozlowski et al. 2019). As such, embeddings represent a powerful instrument from which to measure changing semantic associations over time. By presenting robust, overarching semantic patterns in a singular space, word embeddings have been widely adopted for analyzing complex social and cultural data (Arseniev-Koehler and Foster 2020; Evans and Aceves 2016; Nelson 2021).

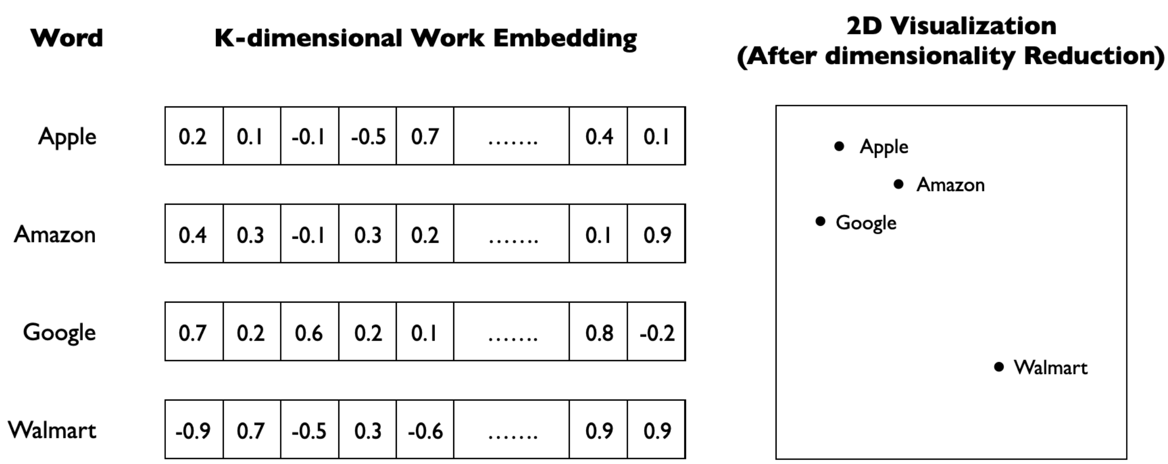

Figure 2.3: Conceptual Figure representing Word Vectors in an Embedding Space.

5 Word embedding models are sometimes considered and referred to as "low dimensional" relative to the number of words used in text (e.g., 50,000) because they reduce this *very* high dimensional word space. Nevertheless, considered from the perspective of one, two or three dimensional models common in the analysis of innovation, these spaces are much more complex, and reproduce much more accurate innovation associations, as shown below.

Word embedding algorithms have increasingly been used to analyze differences between cultures across contexts or over time. When multi-slice embedding spaces are derived from a temporal corpus, they enable analysis of evolving meanings. Approaches to embedding alignment and comparison have become more sophisticated in recent years to create a broader family of dynamic word embedding methods (Hamilton, Leskovec, and Jurafsky 2016; Liu, Zheng, and Zheng 2020; Yao et al. 2018a; Zhang et al. 2016). Outputs from these algorithms are time-stamped word vectors that contain the semantic information for that specific period and are comparable across history.

Dynamic word embedding algorithms follow one of two strategies. Either they train period-specific vectors independently then align them, or they learn embeddings across time jointly. We adopt the joint training paradigm in order to naturally account for changes in the composition of words used over historical time (See Appendix; and Yao et al. 2018a). This approach assumes and enforces a smooth transition between word vectors and their associated meanings in adjacent time periods and trains all periods jointly. Specifically, it alters the objective for the joint word embedding to the following:

$$\min_{U(1),\ldots,U(T)} \frac{1}{2}\sum_{t=1}^{T}\|Y(t) = U(t)U(t)^{T}\|_F^2 + \frac{\lambda}{2}\sum_{t=1}^{T}\|U(t)\|_F^2 + \frac{\tau}{2}\sum_{t=2}^{T}\|U(t-1) = U(t)\|_F^2 \tag{1}$$

The first term in this objective, the L_2 norm of the difference between the empirical word proximities *Y*(*t*) and the modeled proximities *U*(*t*), represents the standard word embedding objective. This objective guides optimization of the algorithm to accurately model word distances within time periods, and the L_2 norm regularizer is used, as in ridge regression, to improve out-of-sample predictive performance of the model. Dynamic word embeddings add two additional terms that: 1) allow the algorithm to minimize unnecessary word vector loadings (the L_2 norm of $U(t)$), and 2) shrink the distance between time-adjacent embeddings (the L_2 norm of the difference between $U(t\text{-}1)$ and $U(t)$).[6] In minimizing this cost function, adjacent word embedding spaces are grown in alignment, with a $t \times n \times k$ matrix, where t is the number of time slices, n is the number of words, and k is the dimensions of word vectors. The business landscape of year t is stored in one $n \times k$ slice, which is optimized to remain similar to the t-1 and t+1matrices. In this way, the jointly trained algorithm retains words that newly appear in the corpus over time and keeps adjacent semantic spaces close in structure. This allows us to construct comparable and historically accurate representations of the structure of concepts within business discourse for each year over the nearly half century we examine.

We use two datasets as input to the dynamic word embedding model of business environments: ProQuest business news, and the USPTO full text patent corpus, 1976-2020. ProQuest business news includes 108 English newspaper and magazine full-text articles, published in the U.S. and falling into the category of "Business and Economics". This contains major business publications, including the *Wall Street Journal*, *Bloomberg BusinessWeek*, *the Economist*, and *American Banker*, among many others. It provides temporal information about relationships between business concepts, including both technologies and their application areas, based on how reportage discussed them in the context of contemporary and imagined future business. In total, our dataset includes over 6 million business news records (1976-2020; see Appendix for additional corpus details). Recent research in finance has demonstrated that word embeddings based on newspapers have accurately captured economic features in the macro

[6] This dynamic word embedding approach relies on the formal equivalence between the standard negative sampling approach to word embedding optimization used by `word2vec` and the low-rank factorization of a pointwise mutual information (PMI) matrix of word vectors (Levy and Goldberg, n.d., 2014), and uses the PMI as input to the objective above.

environment, and can be used to predict economic fundamentals such as GDP, consumption and employment growth, even after controlling for commonly used predictors (van Binsbergen et al. n.d.). Word embedding method has also been widely applied in studies about business niche and market differentiation (Carlson 2023).

Our analysis relies on the knowledge-base underlying new ventures in context not only with the business challenges they seek to address, but the technologies through which they seek to address them. For this reason, we must account not only for knowledge from the lay, business reading public, but expert knowledge about the nature of new venture technologies and their applications. We supplement business news coverage with descriptions from patents owned by businesses within our sample of new ventures. Specifically, we utilize full text descriptions from the USPTO patent dataset (1976-2020) and adjust the weights between news and patents to keep application and technology elements balanced during the training process. This allowed us to represent both business discourse and the technological knowledge basis underlying U.S. venture creation across this period.

The output of our dynamic word embedding is a 45-slice word embedding space, modeling our joint business-technology corpus from 1976-2020. Each slice represents the geometric distribution of business concepts (i.e., words) from a given year, comparable across years.

The Business Landscape.

Our word embedding model generates a 50 dimensional vector space for 227,714 words across 45 years. When the social economic functions of a word change, the element drifts across the landscape. In figure 2.4, we use "Amazon" as an example to show this dynamic. From 1993 to 2017, Amazon transformed, at first dramatically from the Brazilian rainforest surrounded by concepts "tropical", "forest" and "farmland", to an internet retailer in 2001 nearest "bookseller", “etoys” and “priceline”, other online retailers of toys and tickets. Thereafter Amazon continued to evolve. In 2006, the year after its introduction of Prime video and music, it was closer to “youtube”, “video sharing”, and the social media platform “myspace”. With the roll-out of its Fire Phone in 2014, following the success of its Kindle Fire reading and internet devices, Amazon moved closer to “whatsapp”, “smartphone”, “samsung" and “sony”. The emergence of Amazon as the hegemonic retailer it is today, covering all product categories with advertising revenue reaching the billions in 2017, positioned the company near “online”, “advertising”, “ecommerce”, and “facebook”. Other cases of the evolution of company positioning for Apple, Tesla, and Theranos are provided in the Appendix. For example, with the dramatic fall of Theranos in 2016, words surrounding the company before and after dramatically diverged in sentiment. See Appendix for additional validations of our evolving business landscape.

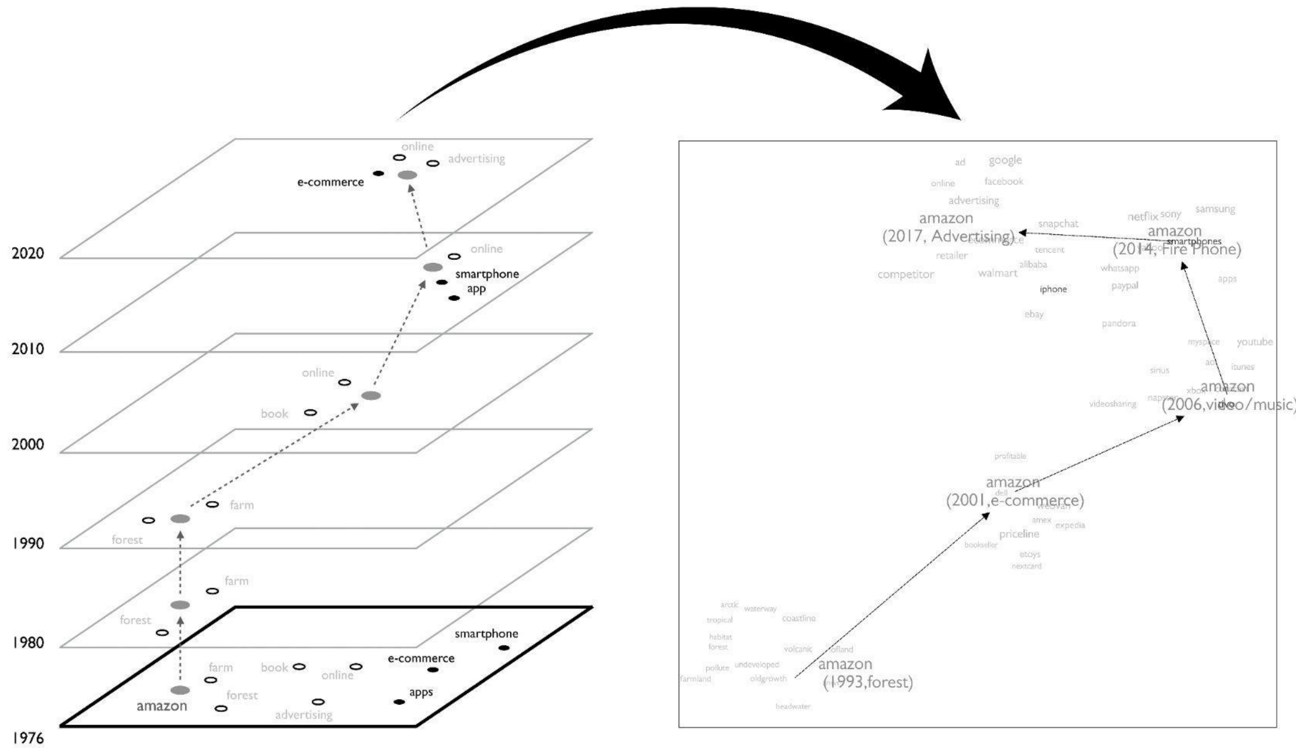

Figure 2.4: Dynamic business landscape. Left, a conceptual representation of the layered construction of our semantic space representing business discourse. Right, an empirical figure, orthogonal to the one on the left, which flattens the overlapping layers, highlighting the movement of Amazon through conceptual space over time and its neighborhoods of most similar concepts in 1993, 2001, 2006, 2014, and 2017.

Mapping New Venture Descriptions within the Evolving Business Space.

In this study, we use two databases to independently assess the position of new venture strategies within the evolving semantic space of business described above. First, we use

CrunchBase (CB), a commercial database serving entrepreneurs and investors (N=298,915 by 2020). Second, for independent validation, we use VentureXpert (VX) from Thomson Reuters (N=63,492 by 2018). Both datasets contain information regarding U.S. companies established 1950 to present. Both datasets provide a long description for each company, with detailed information about its major markets, relevant technologies, and strategies. In Crunchbase, these descriptions are drafted by representatives of the company itself and provided to the database. As companies often use these online platforms to establish their public images, attract potential investors and identify partners, we assume that companies seek to highlight the most relevant and distinguishing information regarding their operations and markets, and have sought to keep them accurate as they themselves evolve. We note that Crunchbase description data has been widely used in strategy-related studies (Chae and Olson 2021; Savin, Chukavina, and Pushkarev 2022). In VentureXpert, another prominent database targeted at investor clients, analysts and Thomson Reuters independently construct descriptions of comparable length and format.

We project company descriptions into the evolving semantic space reflecting business realities by calculating the centroid (or high dimensional average) of all words in the description. We feature results from CrunchBase in the body of the paper, but repeat all analyses with VentureXpert as reported in the Appendix.

We note that new ventures may change over time. Nevertheless, once companies mature such that they enter Crunchbase or VentureXpert, their descriptions rarely change. In order to assess the robustness of our findings to strategic pivoting among new ventures (Kiss and Barr 2015), we collected historical records of the CrunchBase dataset from the Wayback Machine, a non-profit internet archive that preserves historical records of persistent websites. This record allows us to trace changes in company descriptions on a monthly basis. We also carry out validation analysis based on this longitudinal data. All results reported below remain robust and do not change with time-shifting descriptions (see Appendix).

Measures of Lower and Higher-Order Invention.

We then build two measures that characterize new venture creation as lower vs. higher-order invention required to test our hypotheses. These metrics build upon the modular structure of business components within the semantic space constructed from business discourse. Specifically, we need to establish the degree to which new ventures combine (1) lower-level components in novel ways, and (2) higher-level modules in novel ways. We operationalize this by identifying regions within the embedding space, which allow us to measure how tightly business components (e.g., technologies, applications) lie with respect to each other in embedding regions. Business components that cleave together within the space of business discourse represent widely discussed and appreciated modules, available for recombination. We can then measure the distance between modules across the embedding space. Distant modules are those that have not yet become recognized as relevant to one another within business discourse, whose combination represents a surprise to the system. We use a high-performance topic identification method that draws directly upon our word embeddings, the discourse atom algorithm (Arora et al. n.d.; Arseniev-Koehler et al. 2022), to partition the semantic space into coherent semantic regions.

Discourse Atoms Estimate Embedding Regions and Modularity.

Word embeddings geometrically present the semantic relationship between words, but do not naturally capture how meanings and concepts cluster together to form a landscape of uneven density. To take advantage of high-dimensional word meanings and better describe this semantic distribution, Arora et al. (2018) developed discourse atom method, which uses word embedding vectors as inputs and discovers modular structures within semantic space. Based on the embedding space, it uses Singular Value Decomposition to partition words into *k* "discourse atoms" that tile the semantic space into coherent regions. Words close to the same atoms manifest highly correlated meanings, and we assign each word to its closest atom. In our analysis, we identify 200 atoms to balance human interpretation and information loss (Choi, Menon, and Tabakovic 2021; Kaplan and Vakili 2015), but the Appendix shows that our findings are insensitive to this number. Moreover, our results remain robust (see Appendix) when we replace discourse atom approach with k-means clustering built with embedding space loadings as features (Anon 1966).

Specifically, for each period the discourse atom algorithm on the word vectors has cut the space into 200 subspaces based on high within-atom consistency and inter-atom difference. In table 2.1, we use 1990 and 2010 data as an example to show the performance of discourse atom and dictionary-based word filtering.

Table 2.1: Examples for discourse atoms (1990 and 2010).

Example Discourse Atoms	Techno-Social Elements in the Atom
1990 Examples	
Atom 1	gasification, petroleum, hydropower, transportation, manufacturing, construction
Atom 2	cabletv, telegraph, widescreen, polaroid, publishing, broadcasting, roadshow
Atom 3	androgenic, cerebrum, psychographics, dating, boardroom, pawnbroker
Atom 4	archaeology, plato, folklore, mythology, spiritualism, melodrama, filmmaker
Atom 5	thermocouple, disk, pcb, tomography, apparatus, probe, channel, connection
Atom 6	photocurrents, entropy, spectroradiometer, measuring, meter, transition, sense
Atom 7	gemstone, linen, onyx, maraschino, lace, towel, womenswear, pottery, style
Atom 8	biofuel, petrol, uranium, biogas, refine, mineral, papermaking, automaking
Atom 9	cephalosporin, arthroscopy, aspirin, surgery, anesthesia, prevention
Atom 10	hyperboloid, ellipsoidal, optic, vertices, eyelens, intermetal, viewfinder, zigzag

2010 Examples	
Atom 1	pantone, herringbone, textile, fashion, designer, shoe, hat, wear
Atom 2	pycnometer, biofluid, microbiocide, circulation, recycling, fuel, injection
Atom 3	smartphone, cryptocurrency, webcam, chat, video, networking, advertising
Atom 4	ultrasonication,deaggregation, mineralogy, cleanness, crystallizer, fiberizing
Atom 5	overmolded, laminate, caulking, casting, glass, plastic, pressboard
Atom 6	sibutramine, isocarboxazid, linezolid, therapy, treatment, vaccine, preventative
Atom 7	riskmetrics, telebank, microfin, megadeals, coinvest, autofinance
Atom 8	circuitry, multiplexer, synchronous, converter, generator, signal, terminal
Atom 9	lans, uplink, vocoding, websockets, networks, brokering, communications
Atom 10	histologic, vascularity, tracheotomy, neogenesis, surgery, anesthesia

The word embedding space and discourse atom topics provide a description of the macro-cultural environment in which new ventures learn knowledge, enter the market, and make strategic decisions (van Binsbergen et al. n.d.). The recombination of technologies, applications and other socio-economic components across embedding regions allow us to construct two measures to robustly distinguish lower- from higher-order combinations:

Distance Between New Venture Components within Discourse Regions: We measure the distance between company components within a discourse region by assessing the average cosine distance between them. A large distance indicates that the company is engaged in lower-order combination within the semantic regions in which it is engaged. A small distance, by contrast, suggests that company components have been frequently used together in the past, reflecting their recognized modularity within the marketplace. Lower local distance liberates a new venture to engage in higher-order invention, combining modular, applied technologies in potentially novel ways. We predict that a smaller distance between company components within a region will anticipate IPO or high-priced acquisition in accordance with Hypothesis 1.

Global Distance Between Discourse Regions: We measure the average cosine distance between a company's modular components in order to establish the surprise associated with diverse higher-order combinations. High average distance between a company's modules or component regions in semantic space indicates that the combination has not yet been actively discussed, developed or deployed together in industry. Low average distance suggests that a firm has brought together modules frequently present in the marketplace and already considered together in commercial discourse. An expected combination is unlikely to signal a novel contribution that could forge or find a profitable, unfilled niche within the business ecosystem. We predict that larger distances between company modules will anticipate IPO or high-priced acquisition in accordance with Hypothesis 2.

The measurements of local and global distances allow us to determine the levels of invention in a new venture. If a venture shows greater local distance in its business components, it indicates a focus on lower level invention. In contrast, lower local distance paired with a higher global distance suggests that the venture is involved in a more modularized, higher level of invention.

We also conduct a version of analysis where we further distinguish components into two categories: technology and application. We then operationalize lower-order invention as the novel combination of technology and application elements, and higher-order invention as recombination of mature technology-application modules. For that purpose, we employed computational tools to identify technology and application components first. The results are consistent with our main findings, and are reported in Appendix C.

Modeling Business Outcomes with Higher-Order Invention.

Event History Analysis to Estimate New Venture Success. CrunchBase and VentureXpert both record major events in company lifecycles, including foundation, obtaining new funds from venture capital firms, acquisitions, initial public offering (IPO), and closure. Each of these events can be taken as milestones that reveal the competitive strength or weakness of an individual company, and the success or failure of its investors (Ljungqvist and Hochberg 2009).

For our event history analysis, we adopt the multi-outcome survival model to explore how innovation shapes new venture futures. As many outcomes are mutually exclusive—a venture cannot go IPO and bankrupt at the same time—here we call for models that deal with competing risks. The cause-specific Cox model is a possible solution (Blossfeld, Golsch, and Rohwer 2007), but it assumes independence of competing events and goes against assumptions that call for competing risk models. The Fine-Gray model, by contrast, is commonly-used in health and medical research, but not designed with an optimization algorithm suitable for large-scale data. For this reason, here we apply the cause-specific flexible parametric survival model (FPM) developed by (Mozumder, Rutherford, and Lambert 2017). This model identifies the factors that lead to a specific outcome, taking all other possible outcomes into consideration. We use the cause-specific Cox model as a robustness check. The results remain robust and do not change our main conclusion (see Appendix).

Mediation Analysis with Structure Equation Models to Explore Mechanisms. Our final analysis explores mechanisms underlying our finding that higher-order invention is associated with new venture success. For this purpose, we construct a mediation analysis. Here we take each new venture as a single observation. Final outcomes (IPO and high-priced acquisition, funding, low-priced acquisition, and closure) are set as dependent variables, with intermediate events (such as new funding before IPO) ignored. The two types of invention strategies serve as major predictors. Given the categorical nature of the dependent variable, we use generalized structural equation modeling with multinomial outcomes. For the very rare case (N=175) in which two types of final events occur for the same startup, such as low-priced acquisition followed by closure, we code it with the more successful outcome (e.g., low-priced acquisition).

Not every company in Crunchbase has received VC funding. Some received VC investments, but information about transactions is not totally disclosed. As our two mediators are constructed based on information about seed and early rounds (A&B rounds) of funding, in our mediation analysis we only include startups with clear information about these investments. Thus, the sample size for our mediation analysis is markedly smaller than the event history analysis.

We first run regressions on all samples with seed round investment (N=49,304), then run a robustness test with samples including both seed and early round investments (N=19,979). The first method assumes that right-censored data indicates the inability to receive a subsequent round of funding due to low commercialization speed. The second deletes all right-censored samples to rule out other possible explanations. Results show very similar patterns and support hypotheses 3 and 4. Then, we use structural equation modeling to test hypotheses 3 and 4. The results show very robust patterns and support our hypotheses.

Figure 2.5 reveals our analysis workflow, computational tools and datasets for each step. We perform the dynamic word embedding and discourse atom modeling in Python 3, then we transfer these variables for our event history and mediation analysis in Stata 15.

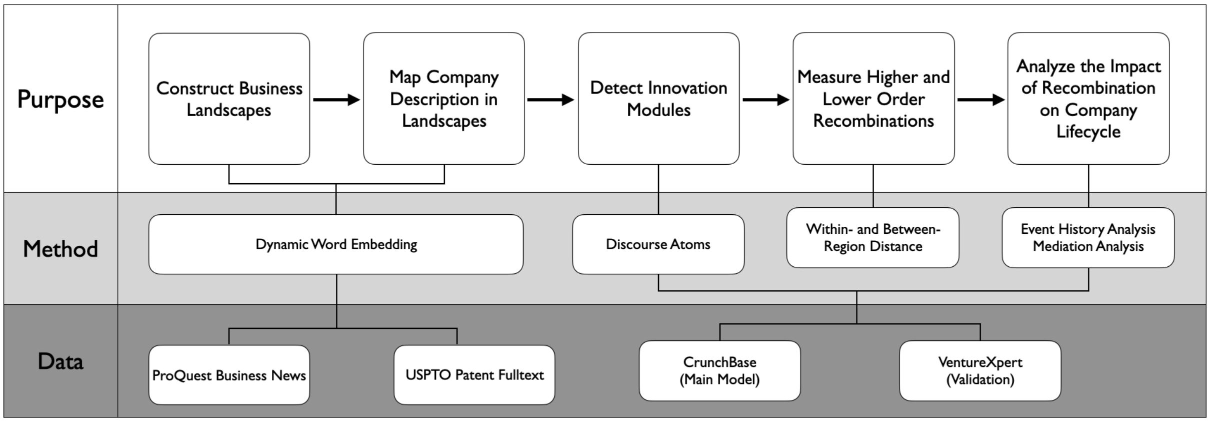

Figure 2.5: Data analysis process.

Variables.

Dependent Variables on Performance. Studies on venture capital finance often measure the success of venture capital firms with successful exits from portfolio companies, defined by number of IPO or M&A events within a given period (Humphery-Jenner and Suchard 2013; Ljungqvist and Hochberg 2009). Although the founding team does not necessarily exit with their investors, following much existing research, we treat IPO and high price acquisitions as proxies for successful growth and development (Chang 2004; Keyhani et al. 2022). Crunchbase contains records for both types of events. We identified M&A transactions with the top 30% price in each industry as high-priced acquisitions, with dollar values deflated by the Consumer Price Index (CPI). New ventures reporting an acquisition, but with missing prices, are assigned to the middle-low priced acquisition classification; a strategy guaranteeing high precision for successful outcomes (i.e., all acquisitions classified as high-priced are verified) at the cost of lower recall. In addition to IPO and M&A, we also take getting new fundings from venture capital firms as a positive sign of venture development (Roche, Conti, and Rothaermel 2020) and bankruptcy as a sign of failure. We treat these four types of events (IPO/high price acquisition, achieving new funding, middle-low price acquisition and closure) as competing outcomes for a venture in any given period. In the appendix, we also re-estimate our models on two alternative codings (to combine all acquisitions into the same category, or separate high-price acquisition and IPO into two independent categories) and the results remain robust.

Following the paradigm of multiepisode event history models (Blossfeld et al. 2007), a company can have several observations (for different developing stages) in the data. However, IPO and closing will be taken as final exits, and any other events after a final exit (like revival after closing) will be ignored.

Independent Variables. The focal independent variables that relate to lower and higher-order invention include (1) local distance between business components within regions of discourse; and (2) global distance between discourse regions. Higher local distance suggests the new venture is engaged in lower-order invention. Lower local distance combined with higher global distance suggests the new venture is engaged in higher-order invention.

Mediators. Our two theoretically motivated mediators include (1) the growth speed of new ventures and (2) the breadth of investment expertise and professional background among early round investors. We measure the growth speed of development with the time difference between seed and early round (A or B round) investment. In the seed funding stage, new ventures typically have begun to pursue their ideas, but remain unproven within the market. By contrast, early round funding occurs when product-market fit has been established and the team is preparing to scale. Thus, the time difference between the two rounds is a proxy for the rapid growth of new ventures. For the breadth of expertise and background, we trace past investment records for each VC firm when a new deal occurs. These include a terminology of Crunchbase industry keywords. The difference between VC backgrounds is calculated as pairwise Jaccard diversity (i.e., 1-Jaccard similarity) for industrial expertise keywords, and the diversity of VC background for one transaction is the average of all pairwise Jaccard diversity for participating VC firms.

Controls. Control variables include (1) text features associated with the business description, including text length (numerical), the existence of very rare words (dummy), and lack of technical words (dummy); (2) features of the company, including funding history (number of historically accepted funding rounds before the observation period), founders' identity (whether the founder is a woman or minority), and the company's media usage (whether the company uses Facebook and/or Twitter platforms); (3) the market environment in the year the venture was founded, measured by the growth rate of companies in the same industry; (4) temporal-spatial features, including year of observation, and geographical region of company headquarters; and (5) familiarity of technology components as a function of frequent, recent usage (Fleming 2001). We incorporate this last technology familiarity variable to capture the recent frequency of their usage in the entrepreneurship community alone.

2.6 Results

Descriptive Statistics.

Descriptive statistics of major IVs and DVs are reported in Table 2.2. The two databases, Crunchbase and VentureXpert, generate similar distributions for most variables. The only difference is that VentureXpert, on average, has a higher number for fund-raising and IPOs. This could be because Crunchbase, which mainly relies on volunteer reporting and automatic information-scraping, has more missing values. Statistics in Table 2.2 guarantee that the two databases are comparable.

Table 2.2: Descriptive statistics of variables.

	Crunchbase (n=197978)		VentureXpert (n=63018)	
	mean	std	mean	std
Dependent Variables				
Average number of accepted funding to date	0.758	1.471	2.455	2.626
Proportion of IPO or high price acquisition	0.038	0.202	0.099	0.300
Proportion of middle or low-price acquisition	0.173	0.378	0.172	0.377
Proportion of closing	0.029	0.168	0.130	0.336
Independent Variables				
Local Distance (lower-order invention)	0.140	0.106	0.128	0.092
Global Distance (higher-order invention)	0.449	0.058	0.450	0.065
Controls				
Element Familiarity	2.399	1.426	1.067	0.680
Dummy: Usage of social media	0.755	0.430	Not Applicable	
Dummy: founded by women or minorities	0.123	0.346	Not Applicable	
Market growth rate in the year of establishment	-0.068	0.228	Not Applicable	
Number of valid elements	30.702	20.564	42.362	22.468
Dummy: no technical elements at all	0.189	0.392	0.164	0.370
Uncommon elements	0.021	0.143	0.011	0.103

A correlation table of the variables are reported in Table 2.3.

Table 2.3: Correlation table for the major variables.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1) Local distance (lower-order invention)	1.000								
(2) Global distance (higher-order invention)	-0.311***	1.000							
	(0.000)								
(3) Element familiarity	0.022***	-0.231***	1.000						
	(0.000)	(0.000)							
(4) Dummy: usage of social media	-0.077***	0.063***	0.144***	1.000					
	(0.000)	(0.000)	(0.000)						
(5) Dummy: founded by women or minorities	-0.029***	0.014***	0.086***	0.108***	1.000				
	(0.000)	(0.000)	(0.000)	(0.000)					
(6) Market growth rate	0.011***	0.052***	-0.231***	-0.053***	-0.123***	1.000			
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)				
(7) Number of valid elements	-0.283***	0.449***	-0.134***	0.094***	0.009***	0.043***	1.000		
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)			
(8) Dummy: no technical elements at all	0.232***	-0.437***	0.152***	-0.020***	-0.003*	-0.004**	-0.335***	1.000	
	(0.000)	(0.000)	(0.000)	(0.000)	(0.086)	(0.021)	(0.000)		
(9) Uncommon elements	0.041***	-0.102***	-0.018***	-0.024***	-0.027***	0.028***	-0.069***	-0.071***	1.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	

*Note: *p < .05, **p < .01, ***p < .001, two-tailed test.*

Table 2.4: FPM model for CrunchBase data.

	Crunchbase Model with Control Variables				Crunchbase Model with Main Effects			
	IPO/high price acquisition	New funding	Other acquisition	Close	IPO/high price acquisition	New funding	Other Acquisition	Close
Local Distance (lower-order invention)					-0.996***	-0.450***	0.520***	0.260*
					(0.143)	(0.028)	(0.044)	(0.122)
Global Distance (higher-order invention)					2.312***	0.900***	-0.102	-2.190***
					(0.275)	(0.057)	(0.095)	(0.230)
Elements Familiarity	-0.177***	-0.001	0.086***	-0.019*	-0.155***	0.025***	0.066***	-0.062***
	(0.015)	(0.002)	(0.003)	(0.009)	(0.016)	(0.002)	(0.004)	(0.010)
Dummy: Usage of social media	-0.017	0.192***	-0.447***	-0.441***	-0.062*	0.174***	-0.404***	-0.399***
	(0.027)	(0.006)	(0.012)	(0.030)	(0.026)	(0.006)	(0.012)	(0.030)
Dummy: founded by women or minorities	-0.413***	0.294***	-0.403***	-0.264***	-0.510***	0.287***	-0.427***	-0.257***
	(0.059)	(0.007)	(0.026)	(0.042)	(0.053)	(0.007)	(0.025)	(0.042)
Market growth rate in the year of establishment	0.013	0.040**	0.039	-0.168	-0.039	0.018	0.043	-0.130
	(0.066)	(0.015)	(0.030)	(0.088)	(0.066)	(0.015)	(0.030)	(0.088)
Text Features (length, missing, unusual words) Year, Location, Industry and Funding History	YES	YES	YES	YES	YES	YES	YES	YES
Constant	-4.215	-4.023	-4.798	-7.185	-5.152	-4.358	-4.724	-6.299
Goodness of Fit	ρ_k^2=,0.40, R^2=0.29				ρ_k^2=0.42, R^2=0.30			

*Note: ***p < .001, **p < .01, *p < .05, two-tailed test.*

We initially explore the relationship between IVs and DVs by correlation. We transform numerical predictors into ten equal-size quantile groups as the *x*-axis, and calculate the average proportion for IPO, M&A and closure (or average number of new funding rounds) for each group as the *y*-axis. The first row of Figure 2.6 graphs the correlation. To show interactive effects of lower- and higher-order inventions, we generate a 3-D wireframe plot for each outcome. Evidence from Figure 2.6 row 1 supports our hypotheses: In general, higher-order invention has a positive correlation with IPO/high price acquisition and new funding attainment, and a negative relationship with other acquisitions and closure. We also plot the estimated subhazard from survival models on the second row of Figure 2.6. The two rows, one based on empirical statistics and one based on our modeling, show similar patterns.

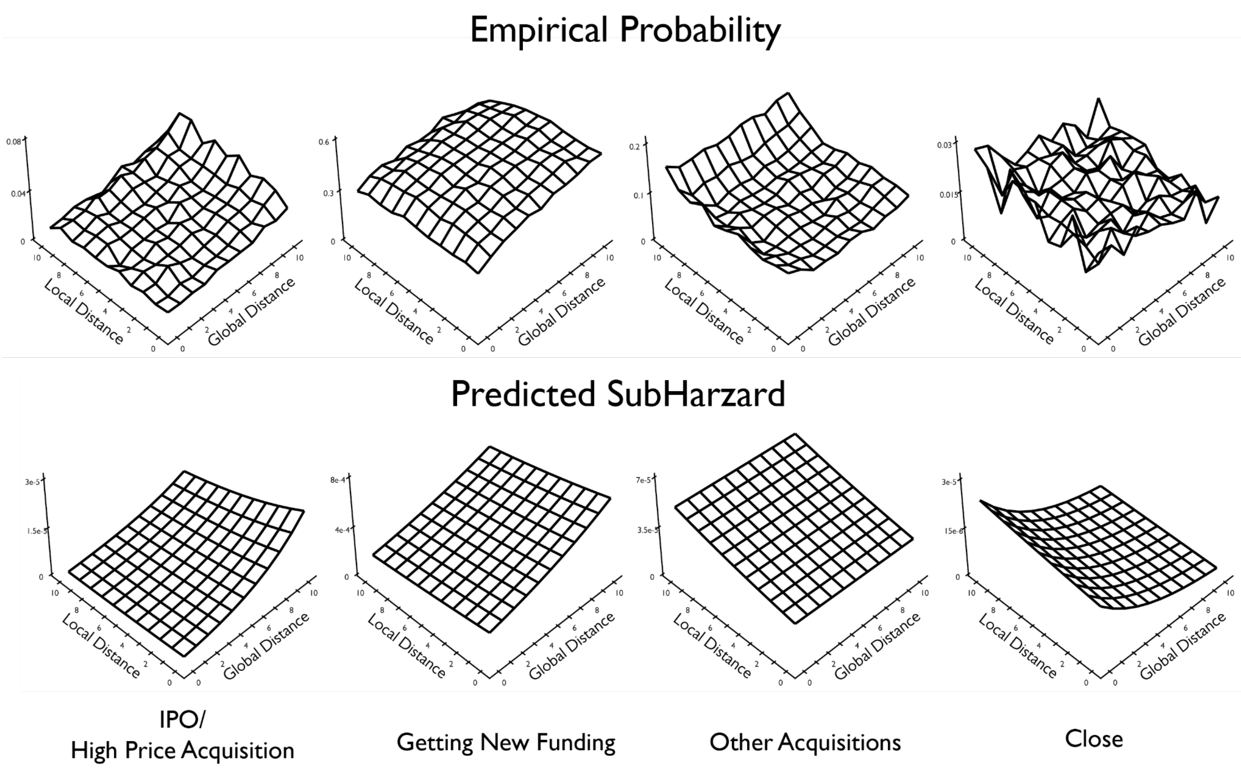

Figure 2.6: Relationship between independent variables and venture outcomes. Upper, equal quantile plots for main effects; Lower, 3D-heatmaps for main effects, smoothed by gaussian kernel.

Event History Model.

We show our event history models based on CrunchBase in Table 2.4. We apply two measurements to assess goodness of fit for our models: variable dependence for censored survival data proposed by (O'Quigley, Xu, and Stare 2005), and a more conservative improvement on this measure by (Royston 2006)). The first is defined as $\rho_k^2 = 1 - exp\left(-\frac{X^2}{e}\right)$, where $X^2 = 2\left(l_{\hat{\beta}} - l_0\right)$ and e equals the number of uncensored observations. The second is defined as $R^2 = \frac{\rho_k^2}{\rho_k^2 + (\pi 2/6) * (1-\rho_k^2)}$, designed as a measure of explained variation for event history models. Although these measures are not as widely used and commonly accepted as adjusted R^2 in linear regressions, they suggest the effect size of the model. Our models perform well according to these goodness-of-fit measures, with ρ_k^2 between 0.42 and .43, and R^2 between 0.30 and 0.31.

These results Table 2.4 are broadly consistent with correlations rendered in Table 2.3 and Figure 2.6. After holding constant all control variables, the relationship between focal independent and outcome becomes clearer for most models, including middle/low priced acquisition. When new ventures engage intensively in lower-order invention, applying technologies to new applications, this leads to a lower probability of IPO/high price acquisition and new funding. Specifically, we estimate a coefficient of -0.996 for local distance on the logged hazard rate that a company will achieve IPO or high priced acquisition. This suggests that with a maximum increase in local distance, from technologies and applications being perfectly related to unrelated or transiting from 0 to 90^{o} degrees of cosine distance, the hazard of going public or being richly bought out decreases by nearly 63.1% ($0.631 \approx 1-e^{-.996}$).[7] This means that as a new venture's components lie far from each other, an indicator of it positioning itself to engage in lower-order invention, its likelihood of ultimate success drops substantially. Similarly, a maximum increase in local distance leads to a 36.2% ($1-e^{-.450}$) decrease in the hazard of obtaining new funding.

Interestingly, when considering competing risks, lower-order invention in terms of local distance increases the hazard of middle and low priced acquisition by 68.2% ($e^{.520}-1$). This suggests that lower order invention, although not the optimal strategy for a founding team, can still make the new venture a desirable target for other companies. These strategies include commercialization of technologies in new application scenarios (see Appendix). If even partially successful, the lower-order activity may create a functional component that can serve as input for a later-stage higher-order invention. Conversely, novelty in higher-order recombination helps start-ups achieve IPO or high-priced acquisition. One unit increase in global distance multiplies the hazard of IPO or high priced acquisition by 10.095 ($e^{2.312}$), yielding a 900% increase. It also raises the hazard of obtaining new funds by 146% ($e^{.900}-1$). By contrast, the same decreases the hazard of ending in a closure by 89% ($1-e^{-2.190}$).

In Figure 2.7, we estimate the hazard rate for two hypothetical companies on the basis of our model. We assume two "ideal types", one with the highest empirically observed higher-order invention (e.g., low local distance, high global distance), and the second with the highest empirically observed lower-order invention (e.g., high local distance, low global distance). According to our theory, these are better and worse companies, respectively. Insofar as our hypotheses hold, we expect much better outcomes to befall the enterprise exclusively focusing

[7] Technically, our model yields the cause-specific sub-distribution hazard rate, which was designed in order to analyze competing risks between different events, as we do here.

on higher-order invention and much worse outcomes those exclusively investing time and energy in lower-order invention. Results from Figure 2.7 meet these expectations: we observe a clear divergence between hazard rates for the two ideal-type ventures. The firm engaged in lower-order invention is much less likely to receive new funding and achieve an IPO, but is much more likely to be sold for a middle to low price or close within a decade.

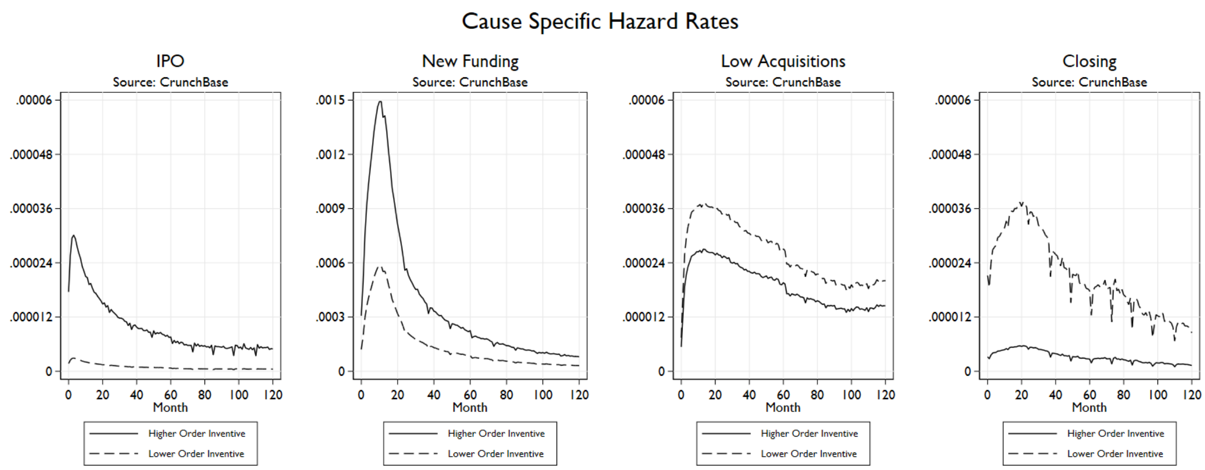

Figure 2.7: Hazard rate prediction for “ideal” companies.

In order to assess the robustness of these event historical findings, we rerun the models using data from VentureXpert. This replicates the model as in Table 2.4 on an overlapping but largely distinctive population of new ventures and events. We further evaluate robustness by accounting for the potential for new ventures to evolve and change with new opportunities in the marketplace. For this, we utilized the Wayback Machine data to account for strategic pivoting in new ventures away from their initial business description and plan. Our results remain robust to these major perturbations and do not change our conclusions (see Appendix).

Mediation Analysis.

Our event history model shows that higher-order invention is associated with new venture achievement of milestone events, while lower-order invention harms it. Here we conduct a mediation analysis to illuminate mechanisms through which the two approaches influence life chances of new ventures through commercialization speed and investment diversity, hypotheses 3 and 4.

We first regress the two mediators on our main independent variables (Table 2.5), and then use the generalized structural equation modeling to estimate and display the effects of two mediators (Table 2.6). In Table 2.7, we decompose that effect into direct effects and indirect effects on startup final outcomes. We also illustrate the path for indirect effects in Figure 2.8.

Table 2.5: Regressing meditators on main independent variables.

	Samples with Seed Round Records (N=49304)		Samples with both Seed Round and Early Round Records (N=19979)	
	Time to Market	VC Diversity	Time to Market	VC Diversity
Local Distance (lower-order invention)	13.546***	-0.105***	-1.039	-0.128***
	(2.313)	(0.019)	(1.687)	(0.038)
Global Distance (higher-order invention)	-32.150***	0.132***	-9.172***	0.134*
	(3.889)	(0.032)	(2.737)	(0.061)
Technology Elements Familiarity	1.438**	0.001	1.596***	0.027**
	(0.540)	(0.004)	(0.444)	(0.010)
Dummy: Usage of social media	1.339**	0.025***	1.135***	-0.006
	(0.468)	(0.004)	(0.292)	(0.007)
Dummy: founded by women or minorities	1.701***	0.035***	1.781***	0.029***
	(0.464)	(0.004)	(0.316)	(0.007)

Market growth rate in the year of establishment	26.868***	-0.018	-1.929**	-0.029
	(1.138)	(0.009)	(0.733)	(0.016)
Text Features (length, missing, unusual words) Year, Location, Industry and Funding History	YES	YES	YES	YES
Constant	43.800***	0.098**	16.437***	0.391***
	(4.318)	(0.036)	(3.568)	(0.080)
N	49269	49269	19976	19976
Goodness of Fit (Adjusted R2)	0.212	0.148	0.183	0.073

*Note: ***p < .001, **p < .01, *p < .05, two-tailed test.*

Table 2.5 shows the effects of invention strategies on the mediation variables. Results are consistent with hypotheses 3 and 4. Holding other variables constant, changing local distance from the empirical minimum to maximum adds approximately 14 months to the period between seed and early round investment (A/B), while changing global distance from its empirical minimum to maximum reduces that period by approximately 32 months. Higher global distance and lower local distance is also associated with the likelihood new ventures obtain diverse support from the investor community, as they attract investors with more distinctive professional backgrounds. Increasing global distance from its empirical minimum to maximum increases the Jaccard diversity score of VC backgrounds by nearly 0.132, and decreasing local distance from its empirical maximum to minimum increases the Jaccard diversity score of VC backgrounds by 0.105. Given the Jaccard diversity score falls in the range [0,1], a growth of 0.105 represents more than an eighth of the total range, suggesting substantial increase in VC knowledge backgrounds. When we replicate these analyses on all samples with both seed round records and early round records, the pattern does not change.

After demonstrating the validity of mediators, we add them into generalized structural equation modeling to identify the effects path. Results are presented in Table 2.6.

Table 2.6: GSEM model for mediation analysis.

	IPO and High Price Acquisition	Other Acquisition	Close	Time to Market	VC Diversity for Early Rounds
Time to Market	-0.009***	-0.003***	0.005***		
	(0.001)	(0.000)	(0.000)		
VC Diversity for Early Rounds	0.613***	0.952***	0.191***		
	(0.085)	(0.040)	(0.062)		
Local Distance (lower-order invention)	-1.780***	0.152	0.427	13.586***	-0.105***
	(0.553)	(0.191)	(0.224)	(2.313)	(0.019)
Global Distance (higher-order invention)	3.191***	0.405	-1.621***	-32.168***	0.131***
	(0.799)	(0.323)	(0.361)	(3.888)	(0.032)
Other Controls	YES	YES	YES	YES	YES
_cons	-1.947	-3.431	0.591	43.967	0.099

*Note: ***p < .001, **p < .01, *p < .05, two-tailed test. Control variables are the same as in FPM. Generalized structural equation model using bootstrapping (1,000 repetitions), bias-corrected coefficients, and robust standard errors.*

For goodness-of-fit measures for GSEM, we report AIC and BIC[8]. To achieve an approximate estimate, we recode the outcome variables as binaries (instead of multi-category outcomes) and re-estimate the equation with an ordinary SEM of the same structure. The output model has an AIC of 784889 and BIC of 786509, compared to 597720 and 599296 for the

[8] For details, see page 3, Intro 7 of SEM in the Stata Manual.

GSEM. This demonstrates that the GSEM improves in its fit of the current data. The ordinary SEM has a RMSEA of 0.028 (<0.05), CFI of 0.99 (>0.95) and SRMR of 0.001 (<0.08), which means our model has fit the data well and captured variation suggesting that it has captured an additional aspect of the social process. Nevertheless, due to the large sample size, the $\chi 2$ test does not work well with a $p < 0.05$.

As we can see in Table 2.6, shorter time to receive early-round funding and more diversified venture capital backgrounds both contribute to the higher probability of IPO/acquisition. Path effects are displayed in Figure 2.8, ignoring direct effects between innovation and performance outcomes. In structural equation modeling, when path A→C co-exists with A→B→C, the total effect from A to C = ac(direct)+ab*bc(indirect). Table 2.7 shows direct, indirect and total effects between our major variables.

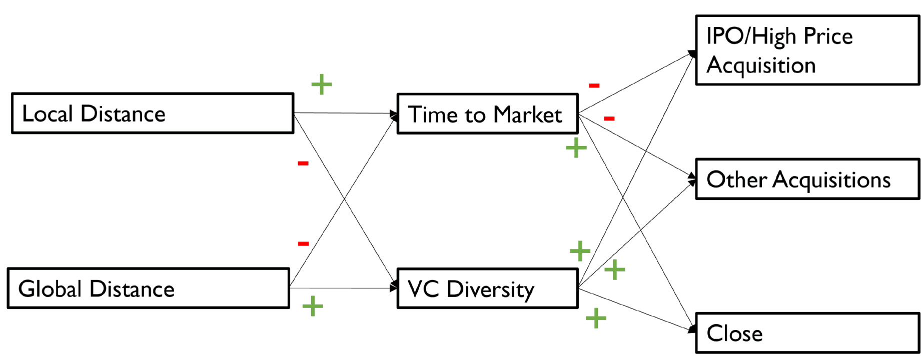

Figure 2.8: Summary of indirect path effects based on generalized structural equation modeling.

Our estimated mediation effects confirm our expectations consistently for all three outcome variables. Higher order inventions significantly accelerate new venture development from idea to market. Higher global distance and lower local distance also bring in broader expertise and insight from investors, and expands the knowledge base accessible to the founding team, which leads to a higher probability for IPO and high-priced acquisition. Interestingly, this broadening effect also significantly adds to the possibility of closure, possibly because heterogeneity among investors brings in disagreements and conflicts, thus higher risks for the company. The direct and indirect effects illustrated in Figure 2.7 echo these patterns.

Table 2.7: Path effects for GSEM.

	Direct Effects	Indirect Effects: Time to Market	Indirect Effects: VC Diversity	Total Effects
DV: IPO and High Price Acquisition				
Local Distance (lower-order invention)	-1.780***	-0.128***	-0.064***	-1.972***
	(0.553)	(0.025)	(0.015)	(0.554)
Global Distance (higher-order invention)	3.191***	0.303***	0.081***	3.575***
	(0.799)	(0.048)	(0.023)	(0.801)
DV: Other Acquisitions				
Local Distance (lower-order invention)	0.152	-0.037***	-0.100***	0.015
	(0.191)	(0.008)	(0.019)	(0.192)
Global Distance (higher-order invention)	0.405	0.087***	0.125***	0.618
	(0.323)	(0.015)	(0.031)	(0.325)
DV: Close				
Local Distance (lower-order invention)	0.427	0.062***	-0.020**	0.469*
	(0.224)	(0.012)	(0.007)	(0.224)
Global Distance (higher-order invention)	-1.621***	-0.148***	0.025*	-1.743***
	(0.361)	(0.022)	(0.010)	(0.361)

*Note: ***p < .001, **p < .01, *p < .05, two-tailed test.*

2.7 Discussion

In this study, we investigate the complex system of organization emergence, and whether it relies on modularity to hedge against the risks of exploratory failure (Simon 1991b) inspired by patterns manifest in the evolution of natural systems. As components coalesce into stable modules, they themselves become objects of macro-recombination, enabling the emergence of not only increasingly complex systems, but those that can support new forms of novel recombination. We demonstrate the value of this complex systems perspective for theorizing the emergence of robust entrepreneurial ventures. We hypothesized that new companies engaged in lower-order invention, crafting new technologies or applying them to new social problems, were unlikely to succeed in the marketplace under the time constraints imposed by venture capital.

Drawing upon developed computational tools, we constructed an evolving map of business discourse and used it to anchor our analysis of the modularity and combinatorial novelty of U.S. entrepreneurial ventures (1976-2020). We projected new venture descriptions (298,915 from Crunchbase; 63,492 from VentureXpert) within this map to evaluate the degree to which they engaged in lower- versus high-order invention.

By testing the relationship between innovation and liquidity events, our empirical evidence demonstrates that lower-order invention, which develops or applies technologies in novel ways, involves higher risks and slows new venture development. It decreases the probability that new ventures receive new funding rounds from venture capital firms, achieve an IPO, or sell for a strong price, and increases the probability that they will shutter. In contrast, those engaged in higher-order invention, assembling diverse applied modules into value-generating systems, were much more likely to obtain rapid new rounds of funding and achieve outsized success through IPO or high-priced acquisition.

In the sociology of innovation, the atypical combination of knowledge often leads to breakthrough science and technology (Cao et al. 2022; Fleming 2001; Shi and Evans 2023; Uzzi et al. 2013), and unexpected combinations of diverse expertise and experience often yield superior product and team performance (de Vaan, Vedres, et al. 2015). The recombination perspective has even been successfully applied to social movements (Wang and Soule 2016) and institutional emergence (Padgett and Powell 2012). Collectively, these studies emphasize surprising similarities between innovation and its outcomes across radically different social domains. Beyond the general association between combinatorial innovation and disruptive success, however, contingency and the compositional structure of innovation has rarely been systematically investigated.

Using modular applied technologies in entrepreneurship as a case, our study operationalizes a framework to examine mechanisms underlying combinatorial innovation across organizations and institutions in society. We theorize and measure social contingencies that vary across contexts of organizational and institutional emergence: (1) composition and integration within lower-order socio-technical modules; and (2) architecture of higher-order modular combinations. Our case serves as an example of both steps in this framework. First, applied technologies serve as functional modules, which form the basis for successful, higher-order entrepreneurial assemblages. Second, startups face a trade-off between higher- and lower-order invention, given the negative correlation between lower-order and higher-order invention indicators (Table 2.3). Finally, higher-order invention prevails as a dominant strategy for entrepreneurship performance.

By examining the social process of new venture creation under this framework, we highlight boundary conditions for recombination in social systems that point to likely heterogeneity across social domains. Domains with a strong incentive for innovative speed will often scaffold their search with previously applied technologies. Consider the famously innovative U.S. defense advanced research projects agency (DARPA), which catalyzed the internet and technology behind Siri and Alexa. DARPA articulates a funding strategy that coincides closely with higher-order invention. It frequently supports attempts to rapidly assemble diverse, stable technologies into novel systems (e.g., the Internet) designed to surprise rather than be surprised by competitors.[9] Other domains without the same capitalist or military pressure for rapid success may rely less on the modularity of applied technologies, or rely on modularity at different levels of the system.

Multiple subfields in sociology might benefit from our theoretical framework. In the sociology of science, knowledge and technology, Latour, Callon and others conceptualize science as a complex, dynamic network. Within this thicket, scientists, institutions, concepts, physical entities and forces "knit, weave and knot" together into an overarching scientific fabric (Callon 1986; Latour 1987; Latour and Centre de Sociologie de L'Innovation Bruno Latour 1999; Latour and Woolgar 1979). Actor network theory highlights the dynamic multi-mode, multi-scale character of scientific networks as intrinsically complex systems. Early work seeking to formalize and measure such networks with data from scientific artifacts resisted the regularity of constituting patterns (Callon, Rip, and Law 1986). Our work highlights that the social actors and nonhuman actants who mobilize one another by attending to stable modules of science are more likely to achieve influence across the network.

[9] "The genesis of…DARPA…dates to the launch of Sputnik in 1957, and a commitment by the United States that, from that time forward, it would be the initiator and not the victim of strategic technological surprises" (https://www.darpa.mil/about-us/about-darpa, accessed June 10, 2023).

Our work also speaks to theory about the importance of diversity in teams, networks, and organizations for innovation (Page 2008, 2019). Most previous work on creativity and novelty emergence demonstrates how social environments desirable for innovation catalyze "creative conflicts" (Cao et al. 2022), either by unleashing the free flow of new ideas (Stark 2011b) or by creating a stable environment that supports higher levels of cognitive heterogeneity and resulting debate (Goncalo et al. 2015). Nevertheless, recent work on information benefits of social networks highlight the danger of information overload (Ter Wal et al. 2016). Without a shared cognitive schema, diversity can bring more chaos than creativity. Our work shows that successful recombination relies on the maturity of the components combined. More specifically, our analysis suggests that modules of problems and solutions may combine in other organizational settings to spur flexible and innovative collective cognition. This perspective inverts the garbage can theory of organizations, through which problems are chaotically and inconsistently paired with solutions in modern organizations (Cohen et al. 1972). Insofar as problems and solutions become stable modules within a complex organization, like technologies and applications in the new ventures examined in our study, the diversity and even incommensurability of those modules may enable that organization to fluidly reorganize in response to higher-order challenges and opportunities.

Finally, our work directly contributes to organizational ecology, positioning theory, and organizational learning in entrepreneurship. Research on firm positioning has explored the notion of cultivating an industrial "niche" (Hannan and Freeman 1977) and achieving "optimal distinctiveness" (McKnight and Zietsma 2018). Our work contributes to this intellectual tradition, and provides an alternative explanation for the inverted-U shaped relationship between differentiation from industry incumbents and organizational performance (Barlow, Cameron

Verhaal, and Angus 2019; McKnight and Zietsma 2018; Taeuscher and Rothe 2021; Zhao et al. 2018). According to that work, successful organizations must remain close enough to the pack to retain legitimacy, but far enough to leverage novel competitive advantage. Alternatively, our theory and findings suggest that success at intermediate differentiation may stem from maintaining stable and even familiar lower-level modules, combined in a novel way.

For entrepreneurship scholarship, our approach helps explain recent empirical evidence about new venture strategy, including the disadvantage of young companies engaged in lower-order invention (Roche et al. 2020) and the low growth rate of new ventures with excessive differentiation (Guzman and Li n.d.) All of the empirical evidence suggests that public funds should be inclined to lower level discoveries and inventions. These involve substantial risk but serve as public goods for successful entrepreneurs engaged in higher-order system assembly (Jones and Summers 2020).

Despite its merits, our work has multiple limitations. We did not have the data to incorporate network effects or geographical considerations into our analysis, despite their proven significance for entrepreneurial success (Sorenson 2018). Future study could examine spatio-temporal contexts to investigate how such factors moderate the relationship between new venture structure and efficiency. Furthermore, although we have incorporated wayback machine data for robustness check, it only captures description changes after 2013, and does not allow us to comprehensively examine the full history of new venture pivoting. Prior research indicates that when longitudinal data is involved, conclusions drawn from VentureXpert remain robust (Y. Shi, Sorenson, and Waguespack 2017). Others have noted that the core business lines and ideas typically remain stable as start-ups grow (Kaplan, Sensoy, and Strömberg 2009) despite frameworks that encourage them to actively experiment and shift on a tactical level (Gans, Scott, and Stern 2022; Ries 2011). Our robustness check on the Crunchbase descriptions based on Wayback machine data suggests the same pattern. Therefore, we assume that when incorporating complete longitudinal data, our main conclusion remains robust. More complete data will be required, however, in order to rule out the interaction between modular company descriptions and company pivots.

Not all inventive combinations are equal. By examining tens of thousands of U.S. venture-funded start-ups over the past half century, we uncover the importance of recombining modular applied technologies to rapidly assemble innovative and persistently value-generating business organizations. Future research should consider how functional modularity scaffolds innovation in fields ranging from science and technology and to finance and business to culture and the arts (Wimsatt 2014).

Such work should also look beyond "modularity" (Simon 1991b) to additional principles through which complex social systems hedge against exploratory failure, such as search along "adjacent possible" innovations one-step from the current frontier (Kauffman 2002). By uncovering the syntax of social innovation, such investigations will not only enable deeper understanding of how actors manage risks and opportunities underlying attempted innovations, they will also enable us to anticipate the range of social inventions most likely to emerge.

CHAPTER 3 DEEP VERSUS BROAD TECHNOLOGY SEARCH AND THE TIMING OF INNOVATION IMPACT

3.1 Introduction

Recombination has become a dominant perspective for understanding and predicting technological evolution (Brian Arthur 2009; Fleming 2001; Weitzman 1998; Xiao et al. 2022). Stemming from the theory of complex, adaptive systems, this approach formalizes technological change as a process of recursive recombination, where more complicated knowledge and technologies are based on the nested combination of modular ideas and techniques from diverse sources (Simon 1991b). This perspective has been used to explain the evolution of many complex knowledge systems, including science (Uzzi et al. 2013), technology (Fleming 2001), and business (Stark 2009). Empirical studies show that these systems can be modeled and predicted, once their underlying components have been identified (Fleming and Sorenson 2001, 2004; Shi et al. 2015; Sorenson, Rivkin, and Fleming n.d.; Sourati and Evans 2023).

Successful recombination requires assembling components together in a way that yields novel capability. Two recent streams of research investigate this phenomenon. The first focuses on *features* of individual components for innovation outcomes, including component newness (Jung and Lee 2016), particularity (Galunic et al. 1998) and familiarity (Fleming 2001). The second examines component *architecture*. When components are brought together, analysts have considered the structure and diversity of knowledge sources. This includes “depth” versus “breadth” (Kaplan and Vakili 2015), hierarchy and “modularity” (Ethiraj, Levinthal, and Roy 2008), and network features (e.g., “centrality and ”structural holes", see Dhanaraj and Parkhe 2006) as important predictors of invention novelty and impact. In this article, we advance on this second line of investigation, formalizing *hierarchical relationships* between knowledge components within the broader technological system. By leveraging computational representation learning and theoretical insights from complex systems, we model the hierarchical tree of technological evolution, and construct rich, intuitive measures of recombination that allow us to explore the impact and mechanisms of strategic search behavior relevant to knowledge systems and collective intelligence.

With this new perspective, we reformulate a classic problem in the literature on recombinant innovation: strategic consequences arising from *depth* versus *breadth* of technological search. The "tension" perspective on recombination (Kaplan and Vakili 2015) argues that distant or diverse knowledge breaks cognitive inertia and stimulates breakthrough ideas (Fleming 2001; Hall, Jaffe, and Trajtenberg 2001; Trajtenberg, Henderson, and Jaffe 1997; Verhoeven, Bakker, and Veugelers 2016). The "foundational" perspective (Kaplan and Vakili 2015), by contrast, celebrates deep immersion in a particular domain, preparing inventors with expertise requisite to identify anomalies and resolve them (Csikszentmihalyi 2009; Kaplan and Vakili 2015; Sternberg and Lubart 1995). Conflicting evidence has led subsequent work to view depth and breadth as more or less strategic contingent on context, such as the pace of change in knowledge (Teodoridis et al. 2019), network structure (Ter Wal et al. 2016) and institutional environment (L. Zheng et al. 2022).

Current challenges to research on depth versus breadth involve the difficulty of robust measurement, rich but conflicting evidence, a lack of clarity on the mechanism relating search strategy to outcome, and an absence of discussion on the level of collective invention and innovation. These challenges cannot be easily addressed through an industry-specific, small-data approach. To address these challenges, we reformulate depth and breadth into a modeling framework that explicitly accounts for the hierarchy of technological components across the entire knowledge system. We do this using the Poincaré embedding method, a machine learning algorithm (Nickel and Kiela 2017a) that facilitates empirical-based construction of a hierarchical knowledge map from the network of combination dependencies. Within this representation of the technological tree, we precisely operationalize *depth* as distance from the location of the knowledge components to the root of the tree, and *breadth* as difference in the distance between tree branches to be combined. In short, depth is about position within the knowledge hierarchy, and breadth about the range between branches. Based on these measures, we empirically examine the individual and collective outcomes of these two search strategies, and the change in effects across time and levels of analysis.

Our work serves as the first to systematically examine deep and broad search and their outcomes using highly precise technological mappings (Aharonson and Schilling 2016; Escolar et al. 2023). While many existing studies on the same topic are based on observations within a single industry, and must reconcile conflicting evidence with competing theories, our novel methods use the structure of the whole knowledge system as a frame of reference, ensuring cross-domain comparability from the outset. This approach enables us to clearly observe the social dynamics of knowledge diffusion and influence accumulation in an intuitive and trackable manner. The empirical patterns revealed through our mapping, with its mathematical accuracy, stability and consistency, provide a simple, clear, and compelling understanding of search outcomes in both deep and broad searches. Rather than focusing solely on interpretation, our work opens opportunities for prediction, making it especially valuable for policymakers and industry decision-makers seeking timely and quantifiable insights into knowledge frontiers.

The remainder of the paper is arranged as follows. We begin by reviewing literature on recombination theory, specifically focusing on relevant debates surrounding deep versus broad search strategies in technological recombination. Next, we introduce the representation learning framework. We discuss how the Poincaré embedding model extracts critical information from empirical data, and produces insightful metrics to understand and evaluate deep versus broad strategies in empirically and theoretically grounded ways. We then validate our depth and breadth measures, nearly independent based on our method, with both human interpretable examples and correlation with classic measurements. Next, we examine the effects of depth vs. breadth of recombination on the impact of patented technologies, and explore their mechanisms through the analysis of large-scale data. Results show that while deep recombination leads to higher short-term citations and immediate economic value, broad recombination stimulates long-term impact and breakthrough discoveries. These effects persist when we strictly control patent quality by including only patent twins, which have exactly the same content but were patented in the United States (i.e., USPTO) and the European Union (i.e., EPO), but in different knowledge systems with distinct histories and structure. Our exploration of mechanisms shows that depth yields short-term impact through immediate within-domain recognition, while breadth effects results in long-term returns through broader scope of diffusion. Finally, we discuss the implication of our results in organizational settings. We demonstrate how inventions within organizations are more rapidly appreciated and built upon, accelerating collective technological advance.

3.2. Depth versus Breadth

3.2.1 Depth and Breadth Search Strategies in Technological Recombination.

The well-known dichotomy of exploration versus exploitation (March 1991) separates those who burrow deep into a singular domain from those who travel and link many, memorably characterized by Isaiah Berlin as “hedgehogs” and “foxes” (Berlin 2013; also see Dyson 2015). Recombinant depth has become associated with local search and the exploitation of prior technological success. Deep strategies foster robust understanding, anomaly detection, and the integration of components underlying a technology (Kaplan and Vakili 2015; Mannucci and Yong 2018; Teodoridis et al. 2019). By contrast, recombinant breadth has become associated with broad and even global exploratory search, achieved through the radical combination of technological components in novel and unexpected ways.

In the context of technological and commercial development, this distinction was rearticulated as the “foundational” and the “tension” perspectives, which respectively favor depth versus breadth of technological search (Kaplan and Vakili 2015). The foundational view celebrates deep immersion in a particular domain, as it prepares inventors with an expert mindset to identify anomalies worthy of resolution (Csikszentmihalyi 2009; Kaplan and Vakili 2015; Sternberg and Lubart 1995). The tension view, by contrast, claims that recombination of distant or diverse knowledge breaks cognitive inertia and stimulates breakthrough ideas (Fleming 2001; Hall et al. 2001; Trajtenberg et al. 1997; Verhoeven et al. 2016).

While early studies sought to distinguish whether deep or broad strategies were more associated with success, later scholarship has aimed to identify important contingencies for these effects. They reveal that depth and breadth effects are highly conditioned by specific strategic contexts. Contexts that mediate the success of search depth versus breadth include the pace of change within a knowledge domain (Teodoridis et al. 2019), network structure shaping information flows (Ter Wal et al. 2016) and the broader institutional environment (L. Zheng et al. 2022). Further studies demonstrate that depth and breadth might beneficially be balanced within a firm (Brahm, Parmigiani, and Tarzijan 2021; Stettner and Lavie 2014).

The heterogeneous menagerie of measures for depth and breadth employed in prior research could also help explain why evidence regarding when to employ each strategy may conflict. In work on how individuals and companies search and what they know, scholars have deployed qualitative information (Boh, Evaristo, and Ouderkirk 2014; Prencipe 2000) and survey methods (Terjesen and Patel 2017; Zhou and Li 2012) to study depth and breadth of expertise. In work analyzing these in terms of knowledge recombination, scholars also rely on existing knowledge classification systems. These include the United States Patent Classification (USPC) or International Patent Classification (IPC) for patents (Kaplan and Vakili 2015; Moorthy and Polley 2010; Nakamura et al. 2015), disciplines or sub-domains for science (Teodoridis et al. 2019; H. Zheng, Li, and Wang 2022), and genres for cultural products, from music to games (Shi, Lim, and Suh 2018; de Vaan, Stark, et al. 2015). Such measures calculate recombination breadth based on component diversity as captured by the Herfindahl index, or the average distance between components based on prior combinations, and assume depth as the inverse of breadth (e.g., Kaplan and Vakili 2015; Lodh and Battaggion 2015) or repeated usage of the same category (Chattopadhyay and Bercovitz 2020; Kim, Guler, and Karim n.d.; Paruchuri and Awate 2017).

In recent years, with the growth and accessibility of open datasets and wide application of scientometric techniques, this second approach to study search strategies has gained momentum (Xiao et al. 2022),. However, existing measures exhibit several major methodological limitations that lead to theoretical misconceptions.

First, existing methods do not describe search paths directly, or construct measures atop untested assumptions. In the case of depth and breadth, such assumptions include placing depth and breadth as poles of a single dimension of search (Kaplan and Vakili 2015; Lodh and

Battaggion 2015; Moorthy and Polley 2010; Mueller et al. 2021; Nakamura et al. 2015; Papazoglou and Spanos 2018; Teodoridis et al. 2019). Depth of search has also been understood as richness or frequency of experience by the inventor (Chattopadhyay and Bercovitz 2020; Kim et al. n.d.; Paruchuri and Awate 2017), but not position of the invention with respect to the technological frontier. In this paper, we argue that depth and breadth should be considered as two independent dimensions of the search path.

Second, most measures of technological recombination use the structure of institutional classification systems (e.g., USPC or I/CPC) to derive technology distances. Categories within these systems, however, can be idiosyncratic and play distinctive roles within the overall combinatorial system, such that they do not consistently cluster. This violates inventors' cognitive experience of similarity. For example, consider patents within the CPC category B ("Performing operations/transporting"). This category includes fundamental processes like boiling, mixing, separating, and sorting that are essential for multiple industries, providing infrastructure critical for a wide array of other technological processes. As a result, category B connects many other inventions. Despite their conceptual similarity in the classification system, components within category B are widely dispersed across industrial applications. In contrast, some technological components, such as "computer hardware" and "computer software," are classified into distinct conceptual categories yet remain closely linked in practice. These examples challenge the "equi-distance assumption," which suggests that the taxonomy system accurately reflects technological practice in inventor communities. This assumption has faced growing criticism in the strategy and economics literature (Ghosh, Ranganathan, and Rosenkopf 2016; Teodoridis, Lu, and Furman 2022) and is particularly problematic over the long term due to concept drift (D'hondt et al. 2014).

Distortions in operationalization are not merely technical errors that introduce minor inaccuracies. One direct outcome of this empirical simplification is that it constrains the ability of existing studies to describe and examine search represented by collective invention—the form of knowledge production that is becoming increasingly dominant in both industry and academy.

3.2.2 Technological Invention as a Collective Effort

Between the end of the 19th and the early 20th century, the nature of invention underwent systematic transformation, from an individual pursuit to a corporate activity. During this period, inventions previously pursued and legally assigned to individual inventors became predominantly targeted, managed and owned by firms and related research organizations (Lamoreaux and Sokoloff 1999). These organizations were driven by incentives to make strategic technological investments that would bolster the success of their research and patenting activities. They not only produce novel investigations and landmark patents, but also clusters of defensive research and patents required to legally protect the former (Shapiro 2000). These developments transformed invention into a collective endeavor, where organizations, rather than individual inventors, now primarily drive both knowledge production and consumption.

This collective nature of invention has been discussed broadly in the literature. In the economics of innovation, a stream of research focuses on estimating the quantity and composition of knowledge spillovers within or across organizational boundaries (Bernstein and Nadiri 1989; Bloom, Schankerman, and Van Reenen 2013; Guillard et al. n.d.). This highlights an important part of invention return as social returns to the public. The strategy literature on combinatorial innovation, however, has predominantly treated technological impact individually in terms of future or "forward" citations (Fleming 2001; Kaplan and Vakili 2015; Verhoeven et al. 2016).

The collective perspective has two key implications for understanding deep vs. broad search. First, from the perspective of knowledge production, depth and breadth should be seen as features of a broad collection of search paths within collective practice. The length of a search is determined not by conceptual differences between categories, but by how closely or distantly they are applied in industrial practice and technological consensus. Collective understanding plays a crucial role in shaping the cognitive landscape of search. Modern inventions are predominantly discovered through continuous exchanges of information and ideas with colleagues and clients, instead of by inventors in isolation, performing experiments or poring over records of past inventions. To capture this practical proximity in collective understanding, existing taxonomies are insufficient because of equi-distance distortions described in section 3.2.1 (Ghosh et al. 2016; Teodoridis et al. 2022). A technological mapping that accounts for how technologies are used is required (Aharonson and Schilling 2016; Escolar et al. 2023; Linzhuo et al. 2020; Teodoridis et al. 2019). In this paper, we calculate search distance within the collective knowledge system using a representation learning algorithm, and capture the search path through intuitive metrics and visualizations that follow from it.

Second, and more importantly, if search has an impact on knowledge consumption, it unfolds through collective evaluations and reactions. Once a focal invention is published, the trajectory and timing of follow-up works is key to understanding the nature and mechanisms of its impact. Recent research has moved beyond numerical measures for invention success, such as citations, and increasingly recognized the diverse roles that different types of knowledge play in collective inventions. Imagine two pieces of knowledge are equally important, but follow different patterns—one remains unnoticed for years before gaining recognition, known as "sleeping beauties" (Ke et al. 2015), while the other receives immediate but short-lived attention

(Kang et al. 2024; Silva et al. 2020). These two forms of knowledge have different strategic values for both inventors and organizations. To date, the link between an invention's characteristics—such as search depth and breadth, and its role within collective intelligence, remains largely unexplored. To better understand this complexity, scholars need to observe knowledge evolution as a dynamic process and systematically track how an invention is received. This includes examining the speed and spread of recognition, intensity and shifts in audience attention, and patterns of diffusion—all of which remain underexplored due to methodological limitations.

In summary, collectives fundamentally influence how knowledge is produced and consumed, but its impact has not been fully examined. Without a method to model the landscape of collective knowledge and the structure of search paths, inventions can only be treated as isolated events, obscuring their collective nature. Our work leverages machine learning algorithms to fill this gap, building on theoretical insights from technological evolution theory and the complexity perspective.

3.2.3 Theoretical Bases for a Hyperbolic Representation of Collective Knowledge Systems

Theorists of technological evolution have examined how past inventions shape the emergence of future knowledge, providing a framework for empirically analyzing knowledge as a collective system. Theoretical discussion of technological evolution (Arthur and Polak 2006; Brian Arthur 2009) and empirical work on technological trajectories (Fontana, Nuvolari, and Verspagen 2008; Huenteler et al. 2016) suggest that technologies evolve progressively and recursively. Some technological components serve as the basis for higher-level systems that emerge from them. Based on this insight, technological knowledge, which is created and stored within human collective understanding, should have a hierarchical, tree-like structure. This notion echoes philosophical discussions about knowledge ranging from Aristotle and Augustine of Hippo to the "tree of knowledge" structuring Diderot and D'Alembert's encyclopedia (Aristotle 1999; Augustine 2003; Weingart n.d.).

This hierarchy goes hand in hand with "modularity," first introduced by Simon (1991b) in the study of complex systems and later expanded by Arthur (2009) in the context of technology. In hierarchical, modular knowledge systems, lower level components integrate into robust, tightly connected subsystems, which are recursively combined to create more complex technological designs. In this context, if depth and breadth are seen as characteristics of search paths, then deep local search involves diving into the underlying technological infrastructure—refining knowledge components within submodules and strengthening their architectural connections. Broad search, on the other hand, incorporates knowledge from other domains. These two processes are not necessarily in conflict; rather, they may reinforce each other. Depth produces more robust modules that can be widely reused, enabling both the overall system and individual corporate innovators to support broader and more exploratory recombinations. Conversely, the outputs of broad search may initially lack robustness, requiring a phase of refinement and integration before they can be effectively recombined in a durable way.

We first consider the value of deep expertise for building integrated technologies. Existing literature reveals that deep local search helps individuals and companies better understand underlying connections between components. This cultivates the familiarity and expertise required to identify anomalies, challenges, and limitations that become the targets of innovation (Kaplan and Vakili 2015; Mannucci and Yong 2018; Teodoridis et al. 2019). In sum, depth entails a comprehensive understanding of the underlying technological principles and infrastructure within a domain.

Theories for technological evolution posit that complex technological structures emerge from the recursive combination of preexisting lower-level subsystems. An example is a wind turbine, which is made up of four subsystems: the turbine rotor, power train, mounting and grid connection. The rotor, in turn, is made up of more granular components such as blades, hub and rotor control systems, etc. (Huenteler et al. 2016). To invent in the field of wind turbines requires re-design of a sub-component or making a new combination of existing sub-components. In this view, the depth of an expert or firm's understanding about a technology is proportional to their innovative recombinations of "deep", lower level components. Depth does not imply narrowness. Consider an inventive team that combines an electrical engineer and a material scientist who themselves bring together deep expertise in processors and 2D materials to invent a next generation semiconductor. In this case, the team's collective deep understanding of two very different fields can be combined into an invention both deep and broad. The same can be assessed at the level of a firm patent portfolio.

Now consider search breadth. In a multi-layered, recursive knowledge system, broad search involves not different fields, but fields falling on diverging branches of the knowledge tree. For example, integrating quantitative sociology with statistics is not as broad as combining it with biology, even if both may seem equally distant in the conceptual taxonomy. This is because statistics has contributed to the foundation, or "knowledge roots", of applied social science, whereas biology belongs to a distinctive branch of knowledge.

Breadth does not imply shallow recombination. Some individual scientists and inventors are generalists who master multiple areas of knowledge and can work across them creatively. Despite the rise of specialists in science (Jones 2009), generalists continue to play a critical role in transformational research (Milojević 2015; Risha et al. 2023). The assemblage of multidisciplinary teams make the independent consideration of depth and breadth a necessity, with teams explicitly designed to vary expertise breadth across researchers and depth within them (Ahmadpoor and Jones 2019). This can also be assessed at the firm-level across all inventor teams. We take this into consideration by creating a breadth measure that directly involves separation between combined domains, whose assemblage within novel technology reflects and often results from creative conflict (Stark 2009).

Therefore, we argue that invented technologies may be broad, deep, both or neither in knowledge recombination, and that depth and breadth should be measured independently for conceptual and strategic clarity. Depth involves comprehensive understanding and full utilization of components based on the evolution of knowledge, while breadth involves familiarity with distinct domains that could yield novel integration. In this way, both depth and breadth are measured as features of search paths, instead of features of the inventors who pursue them. From this standpoint, we base our measures of depth and breadth on an empirical mapping of technology, and operationalize depth and breadth as independent dimensions of technological recombination. We detail this method in section 3.2.4.

3.2.4 Constructing a Hyperbolic Representation of the Technological Tree

Current research calls for a model that accurately represents the hierarchy of technical knowledge while allowing us to project the influence trajectory of an invention within it. This new representation should construct an empirically verifiable basis for distances between components, which goes beyond *a priori* classification systems, and thus facilitates accurate measurements of collective invention trajectories. To this end, we turn to hyperbolic geometric representations that natively encode hierarchy and allow for the independent measurement of hierarchical depth and breadth.

Hyperbolic space exhibits negative curvature, with the distance between data points increasing exponentially from the center to the margin. Embedding algorithms that project hierarchies within hyperbolic space do so without distortion and have successfully captured the structures of many significant cultural and social hierarchies (Chamberlain, Clough, and Deisenroth 2017; Chami et al. 2019; Fellbaum 2005; Linzhuo et al. 2020; Nickel and Kiela 2017b). For example, when researchers embedded WordNet, a hierarchical graph of words connected by shared meanings of synonymy, into into a hyperbolic space, it required only a few dimensions (2-5) but hundreds in Euclidean space (Chamberlain et al. 2017; Fellbaum 2005; Nickel and Kiela 2017a). Moreover, the hyperbolic space model has been shown to unify both "small-world" (locally clustered but globally connected) and "scale-free" (uneven popularity) properties of social and knowledge networks (Krioukov et al., 2010; Papadopoulos et al., 2012).

Figure 3.1 offers a conceptual explanation of hyperbolic geometry using popular hyperbolic embedding representation, the two dimensional Poincaré disk, using machine learning. When lines and points on a hyperboloid surface are projected onto such a disk representation, lines transform into circular arcs, interior angles within a triangle sum to less than 180°, and distances increase exponentially. Now, consider a tree-like hierarchical structure on the hyperboloid surface. When projected onto the disk to obtain a polar coordinate (r, θ), the radius r intuitively represents a component's depth in the hierarchy, while the angle between coupled components represents the breadth of the invention that couples them. This clear mathematical distinction between the two dimensions is what we leverage in our work to facilitate direct and accurate measures of recursive combination within a hierarchical system.

Hyperbolic embedding outperforms traditional representations, such as flat, Euclidean space. Consider a hierarchical tree mapped to a flat, two dimensional Euclidean space. As hierarchical layers (x) increase, the number of leaves at each layer grows exponentially (4^x) and the distance between leaves at each layer must shrink as the area of a circle grows quadratically relative to its radii. More Euclidean dimensions cannot solve this problem, failing to keep pace with exponential growth. In this representation, tree depth—its vertical span from center to margin—negatively correlates with breadth—the distance between two marginal nodes. This distorted cognitive map suggests a necessary trade-off between depth and breadth, which finds frequent expression in the operationalization of depth and breadth from existing literature described above (Kaplan and Vakili, 2015; Lodh and Battaggion, 2015; Moorthy and Polley, 2010; Mueller et al., 2021; Nakamura et al., 2015; Papazoglou and Spanos, 2018; Teodoridis et al., 2019).

In the following methods section, we first delve into the technical details and mathematical foundation for the Poincaré embedding algorithm. Second, we construct a directed citation network, following scientometrics standards, between technological categories to trace knowledge diffusion (Huenteler et al. 2016; Malhotra et al. 2021). Third, we use neural network techniques to embed this network onto a Poincaré disk, which retains the hierarchical network information with minimal distortion, and displays it intuitively in two dimensions. The Poincaré disk allows us to identify both the position of specific categories and the structure of recombination, in a geometrical way. Finally, we demonstrate how we can trace knowledge diffusion paths for a granted patent, and observe when and where follow-up work emerges across the hyperbolic space of inventions. Building on this, we compile a longitudinal dataset that allows us to test our hypothesis about collective invention, and calculate the temporal change of depth and breadth effects on technological impact.

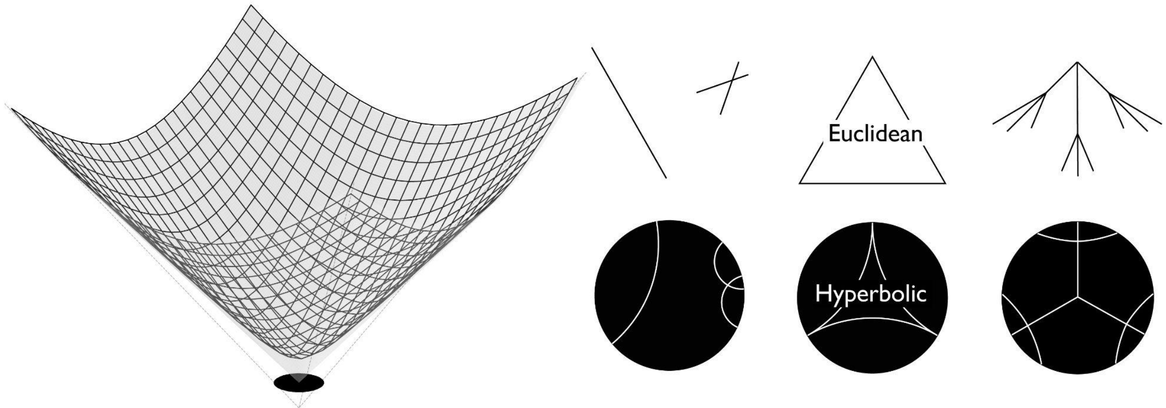

Figure 3.1: Hyperbolic, poincaré disk embedding

3.3. Measuring Hierarchy with Hyperbolic Geometry to Assess Search Strategies in Patents

In this section, we propose complementing classic recombination measurements by introducing a model that accounts for latent knowledge hierarchy. We first present the Poincaré embedding method using hyperbolic geometry, followed by a discussion of novel measures that natively address the theoretical requirements described above. Our approach introduces a theory of depth and breadth complementarity and opens new possibilities for future study of combinatorial innovation, as also social and technological networks more broadly.

3.3.1 Representation Learning and the Embedding Method

Machine learning has become increasingly popular in management research in order to (1) aid human coders in processing raw data (Harrison et al. n.d.), (2) extract subtle features from complex unstructured data that are otherwise difficult to quantify (Choi et al. 2021; Choudhury et al. 2019; Sgourev, Aadland, and Formilan 2023; Vicinanza et al. 2023), and (3) reveal robust patterns from structured data to support theory-building, as a supplement to traditional statistical models (Choudhury, Allen, and Endres 2021; Kumar, Liu, and Zaheer 2022; Shrestha et al. 2021). The rapid growth of computational resources and data availability (Salganik 2019) has motivated researchers to validate existing theories and test new hypotheses in a novel, computational way (e.g., Goldberg et al. 2016).

Among these machine learning methods, embedding algorithms have become a powerful way of representing complex structured and unstructured data. Embeddings are a form of representation learning, a branch of machine learning that has garnered significant attention in the field of organizational research (Aceves and Evans n.d.). Representation learning aims to automatically discover meaningful and predictive representations of raw data that enable the extraction of features for analysis. Representation learning algorithms most notably include "auto-encoders" that automatically encode data into a representation of reduced dimensionality. In principle, these algorithms are similar to matrix factorization approaches such as factor analysis, principal components analysis, and singular value decomposition (Levy and Goldberg n.d.), but tuned for efficient computing with large-scale data and allowing for representational flexibility that tune them to particular empirical and theoretical demands (Baevski et al. 17--23 Jul 2022; Hinton and Salakhutdinov 2006; Vincent et al. 2008). These algorithms reduce high-dimensional inputs into lower-dimensional expressions, producing outputs suitable for subsequent analysis and visualization. With their ability to distill complex information into clear, robust measures, representation learning has the potential to complement traditional statistical models within organizational studies.

Organization science studies often involve text data, and so representation learning for natural language processing ("NLP") has become particularly prominent in strategic management literature, especially topic modeling, word and contextual embedding methods (Aceves and Evans n.d.). Topic models employ probability models to analyze the collocation of words within a corpus and generate latent, low-dimensional "topics" tuned for human comprehension (Blei, Ng, and Jordan 2003; Evans and Aceves 2016; Mohr and Bogdanov 2013). Word embedding, by contrast, uses more precise word proximity and sequence information to capture semantic relationships, representing words as high-dimensional vectors (Aceves and Evans n.d.; van Binsbergen et al. n.d.; Joulin et al. 2017; Kozlowski et al. 2019; Mikolov, Chen, Corrado, and Dean 2013; Pennington, Socher, and Manning 2014b) or context-contingent vector clouds in continuous geometric space (Devlin et al. 2019; Radford et al. 2018).

In this paper, we use another branch of representation learning—network embedding—which simplifies a network representation for analysis. Like word embedding, network embedding aims to learn simple and interpretable continuous representations from complex discrete data. While word embedding uses words or words-in-context as basic units for representation, network embedding is designed to represent nodes, edges, or entire networks as units for representation. Many popular network embedding methods represent networks in flat, Euclidean space, but require many dimensions (Grover and Leskovec 2016; Hamilton, Ying, and Leskovec 2017; Perozzi, Al-Rfou, and Skiena 2014). Insofar as knowledge diffusion trajectories manifest hierarchical structure, with core technology categories serving as foundations and building blocks for others, we adopt the Poincaré embedding algorithm designed to natively encode network hierarchy with hyperbolic geometry (Chamberlain et al. 2017; Nickel and Kiela 2017a).

3.3.2 Technological Evolution, Hyperbolic Geometry, and Poincaré Embedding

Technological progress does not occur simultaneously across all domains. Instead, technology evolves unevenly, with certain elements changing first and then stabilizing, laying the foundation for complementary change. For example, a study conducted by Huenteler and colleagues (2016) found that the wind turbine underwent sequential development through its four subsystems: its rotor (core subsystem), power train, mounting and encapsulation, and grid connection, over the past half century (1963-2009). Rotor parameters varied widely between 1975 and 1990, but no significant changes emerged after 1995. In later years, as the rotor stabilized, attention shifted towards innovation in the powertrain, then the other two subsystems in sequence. This hierarchical sequence of technological development, with central components stabilizing, incentivizing subsequent advances to build upon them, has been observed across several subdomains of invention (Brian Arthur 2009; Fontana et al. 2008; Huenteler et al. 2016; Verspagen 2007).

A common approach to quantify technological trajectories is to construct and analyze networks of knowledge diffusion between patented inventions (Chang, Lai, and Chang 2009; Feldman and Yoon 2012). Technical classes reveal the inheritance of knowledge inheritance, as citing patents draw inspiration from those cited. Unlike citations within research articles, patent citations are countribed in roughly equal parts from authors and patent examiners (Alcácer and Gittelman 2006). Patent examiners possess technical expertise in the patented area. Insofar as they require the addition of influential patent citations intentionally or accidentally excluded, this adversarial process results in a more balanced crediting of historical technical influence.

Technological networks can be extracted in two different ways. Citations linking focal patents and source patents form a one-mode network that traces the directed network of claimed technical influence (Huenteler et al. 2016; Verspagen 2007). Both the weights and structure of citation networks offer meaningful information about knowledge flows (Fontana et al. 2008; Funk and Owen-Smith 2017; Park, Leahey, and Funk 2023). By contrast, patent references to classes form a two-mode network (Fontana et al. 2008). Within technology class-level networks, when a class is cited by and combined with diverse other classes, the technology it represents has been deployed in a wide variety of products and processes, suggesting a higher level of "generality" (Feldman and Yoon 2012; Petralia 2020; Sterzi 2013; Trajtenberg et al. 1997). At the extreme, such technologies become "general purpose" technologies like electricity, the internet, and most recently artificial intelligence techniques that have application across the productive economy (Crafts 2021; Petralia 2020). By contrast, when a technology class is cited less, but cites other classes more more, this suggests a higher level of specificity, such as an implementation or application of an upstream technological element (Fontana et al. 2008). These studies demonstrate that knowledge networks have a characteristic hierarchical structure, stemming from the history of technological evolution.

A discrete network representation may not facilitate the intuitive discovery or analysis of hierarchical structure in complex networks. In dense networks, for example, even the most remote nodes retain edges between them, obfuscating the knowledge tree from traditional network measures and analyst's awareness. "Trimming", or deleting less important nodes or infrequent edges results in a loss of information that distorts the representation structure. The embedding method serves as a solution to this challenge. By transforming each node into a vector output, it utilizes all information to calculate the relevant density of nodes and edges that immediately reveals hierarchical structure if present. Here we turn to the Poincaré embedding algorithm based on hyperbolic geometry, which is specifically designed for modeling discrete hierarchical systems without distortion.

In the top three panels of Figure 3.2, we contrast a network representation with hyperbolic and Euclidean embeddings. In the 2D Poincaré disk, distances grow exponentially due to a constant negative curvature revealing hierarchy as a fundamental property of the space. In this way, the hyperbolic projection maps the trinary tree network without distortion, lucidly revealing its hierarchical structure. The hierarchical structure naturally facilitates measurements for depth (bottom panel 1) and breadth (bottom panel 2), as functions of r and θ in hyperbolic space. Details on measurement construction are provided in Section 3.3. Different combinations of depth and breadth generate four types of inventions: deep broad (blue), deep narrow (green), shallow broad (red), and shallow narrow in the bottom panel 3 of Figure 3.2. By contrast, the network is distorted when projected to a comparable 2D Euclidean space where lower-level nodes in the tree violate their proximity to “parent” nodes in levels above.

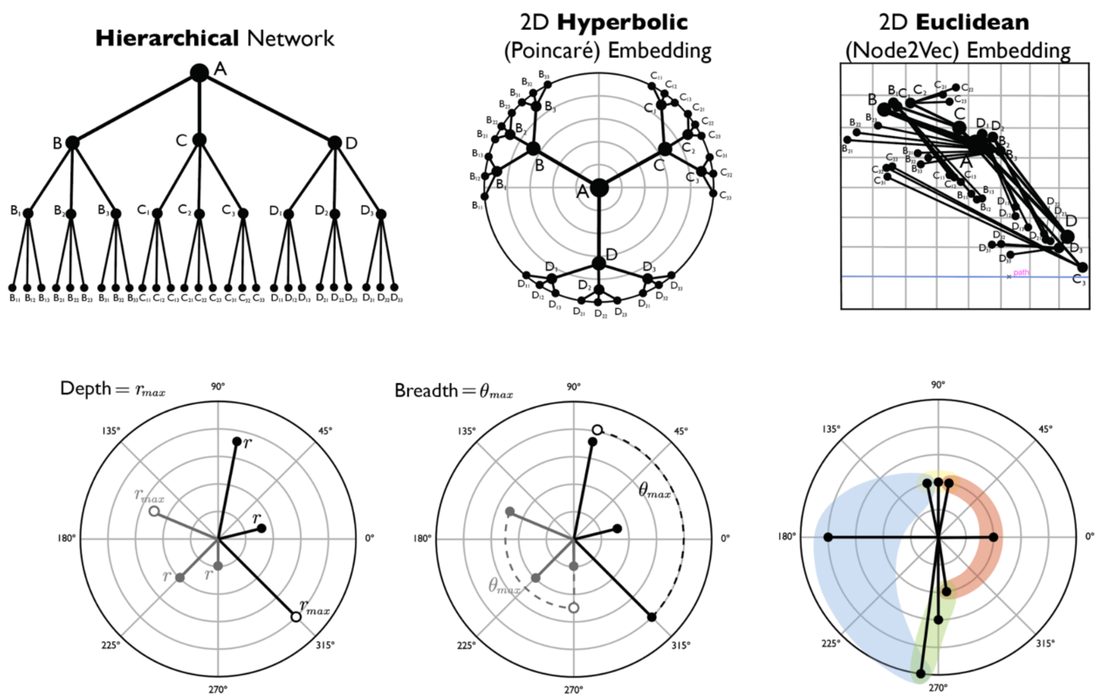

Figure 3.2: Hierarchical depth and breadth on a poincaré disk

In a 2-D Poincaré embedding disk, nodes possess two dimensions: the “generality” dimension pointing from most to least centered components, and the “similarity” dimension along the periphery where related components are locally clustered, thereby exhibiting a high level of homophily. This is illustrated in the lower 3 panels of Figure 3.2. The left panel illustrates differences in the radius (r) of the Poincaré disk at which embedded technology classes lie. Higher r suggests a "deeper" or more specific technology class and its associated patents. The middle panel illustrates differences in the angle (θ) across the Poincaré disk that assess distance across the hierarchy of technology anchoring classes with a patent. The right panel illustrates examples of four technology patents that are narrow, wide, deep, or narrow as a function of their varying combination of class r and θ. This figure highlights the independence of depth and breadth within this representation, and its ability to facilitate analysis of their complex combination within patents, and within firms. Utilizing this embedding, we model the recombination of technological classes, and re-examine the classical hypothesis of deep versus wide recombination.

3.3.3 Constructing Measures for Depth and Breadth for Patents

We investigate the implications of deep and wide recombination based on 4,992,915 granted patents with identifiable prior art categories in the Poincaré disk. For our analysis, we utilize patents granted between 1976 and 2017 to facilitate the calculation of outcome variables such as five-year forward citations. We first construct weighted, directed networks between technological categories based on their citation within patents for each 3-year period (with 1976-1978 as the first period, and 2015-2017 as the last), then embed them into Poincaré disks to map the hierarchy of technical knowledge. Then, we deploy measures for recombination depth (r) and breadth (θ), accounting for the hierarchy of evolving technologies. Next, we incorporate these measures into regression analyses, estimating the effects of deep and wide recombination on patent short and long term impacts. Lastly, we examine potential mechanisms by projecting subsequent work onto the Poincaré disk and tracing the trajectory of knowledge diffusion. The citation network and Poincaré embedding space are built using Python 3, and regression analyses are conducted with Stata 15.

3.3.3.1 Citation Networks and Poincaré Disks for U.S. Patents, 1976-2017

The majority of existing literature on technological evolution and recombination employs the United States Patent Classification ("USPC") system (Fleming and Sorenson 2004; Petralia 2020) or the International Patent Classification ("IPC") system (Corrocher, Malerba, and Montobbio 2007; Fontana et al. 2008) as primary classification systems. In 2013, USPTO introduced the Cooperative Patent Classification ("CPC") system and gradually replaced the USPC. Based on the IPC, the system used by the European patent system, the CPC system boasts greater granularity than USPC and is favored by the USPTO for future use in its efforts to defend U.S. intellectual property in alignment with Europe. Empirical evidence reveals high consistency between measures built on USPC and CPC classifications (Lobo and Strumsky 2019). Here we employ the CPC schema in our first models and consider each level-4 classification as an analytical unit (e.g., A01B33), resulting in approximately 20,000 groups as nodes in the citation network. In a second analysis based on twin-patents filed and granted by both EPO and USPTO, we use USPC codes for US patents and IPC codes for European patents, based on independent embedding spaces. We provide details in the data section below.

Based on the CPC indexes, we construct weighted, directed knowledge diffusion networks between technological categories for 14 periods: 1976-1978, 1979-1981, 1982-1984, and so on, with the last period 2015-2017. The networks are directed, because each diffusion tie originates from cited patents and ends with citing ones. They are weighted, as multiple diffusions can occur between two technological categories within the same period. We also eliminate self-loops, because they are irrelevant to the positioning of classes in relation to one another and do distort the structure of citation networks. In this network, the most frequently and diversely cited categories occupy the center of the hierarchy, indicating a high potential for downstream application use. We then embed this network into a 2-D disk using the Poincaré Embedding algorithm with training parameter settings that follow (Linzhuo et al. 2020)[10].

The embedding space parsimoniously represents the hierarchical structure of technical knowledge. Within this disk, each node is represented by polar coordinates (r, θ). If a technological category receives diversified citations from various fields—characterizing it as an early defined, general-purpose component supporting numerous applications (Feldman and Yoon 2012; Petralia 2020; Sterzi 2013; Trajtenberg et al. 1997), the node approaches the center of the Poincaré disk with a small r. Conversely, a large r signifies that the category is closer to the margin, predominantly used in a focused cluster of downstream applications. A smaller difference in θ with respect to follow-up patents indicates a narrow set of applications, while larger θ among follow-up work indicates "wide-angle" diversity in applications across domains.

We show an example in Figure 3.3 (left panel), which presents the results of the Poincaré embedding of patented technologies from 2015 to 2017. We present embeddings for all 14 periods in the Appendix A, aligning them with the Procrustes procedure (Hamilton et al. 2016) for visualization. Coloring of the nodes is based on the "section symbol"—the highest-level category according to the CPC system.

[10] Batchsize=30, learning rate=0.2, the model is evaluated every 10 epochs. Given the asymmetrical nature of the citation network, we set the parameter symmetrize to be False. We also set the training epochs to be 1000 to guarantee sufficient training, of which the first 20 were "burn-in" epochs.

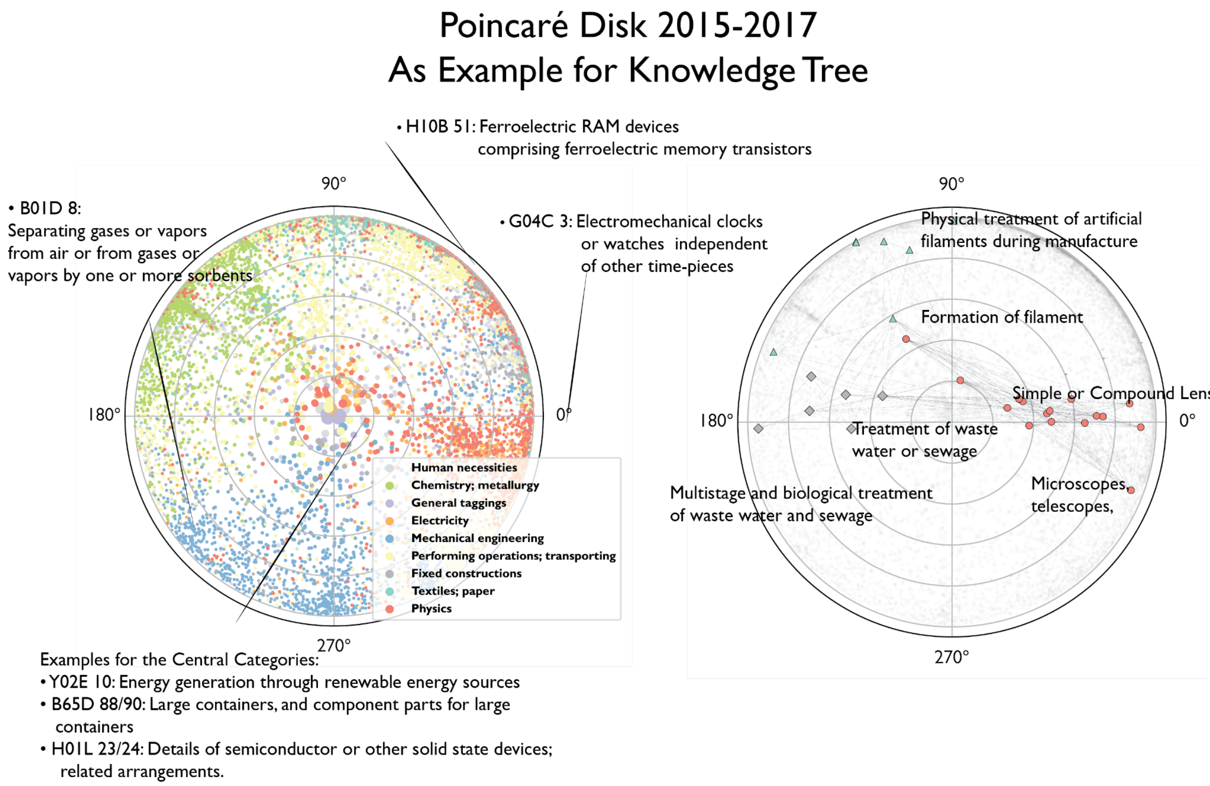

Figure 3.3: Poincaré disks for embedding technological categories

Results indicate that physics (G) and electricity (H) are the most prevalent foundations for other technologies, and they also form an important cluster in downstream applications. Beyond these, chemistry/metallurgy (C), mechanical engineering (F) and performing operations/transportation (B) constitute three other major clusters for downstream application. Performing operations/transporting (B) are dispersed throughout the space, acting as an adhesive for other technological components. The right panel of Figure 3.3 illustrates three examples for technological tree branches, with subclass families C02F (black), D01D (blue), and G02B (red). Each point represents a low-level technological group, and groups within each subclass scatter across the disk in cones, representing the evolutionary history through which general-purpose components have combined into increasingly complex applications. For example, within C02F (treatment of water, waste water, sewage or sludge), the general group C02F1 for water treatment expands into specific methods like multistage treatment (C02F9) and biological treatment (C02F3). Within D01D (methods or apparatus in the manufacture of filaments, threads, fibers), formation of filaments (D01D5) is the basis for treatment of filaments-forming (D01D1) and physical treatment of filaments during manufacturing (D01D10). Within G02B (optical elements), simple and compound lenses (G02B3) make the basis for microscopes (G02B21) and telescopes (G02B23).

Statistical validations of this method can be found in the Appendix A. The results of our robustness checks suggest that our methods efficiently and precisely capture the distance between technological categories.

3.3.3.2 Constructing Measures for Deep and Wide Recombination

Following prior research on scientific and technological innovations, we consider the technical categories of prior art (cited patents) as representing the space of possible combinations for each focal patent (Kaplan and Vakili 2015; Uzzi et al. 2013; Verhoeven et al. 2016). First we identify all combined categories for each patent and determine their positions on the Poincaré disk. Then we calculate depth and breadth measures for each focal patent based on the relevant categories' positions and their geometrical relationships in the embedding space.

In a 2-D Poincaré disk, each node is represented by a polar coordinate (r, θ). As r measures vertical depth in the knowledge tree and θ represents diversity, we construct depth and breadth measures to capture the maximum depth and breadth of the invention based on hyperbolic embeddings as follows:

$Depth = 90th\ Percentile(r_1, r_2, r_3, ...r_n)$ where $r_n = |r_n - r_{origin}|$

$Breadth = 90th\ Percentile(\theta_1, \theta_2, \theta_3, ...\theta_n)$ where

$$\theta_n = |\theta_i - \theta_j|, \forall i = i_1 ... i_k, \forall j = j_1 ... j_k, i \neq j$$

Depth measures the specificity of knowledge elements within the knowledge tree, while breadth evaluates the span of knowledge elements across domains. In contrast to previous studies that calculate recombination depth and width based on existing classification systems, our measures consider search styles as a joint distribution in a polar 2-D space. Consequently, attaining high depth and high breadth simultaneously is not only conceptually, but also empirically feasible. The Pearson correlation coefficient between depth and breadth for U.S. patent data is approximately 0.05, suggesting approximate conceptual independence. Their correlation with prior art categories are 0.09 and 0.20, respectively. Therefore, our measures have captured information on knowledge structure beyond technological diversification.

3.3.3 Two Examples

To explain our measures intuitively, we illustrate two examples for recombination with the Poincare disk. The left one, patent US8837824, was filed by Microsoft Corporation for a method of choosing the best encoding codec for images, which reduces network bandwidth consumption in communication. Among the prior art technology categories underlying this patent, H04N/L/W (pictorial communication, transmission of digital information, communication networks) and G06F/Q (digital data processing, information technology for managerial and administrative purposes) are both categories frequently related to digital information. G06F/Q lies upstream of H04N/L/W, with methods to increase the efficiency and quality of communication (H04N/L/W) improving overall data processing (G06F/Q). This patent exemplifies deep but narrow recombination, as all categories fall along the same branch of the technology tree and draw upon technology components well-used within the specific domains of digital image analysis.

In contrast, patent US9513224 (assigned to Labrador Diagnostics LLC) provides an image analysis and measurement method for biological samples, which uses specialized sample holders and systems for examining samples in different settings. This patent (1) aims to improve image data processing (G06T); and (2) targets laboratory settings, focusing on the chemical and physical properties of materials (G01N and B01L). This exemplifies wide but shallow recombination with elements spanning distinct branches of the technology tree, unexpectedly combining general designs across candidate technologies, rather than drawing upon nuanced technical implementations that lie deep in the thickets of a specific application.

Figure 3.4 shows the location of the two patents and the prior art categories they combine. It illustrates the rationale and inspiration for our measures.

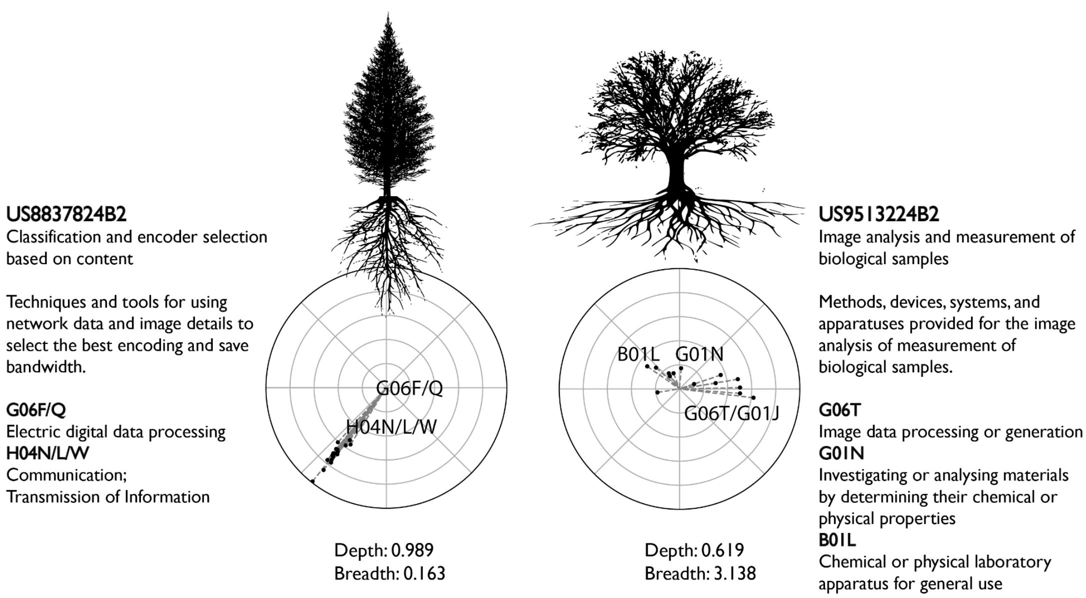

Figure 3.4: Two examples for deep and broad recombinations

3.3.4 Comparison to Canonical Measures of Recombination Depth and Breadth

We next correlate our measure with classic measures for depth and breadth in previous literature. The most widely applied measures include Herfindahl diversity (Hall et al. 2001; Kaplan and Vakili 2015), technological distances (Trajtenberg et al. 1997), combination familiarity (Fleming 2001) and a dummy for whether the combination is new to the world (Verhoeven et al. 2016). Details of these traditional measures are reported in the appendix C, and correlation results are reported in table 3.1.

Table 3.1: Correlation of poincaré depth and breadth with previous measures.

	Herfindahl Diversity	Technological Distances	Original Combinations (dummy)	Combination Familiarity
Depth (max r)	0.376	0.272	0.104	-0.242
Breadth (max θ)	0.415	0.410	0.096	-0.160

Note:N=4,992,915, all p-values<0.001.

Results show that both our measures for depth and breadth exhibit positive correlations with conventional operationalizations. A significant proportion of recombinatory breadth identified in prior research stems from the horizontal (cross-domain) dimension of Poincaré embedding. This suggests that when inventors combine multiple technological categories, they often bridge several branches of the knowledge hierarchy. Paradoxically, combination originality displays a larger correlation with depth than breadth (r_{depth}=0.104 > $r_{breadth}$=0.096). This demonstrates that most new combinations arise from exploitative search within a single knowledge domain, confirming patterns discovered by Kaplan and Vakili (2015).

3.4. Linking Depth and Breadth with Short versus Long Term Impact

3.4.1 Dependent and Independent Variables

Dependent Variables. We draw from previous literature to use forward patent citations, the most widely accepted measures for patent impact. This refers to the number of citations a focal patent receives in subsequent years, which correlates with patent licensing revenue (Sampat and Ziedonis 2005). More citations suggest the technology underlying the patent is attributed significant influence and is frequently reused, revised, or expanded upon in subsequent inventive endeavors to which its knowledge diffuses (Fontana et al. 2008). Approximately half of a patent's citations are added by patent inventors, the others coming from patent examiners through a process that critically assesses the novelty of the invention.

Due to the right-censored nature of citation growth, technological impact is measured by the number of citations received within specific period after the focal patent is granted, often 3-7 years in empirical studies (Fleming 2001; Kaplan and Vakili 2015; Verhoeven et al. 2016). Here we use 5 years as the time frame for short-term citations. Nevertheless, accumulating a high number of short-term citations may not ensure long-term success. Collective invention mechanisms—like preferential attachment and the incorporation of ideas into later work—exert influence over the long term (Wang, Song, and Barabási 2013). To account for differences between short and long-term impact, we constructed four citation variables: 0-5 year, 5-10 year, 10-15 year, and 15-20 year forward citations. These variables trace impacts ranging from the most immediate to the most enduring. In our analysis, we first use patents before 2017 as the sample to estimate depth and breadth impact on 0-5 year citations. Then, we include only patents granted before 2002, allowing 20 years to estimate the change of depth and breadth effects across time.

Independent Variables. We employ depth and breadth measures based on hyperbolic embedding space as independent variables, as operationalized in Section 3.3.3.2. To statistically account for factors pertinent to a patent's short and long-term impact, we include control variables drawn from prior research on technological innovation that allow us to examine the marginal contribution of depth and breadth. Controls include (1) familiarity of components and combinations from prior inventions (Fleming 2001); (2) number of technological categories credited within both focal patents and prior art (Verhoeven et al. 2016); (3) inventor context, including indicator variables for organizational assignee, team collaboration, and inventor experience; and (4) bibliometric indicators including non-patent references, number of claims, and family size (number of patents within a collection covering the same or similar technical content) of the focal patent. As the number of prior art citations is almost perfectly correlated with the total number of technological categories represented in prior art ($r=0.93$), we do not include the former in the regression. We also control for year fixed effects and the average domain location θ of all technological categories of a focal patent, to account for heterogeneity of technological domains.

3.4.2 Data and Model Specification

In our first analysis, we estimate the impact of depth and breadth on citations with all US patents granted between 1976-2017 whose prior categories are identifiable. In total, this yields a sample of 4,992,915 patents. Citations are characterized by over-dispersed non-negative integer values. In accordance with Fleming (2001) and Verhoeven et al. (2016), we utilize a generalized negative binomial regression to estimate both the average and dispersion of citations, and use Poisson Maximum Likelihood Estimator ("QMLE") as a robustness test in the Appendix B. The first set of models are conducted at the patent level:

$$log(Citations_i) = \beta_1 \times Depth_i + \beta_2 \times Breadth_i + \beta_3 \times Controls_i + \epsilon_i,$$

where $Depth_i$ is the depth measure for the focal patent i, whereas $Breadth_i$ captures patent breadth. Both categories are calculated from the CPC classification system, either issued at granting or in retrospect. This model estimates the combinatorial effects across the broadest samples available. Here, $Citations_i$ includes $Citations_{i,0-5}$, $Citations_{i,5-10}$, $Citations_{i,10-15}$ and $Citations_{i,15-20}$, allowing us to estimate the direct effects of depth and breadth on short- and long-term impacts, respectively. When we compare effects across time, we only include patents granted before 2002 to avoid right censorship of the citation data.

Direct effects of depth and breadth estimated in this simple model may unintentionally reflect heterogeneity in the technology underlying the patents themselves, reflecting their quality or technological complexity. To rule out these confounders, we adopted a second, twin-patent design (Bikard 2020; Kovács, Carnabuci, and Wezel 2021). For this analysis, we only retained patents submitted to both the USPTO and EPO, and granted at both patent offices. These patents have identical quality and patent features, but are granted and evaluated within different knowledge systems with different histories, conditioning distinct expectations. Therefore, while the quality of twin inventions remains the same, the search paths of inventors are perceived differently in the U.S. and Europe due to distinct histories of technological evolution encoded within their distinct collective knowledge systems. If our argument holds that collective understandings play a crucial role, we expect these effects to persist even after controlling for the intrinsic quality of an invention, as citations are shaped by the subjective decisions of other inventors.

We followed previous literature to filter valid patents. We identified one-to-one US and European patent matches, and then excluded patents with different titles or one-to-many or many-to-one matches. The resulting sample includes 84,916 patent pairs.

For patents granted by the USPTO, citations may be added both by applicants and examiners, while patents granted by EPO only receive citations added by examiners. To make the pairwise samples comparable, we only count citations added by examiners for both USPTO and EPO samples. Distinct from our first analysis, here we calculated depth and breadth for the US patents based on the USPC classification system and for EPO patents based on the IPC system in order to intentionally keep different operationalizations between the twin samples. European categories are embedded in separate Poincaré disks within 3-year windows. We restricted our sample to patents granted between 2001-2012, as 2001 is the first year when data on examiner-added citations is available for the USPTO, and 2013 is the year when USPTO changed the standardized classification system from USPC to CPC, which shifts the comparison baseline.

An invention's citation counts at the USPTO can be specified as:

$$log(Citations_{i,US}) = \beta_1 \times Depth_{i,US} + \beta_2 \times Breadth_{i,US} + \beta_3 \times Controls_{i,US} + Quality_i + \epsilon_{i,US}$$

while the impact of its patent twin at the EPO can be expressed as:

$$log(Citations_{i,EU}) = \gamma_1 \times Depth_{i,EU} + \gamma_2 \times Breadth_{i,EU} + \gamma_3 \times Controls_{i,EU} + Quality_i + \epsilon_{i,EU}$$

We substitute the second equation into the first equation. The resulting equation has the following form:

$$log(Citations_{i,US}) = \alpha_1 \times Depth_{i,US} + \alpha_2 \times Breadth_{i,US} + \alpha_3 \times Controls_{i,US}$$
$$+\alpha_4 \times log(Citations_{i,EU}) - \alpha_5 \times Depth_{i,EU} - \alpha_6 \times Breadth_{i,EU} - \alpha_7 \times Controls_{i,EU}$$
$$+\epsilon_{i,US} - \epsilon_{i,EU}$$

We estimate this model on the pairwise samples to get the depth and breadth effects when controlling for patent quality and other individual-level heterogeneities as a robustness test.

Finally, we conduct an analysis to identify mechanisms by which depth and breadth differentially associate with short and long-term impact. For this part of the analysis, we perform a mediation analysis based on regressions[11] on the full USPTO sample (N=4,992,915). We build two mediator variables from our embedding spaces: (1) local concentration, calculated as the citations an invention receives nearby in its domain (i.e., citations divided by the standard deviation of $\Delta\theta$ between focal and citing patents), and (2) diffusion distance, computed as the average dispersion of follow-up works across the inventive space (i.e., the standard deviation of $\Delta\theta$ between focal and citing patents). Details of these mediator measurements are explained in the mechanism analysis section. We estimate GSEM on citations received in different periods after the focal patent is granted, including 0-5 year citations, 5-10 year citations, 10-15 year citations, and 15-20 year citations, respectively. Based on these models, we track shifts in effects mediating patent depth and breadth with citation outcomes.

In the final section, we scale our analyses of technological search to the organizational level. We first estimate the effects of focal patents' depth and breadth on the depth and breadth of follow-up work, both within organization and at the patent population level. Second, we aggregate depth and breadth to the inventor, organization and domain (5-digit CPC technology category) levels, and test depth and breadth effects on citation impact at those levels. We compare the coefficients for depth and breadth on citation impact across levels, to reveal how

[11] These models, when considered together, can be interpreted as a generalized structural equation model (GSEM). Given the complexity of effect paths, the estimation and effect sizes remain unchanged compared to estimating these regressions separately. For methodological details, see the Stata manual at https://www.stata.com/manuals/semgsem.pdf.

collective strategic efforts enable collective technological progress. For these models, we use negative binomial regressions with cluster-level fixed effects, specified as:

$$AverageCitations_{i,t} = \beta_1 \times Depth_{i,t} + \beta_2 \times Breadth_{i,t} + \beta_3 \times Controls_{i,t} + \alpha_i + Year_t + \epsilon_{i,t}$$

where $AverageCitations_{i,t}$ is the average 5-year citations received by patents invented by inventor *i* (or filed by organization *i* or in domain *i*) at time *t*, α_i represents inventor/organization/domain level fixed effects, and $Year_t$ captures year fixed effects.

3.4.3 Results for the Negative Binomial Regressions

Table 3.2 reports basic descriptive statistics for all variables in different samples, while table 3.3 presents the correlation matrix. Table 3.4 reports results from our models.

Our negative binomial models indicate that, overall, depth positively predicts short-term technological impact, while breadth exhibits a contrasting effect. A one-unit increase in depth (from theoretical minimum to maximum) results in roughly a 0.196 increase in the five-year forward citation count. At first glance, this might seem unimpressive. However, considering the highly skewed nature of the forward-citation distribution, with 1 as the median, a 0.196 increase represents an approximately 20% rise for more than half of the patents in the sample. In contrast, a one-unit increase in recombination breadth corresponds to a decrease of 0.043 in citations. Given that the theoretical maximum change in breadth is about 3.14 radians (180°) in the Poincaré disk, the maximum increase in recombination breadth would lead to an approximate 0.135 decrease in citations.

Table 3.2: Descriptive statistics.

	Full Samples (N=4,992,915)				**Twin Patent: US Sample (N=84,916)**				**Twin Patent: European Sample (N=84,916)**			
	MEAN	**SD**	**MIN**	**MAX**	**MEAN**	**SD**	**MIN**	**MAX**	**MEAN**	**SD**	**MIN**	**MAX**
0-5 Year Citations	3.443	10.156	0	1142	1.087	1.789	0	45	0.521	1.592	0	45
5-10 Year Citations	5.320	19.205	0	2555	1.144	2.059	0	125	0.888	3.025	0	365
10-15 Year Citations	4.115	18.029	0	2542	0.639	1.476	0	82	0.709	4.011	0	379
15-20 Year Citations	2.926	16.013	0	2468	0.242	0.935	0	47	0.357	2.880	0	319
Depth	0.817	0.160	.034	.999	0.680	0.172	0	0.999	0.671	0.245	0	1
Breadth	1.052	0.995	0.000	3.141	0.923	0.908	0	3.140	1.004	0.984	0	3.14
Horizontal Location (mean theta)	0.201	1.551	-3.139	3.138	-0.072	1.653	-3.313	3.129	0.211	1.549	-3.127	3.134
Ln (Component Familiarity)	11.426	2.546	0	17.494	7.850	1.483	0	11.614	10.389	1.376	4.359	14.299
Ln (Combination Familiarity)	4.279	3.358	0	13.496	2.723	2.223	0	8.782	5.852	2.661	0	12.174
Number of Tech Classes	5.372	4.352	0	17	5.345	4.040	0	16	4.586	3.096	1	12
Number of Prior Art Classes	56.605	75.070	2	295	47.200	50.063	0	196	206.335	159.531	9	626
Assigned to Organizations	.897	0.304	0	1	0.973	0.162	0	1	0.973	0.162	0	1
Invented by Team	.628	0.483	0	1	0.709	0.454	0	1	0.709	0.454	0	1
Ln (Inventor Experience)	2.662	1.854	0	12.219	1.975	1.179	0	7.510	1.850	1.295	0.693	6.685
Non-Patent References	3.081	5.999	0	23	1.997	3.211	0	12	0.481	0.911	0	3
Number of Claims	15.272	8.907	1	36	15.724	8.982	1	37	4.43	6.461	0	20
Family Size	.439	0.611	0	2	1	0	1	1	1	0	1	1

Note: N=4,992,915 for all patents granted by USPTO between 1976-2017, and N=84,916 for patents granted at both USPTO and EPO between 2001-2012. Patents with < 2 prior art categories among the patents they cite are excluded as we cannot construct meaningful independent variables.

Table 3.3: Correlation matrix of major variables (full sample, N=4,992,915).

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
(1) 5-Year Citations	1.000													
(2) Depth	0.017	1												
	[0.000]													
(3) Breadth	0.034	-0.160	1											
	[0.000]	[0.000]												
(4) Horizontal Location (mean theta)	0.018	-0.060	0.010	1										
	[0.000]	[0.000]	[0.000]											
(5) Ln (Component Familiarity)	0.059	-0.421	0.003	0.037	1									
	[0.000]	[0.000]	[0.000]	[0.000]										
(6) Ln (Combination Familiarity)	0.013	-0.223	-0.122	0.016	0.314	1								
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]									
(7) Number of Tech Classes	0.086	-0.118	0.122	0.029	0.337	-0.371	1							
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]								
(8) Number of Prior Art Classes	0.169	0.043	0.274	0.016	0.254	-0.063	0.382	1						
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]							
(9) Assigned to Organizations	0.025	-0.094	-0.049	0.017	0.228	0.079	0.094	0.084	1					
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]						
(10) Invented by Team	0.035	-0.067	0.004	0.002	0.174	0.008	0.128	0.113	0.291	1				
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]					
(11) Ln (Inventor Experience)	0.040	-0.139	0.036	0.005	0.305	0.088	0.201	0.223	0.224	0.178	1			
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]				
(12) Non-Patent References	0.113	0.050	0.014	0.007	0.235	-0.030	0.235	0.456	0.112	0.149	0.165	1		
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]			
(13) Number of Claims	0.098	-0.001	0.040	-0.011	0.154	0.039	0.097	0.214	0.093	0.098	0.089	0.183	1	
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]		
(14) Family Size	-0.071	-0.086	-0.042	0.014	-0.000	0.008	0.002	-0.184	0.158	0.056	0.081	-0.145	-0.166	1
	[0.000]	[0.000]	[0.000]	[0.000]	[0.544]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	

Note: *p*-values are between square brackets. All tests are two-tailed.

Table 3.4: Effects of recombination depth vs. breadth on patent short- and long-term impact.

	Model with All Patents: 1976-2017	Model with Early Patents: 1976-2002			
	Model 1: Predicting 0-5 Year Citation	**Model 2 Predicting 0-5 Year Citation**	**Model 3 Predicting 5-10 Year Citation**	**Model 4 Predicting 10-15 Year Citation**	**Model 5 Predicting 15-20 Year Citation**
Main Models					
Depth	0.196***	0.095***	-0.000	-0.164***	-0.434***
	[0.000]	[0.000]	[0.950]	[0.000]	[0.000]
	(0.005)	(0.006)	(0.007)	(0.008)	(0.009)
Breadth	-0.043***	-0.036***	-0.019***	-0.001	0.037***
	[0.000]	[0.000]	[0.000]	[0.549]	[0.000]
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Horizontal Location (mean theta)	0.035***	0.077***	0.073***	0.069***	0.087***
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)
Ln (Component Familiarity)	0.033***	0.029***	0.033***	0.034***	0.023***
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)
Ln (Combination Familiarity)	0.030***	0.035***	0.025***	0.019***	0.015***
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)
Number of Tech Classes	0.028***	0.032***	0.040***	0.046***	0.050***
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Number of Prior Art Classes	0.003***	0.003***	0.004***	0.004***	0.005***
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Assigned to Organizations	0.075***	0.115***	0.072***	-0.011***	-0.085***
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
	(0.002)	(0.002)	(0.003)	(0.003)	(0.003)
Invented by Team	0.076***	0.086***	0.094***	0.086***	0.084***
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)
Ln (Inventor Experience)	0.010***	-0.000	-0.023***	-0.039***	-0.047***
	[0.000]	[0.880]	[0.000]	[0.000]	[0.000]
	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)
Non Patent References	0.009***	0.007***	0.017***	0.019***	0.017***

	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Number of Claims	0.018***	0.018***	0.021***	0.023***	0.024***
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Family Size	-0.169***	-0.101***	-0.203***	-0.268***	-0.308***
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)
Year Fixed Effects	YES	YES	YES	YES	YES
Constant	0.049	0.099	0.036	0.206***	0.439***
	[0.367]	[0.068]	[0.525]	[0.001]	[0.000]
	(0.054)	(0.054)	(0.057)	(0.061)	(0.060)
Dispersion Model	***With Same Independent Variables, Reported in Appendix G***				
Observations	4992915	2010031	2010031	2010031	2010031

Note: *p*-values are between square brackets. Robust standard errors are in parentheses. All tests are two-tailed.

Models 2-5 present regressions based on samples from before 2002, allowing 20 years (2002-2022) to observe short- and long-term citations. For this sample, we constructed four dependent variables for citations from different periods and estimated the effects of depth and breadth over time: 0-5, 5-10, 10-15, and 15-20 year citations. While the 0-5 year citations model (model 2) aligns with the general model (model 1) results, effects for both depth and breadth are reversed for long-term citations (models 3-5). Over time, the depth effect transitions from positive to negative (from β=0.095 to -0.434, both $p<10^{-5}$), whereas the breadth effect turns from negative to positive (from β=-0.036 to 0.037, both $p<10^{-5}$). The shift in coefficients for depth and breadth, as citation lags grows, are both monotonous, with a modest acceleration from model 2 to model 5. We visualize the temporal change of depth and breadth effects in Figure 3.5.

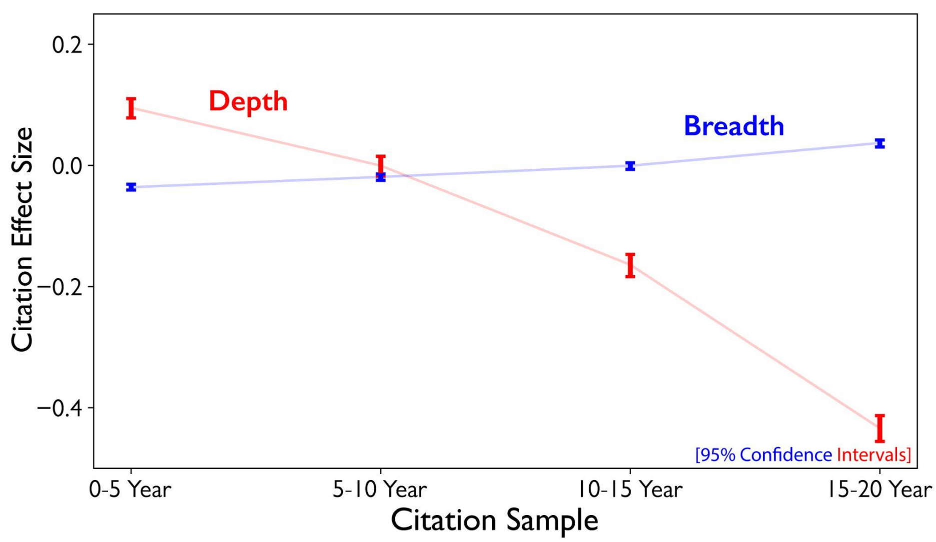

Figure 3.5: Depth and breadth effects on short-term and long-term patent impact

The dispersion model (Appendix G, Table A11) reports the effects on citation dispersion. In general, both depth and breadth reduce the standard deviation for citation (β_{depth} = -0.117, $\beta_{breadth}$= -0.031, as seen in model 1). Nevertheless, the two variables manifest different cross-time variation: depth boosts variance first and then reduces monotonously as it turns negative; breadth remains stable and negative throughout. We have also estimated models 1-5 on only within-organization citations, and other sub-samples (patents before 2012, and patents before 2007) for robustness, manifesting a similar pattern of results. These results are reported in the Appendix B.

We estimate the same set of models on US-EU twin patents (*N*=84,916) as a stricter specification. By substituting the European citation equation into the original model, we strictly control for patent quality, and so estimate the perception of depth and breadth effects independent of the inventive process. We report the results of this model in Table 3.5. If our argument is robust, we expect (1) the depth and breadth effects to persist with the same signs, as in the previous models from Table 3.4; (2) US and European citation counts to be positively correlated in each of the target periods; and (3) within the US citation model, the European independent variables (e.g., European depth and breadth for the same patent) will have opposite signs as the US variables, suggesting that they manifest the same effects, but flipped by our model specification. We adopt the standards set by (Kovács et al. 2021). The results, as reported in Table 3.5, meet these expectations, and show the robustness of our model from Table 3.4 when strictly controlling for the quality of the underlying invention.

Table 3.5: Patent-twin tests: five-year forward citations added by examiners at the USPTO while controlling for EPO measures.

	Model 6	Model 7	Model 8	Model 9
	Model: Poisson QMLE		Model: Negative Binomial	
USPTO Depth	0.750***	0.624***	0.548***	0.571***
	[0.000]	[0.000]	[0.000]	[0.000]
	(0.038)	(0.040)	(0.031)	(0.036)
USPTO Breadth	-0.053***	-0.050***	-0.020***	-0.043***
	[0.000]	[0.000]	[0.000]	[0.000]
	(0.007)	(0.007)	(0.005)	(0.006)
European Five-Year Forward Citation		0.079***		0.103***
		[0.000]		[0.000]
		(0.003)		(0.003)
European Depth		-0.075**		-0.063*
		[0.006]		[0.010]
		(0.027)		(0.025)
European Breadth		0.012		0.008
		[0.065]		[0.166]
		(0.006)		(0.006)
US/EU Patent Controls	YES	YES	YES	YES
US/EU Year Fixed Effects	YES	YES	YES	YES
Dispersion Model For Negative Binomial			YES	YES
N	84,956	84,956	84,956	84,956
Log-Likelihood	-137492.823	-131271.191	-118030.06	-116809.3

Note: P-values are between square brackets. Robust standard errors are in parentheses. All tests are two-tailed.

Our models reveal that a patent's depth and breadth in knowledge structure does influence how inventors perceive and understand its significance, and shape their decision of whether to adopt it in their own work. Consistent with category-spanning theory (Zuckerman 1999), our models show a short-term punishment for inventions that span many categories distant from one another. Nevertheless, our work also reveals temporal dynamics less discussed in previous literature: while category spanning inventions appear ambiguous and less recognized in the short-term, they are more likely to be noticed in future, while deep inventions shine first but exhaust their potential rapidly. To explore the processes underlying these patterns, we turn to an analysis of mechanisms.

3.5 Exploring Mechanisms with Diffusion across Embedding Space

Our analysis indicates that search depth correlates with higher forward citations, while breadth exhibits the opposite effect with the direction of these effects reversing further into the future. Why does this occur? The path and pattern of citation growth can be tracked in the Poincaré embedding space. With each technological category and each patent assigned a location, the embedding model provides a window on the underlying mechanisms.

We first compute the position of each patent by averaging the coordinates of its categories. We then repeat this process for follow-up, citing patents. By comparing the location of a focal patent with its citers, we can quantify knowledge diffusion distances and traces. The left panel of Figure 3.6 presents a conceptual depiction of the collective invention process. After the emergence of a focal invention (represented by the dark blue dot), subsequent patents may arise either along the same intellectual vein (light blue dots) or from different domains (yellow dots). We use the absolute value of the difference between the focal patent's θ and a follow-up patent's θ (denoted $|\Delta\theta|$) to measure the diffusion distance between domains.

The right panel of Figure 3.6 depicts the conceptual distribution of this distance: in general, follow-up works tend to cluster around the focal work, and diffuse to more distant areas with diminishing probability. The specific shape of this distribution is influenced by both the unique attributes of the focal patent and the diffusion time. As a focal patent becomes more visible following publication, both the average $\Delta\theta$ and its dispersion tend to increase.

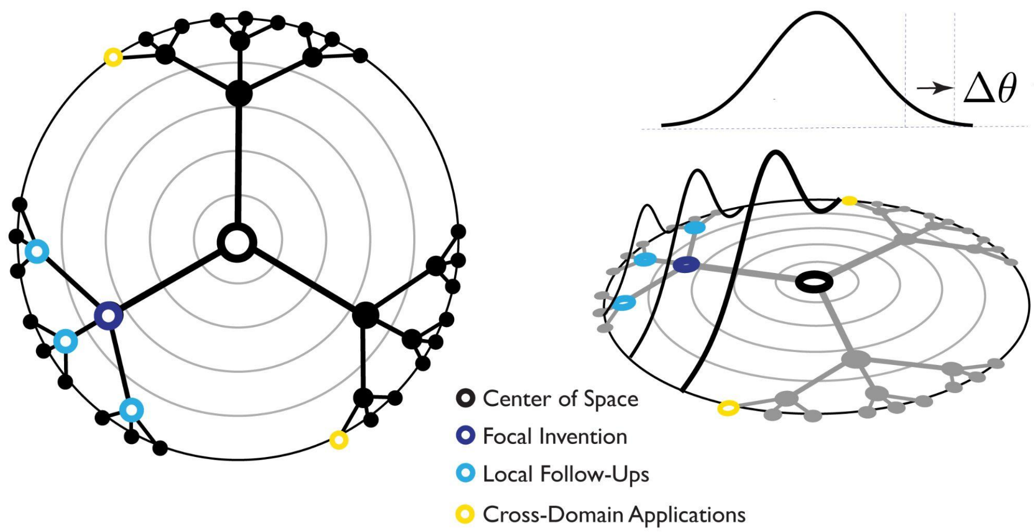

Figure 3.6: Conceptual figure for post-publication diffusion

Two factors determine whether an invention will be highly cited in the future: (1) the citation yield in its local domain, and (2) the scope of its diffusion in other domains. Each of these aspects is underscored by a competing theory regarding depth vs. breadth. The "foundational" view contends that local search helps identify anomalies and facilitates breakthroughs within the same intellectual vein, thereby primarily increasing (1) local impact. In contrast, the "tension" view underscores "lock-in" as a potential consequence of local search, leading to a decrease in (2) diffusion distance. Drawing on these theoretical perspectives, we introduce two variables as mediators. We employ the standard deviation (SD) of diffusion distance as a proxy for (2) diffusion distance. We calculate citation concentration by dividing the period-wise citation count by this SD, generating a proxy for (1) local impact.

We conduct a mediation analysis for all samples (Table 3.6) and assess the cross-time variation of coefficients based on patents granted prior to 2002 (reported in Appendix E). Model results are presented in Model 10, with indirect effects of depth and breadth through the local concentration and dispersion paths visualized in the bottom row of Fig 3.7.

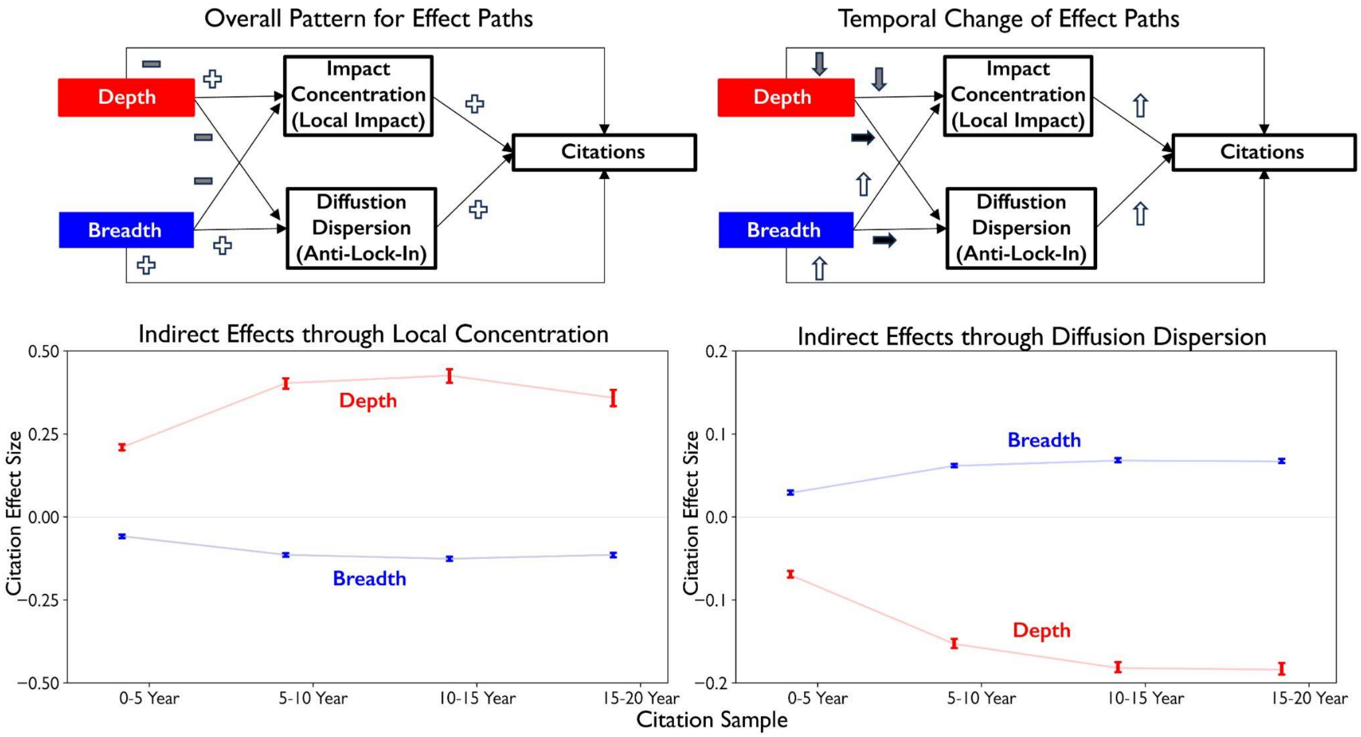

Figure 3.7: Effect paths for generalized structural equation modeling

Table 3.6: Mediation analysis for impact accumulation (five years).

		Model 10	
	Impact Concentration	Diffusion Dispersion	5 Year Forward Citation Count
Concentration			0.165***
			[0.000]
			(0.001)
Dispersion			0.749***
			[0.000]
			(0.004)
Depth	1.214***	-0.103***	-0.076***
	[0.000]	[0.000]	[0.000]
	(0.020)	(0.001)	(0.006)
Breadth	-0.473***	0.059***	0.019***
	[0.000]	[0.000]	[0.000]
	(0.002)	(0.000)	(0.001)
Horizontal Location (mean theta)	0.006***	0.001***	0.018***
	[0.000]	[0.000]	[0.000]
	(0.002)	(0.000)	(0.000)
Ln (Component Familiarity)	0.134***	-0.013***	0.013***
	[0.000]	[0.000]	[0.000]
	(0.001)	(0.000)	(0.000)
Ln (Combination Familiarity)	-0.057***	0.002***	0.018***
	[0.000]	[0.000]	[0.000]
	(0.001)	(0.000)	(0.000)
Number of Tech Classes	0.028***	-0.001***	0.012***
	[0.000]	[0.000]	[0.000]
	(0.001)	(0.000)	(0.000)
Number of Prior Art Classes	0.001***	-0.000***	0.002***
	[0.000]	[0.000]	[0.000]
	(0.000)	(0.000)	(0.000)
Assigned to Organizations	0.165***	-0.008***	0.032***
	[0.000]	[0.000]	[0.000]
	(0.008)	(0.001)	(0.002)
Invented by Team	0.053***	0.001***	0.037***
	[0.000]	[0.000]	[0.000]
	(0.005)	(0.000)	(0.002)
Ln (Inventor Experience)	0.001	-0.002***	0.009***
	[0.422]	[0.000]	[0.000]
	(0.001)	(0.000)	(0.000)
Non Patent References	0.014***	-0.000***	0.007***
	[0.000]	[0.000]	[0.000]
	(0.000)	(0.000)	(0.000)
Number of Claims	0.007***	0.000***	0.007***
	[0.000]	[0.000]	[0.000]
	(0.000)	(0.000)	(0.000)
Family Size	-0.102***	-0.001*	-0.096***
	[0.000]	[0.013]	[0.000]
	(0.004)	(0.000)	(0.001)
Year Fixed Effects	YES	YES	YES
Constant	1.634	0.331	0.609
	[0.000]	[0.000]	[0.000]
	(0.130)	(0.017)	(0.048)

var(Concentration Residual)	10.519 [0.000] (0.060)
var(Dispersion Residual)	0.046 [0.000] (0.000)
cov(Concentration Residual, Dispersion Residual)	-0.179 [0.000] (0.000)
N	2178535

Note: P-values are between square brackets. Robust standard errors are in parentheses. All tests are two-tailed.

Our results in model 10 shows that depth enhances local impact, while breadth expands diffusion range. Although the overall pattern has been predicted by existing theories, our models reveal three novel observations. First, the positive effect of depth on local impact diminishes over time (from $\beta_{depth\_to\_concentration} = 1.592$ for 0-5 year citations, to $\beta_{depth\_to\_concentration} = 1.027$ for 15-20 year citations, monotonously decreasing). Moreover, the negative effect of breadth on local impact diminishes (from $\beta_{breadth\_to\_concentration} = -0.438$ for 0-5 year citations, to $\beta_{breadth\_to\_concentration} = -0.326$ for 15-20 year citations, monotonously increasing). Meanwhile, the effects of depth and breadth on diffusion range display distinct temporal trends. Inventions based on deep technical understanding and local search remain locked-in within a small region of technology, accruing outsized rewards in the short term. Second, both local impact and diffusion range positively influence forward citations ($\beta_{concentration\_to\_citation} = 0.165$, and $\beta_{dispersion\_to\_citation} = 0.749$ in model 10), and their effect sizes both magnify over time. The contribution from diffusion range outpaces that from local impact, however, with $\beta_{concentration\_to\_citation}$ growing from 0.132 for 0-5 year citations to 0.350 for 15-20 year citations, and $\beta_{dispersion\_to\_citation}$ growing from 0.498 for 0-5 year citations to 1.245 for 15-20 year citations), suggesting that as time progresses, cross-domain diffusion assumes greater significance than local impact. Over greater lags, citations stemming from breadth dominate. Lastly, the direct effect wanes for depth but amplifies for breadth over time (β_{depth_direct} decreases from -0.066 to -0.472 while $\beta_{breadth\_direct}$ increases from -0.003 to 0.060). These direct effects likely also trace influences beyond the diffusion process itself. These could include unobserved aspects such as patent quality and the process of obsolescence.

3.6 Depth and Breadth Effects on Collective Invention

Our analysis helps reveal patterns of impact accumulation, but has heretofore focused on features and impact of individual patents. Nevertheless, from corporate and societal standpoints, inventions are not independent products, but weave a collective fabric that conditions future innovation (Shi et al. 2015). To gauge how patent depth and breadth influence the search for technology in subsequent work, we conduct two additional analyses, to further examine how collective contexts have enhanced or hindered the depth and breadth effects.

First, we explore the relationship between depth, breadth, and the knowledge structures of follow-up work. We begin by calculating average depth and breadth of follow-up patents within a 5-year span. We then employ OLS models to estimate the association between depth, breadth, and features of collective search. As visualized in the upper panel of Figure 3.8 and reported in Appendix Table A9-A10, our results reveal two patterns: (1) depth breeds more depth as breadth breeds more breadth. One unit increase of focal depth leads to 0.414 units increase in followup depth (model A25), while one unit increase of focal breadth leads to 0.505 units increase in followup breadth (model A29). Both associations are more pronounced within the same organization ($\beta_{\text{depth\_to\_depth}}$ = 0.515>0.414 in Model A26, and $\beta_{\text{breadth\_to\_breadth}}$ = 0.595>0.505 in Model A30), likely due to diminished barriers for knowledge diffusion. (2) Model A31 (addressing follow-up breadth) surfaces a positive interplay between focal patents' depth and breadth, with larger coefficients for their interaction estimated within the same organization. It shows that depth and breadth enhance and complement one another within organizations. This interaction likely implies that a "deep" invention has greater robustness as a reliable component for future recombination with other technologies ($\beta_{\text{depth x breadth}} = 0.404$ in model A31 and $\beta_{\text{depth x breadth}} = 0.411$ in model A32). This moderating effect has a larger effect size within organizations, suggesting that without crossing organizational boundaries, invention depth more efficiently prepares the focal invention to become incorporated into future work.

These findings suggest that collective inventions are constrained by the distribution of attention. In order to examine this further, we conduct a second additional analysis to replicate our results on different analytical levels, i.e., all natural units of collective innovation in technology development—inventors, organizations (from original assignees), and technological domains (defined as 5-digit CPC categories), from small to large in scale. If collective thinking operates via attention—guiding inventors' focus toward relevant components and speeding the process of novel incorporation—we should see an increasing influence of search depth and breadth on citations as we move from the patent to the inventor, organization, and ultimately domain levels. This is because larger, coherent units possess more excess capacity to exercise strategic intelligence. For example, a science-based company can organize their researchers to follow new technologies to the more complex and advanced technologies to which they lead. While a singular inventor has a fixed budget of time and resources for follow-on development, organizations have more, and technological communities have still more. The lower panel of Figure 3.8 displays this pattern. We first aggregate the patent-level data to inventor, organization and technological domain levels, averaging for control variables, and taking the maximum for depth and breadth. Then, we estimate fixed-effects negative binomial models with fixed effects for inventor / organization / domain, year and career year. From patent level models to domain level models, we observe an apparent growing influence of both depth ($\beta_{\text{patent\_depth}} = 0.200$, $\beta_{\text{inventor\_depth}} = 0.433$, $\beta_{\text{organization\_depth}} = 0.494$, $\beta_{\text{domain\_depth}} = 0.992$, all with $p<10^{-128}$) and breadth

($\beta_{\text{patent\_breadth}} = -0.035$, $\beta_{\text{inventor\_breadth}} = 0.033$, $\beta_{\text{organization\_breadth}} = 0.061$, $\beta_{\text{domain\_breadth}} = 0.059$, all with $p<10^{-186}$) across levels, supporting our hypothesis regarding the mechanism of collective attention.

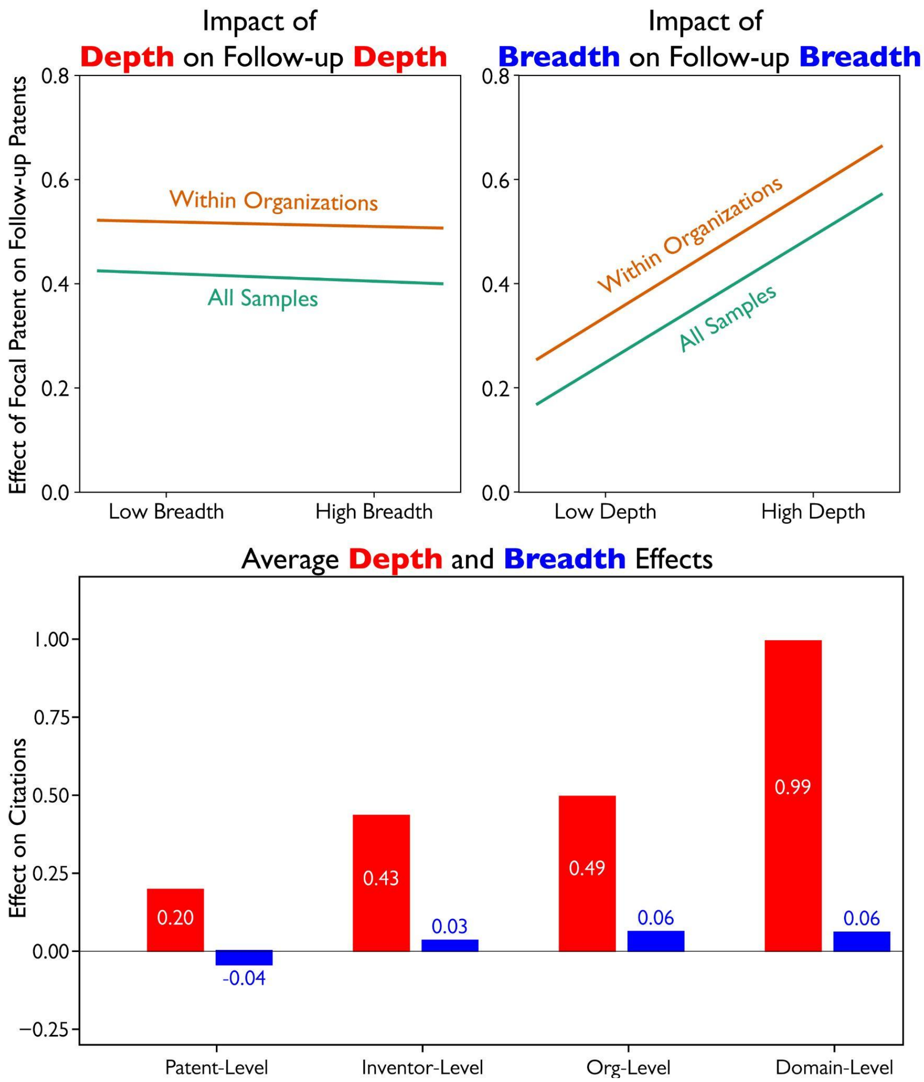

Figure 3.8: Organization and collective effects of depth and breadth

3.7 Discussion And Conclusion

3.7.1 Implications for the relationship between Depth and Breadth in Technological Search

The impact of the combinatorial structure of inventive search on the quality of technological innovations has long been debated, with conflicting evidence for the efficacy of broad versus local search. Kaplan and Vakili (2015) describe this debate in terms of the "foundational" and "tension" views. Our study employs computational representation learning to model the technology landscape, enabling us to develop fine-grained measures of recombination structure and pathways of diffusion. This approach allows us to empirically test mechanisms previously difficult to observe. We first utilize the Poincaré embedding method, based on hyperbolic geometry, to construct a knowledge tree of technological categories within a Poincaré disk. In this representation, the vertical dimension r signifies the continuum from general-purpose technological components to specific applications, the depth of technological combinations, while the horizontal dimension θ represents the breadth or diversity of cross-domain combinations.

We then develop measures for recombination depth and breadth. Contrary to traditional methods that treat these factors as trade-offs, we measure depth as 90th percentile of Δr and breadth as 90th percentile of $\Delta\theta$, as defined geometrically in the embedding space. The former measures the extent to which an inventor or corporation incorporates and revises existing technological infrastructure, and the latter the extent to which an inventor or corporation combines components across domains in unexpected ways. This approach better aligns with the real-world observation that many combinations can be both or neither deep and wide simultaneously.

Our findings offer new insights into the mechanisms of combinatorial innovation, and carry significant theoretical implications. First, we observe a clear and consistent pattern that depth positively correlates with short-term and negatively with long-term impact. Breadth demonstrates the opposite effect. This suggests that the fundamental and tension views are not contradictory, but capture distinct facets of reality. The foundational view is correct that local search garners more local recognition. The tension view is correct that local search leads to "lock-in". Nevertheless, these two outcomes do not occur simultaneously. Local recognition occurs immediately upon patent grant, but the negative "lock-in" emerges and intensifies over time. Technological fields vary widely in their pace of evolution—5 years might be "long-term" within the rapidly advancing AI domain but "short-term" in conventional chemistry. In this way, previous domain-specific studies may have captured only one aspect of the story, even with identical forward citation metrics, potentially yielding divergent conclusions. Our study underscores the role of time in organizational learning, offering a resolution to longstanding debates over depth vs. breadth.

Second, our models on follow-up patent features reveal the unfolding interplay of exploitation and exploration, highlighting their synergistic relationship, especially within companies that can strategically allocate engineering resources to build on prior advances. On the tree of technological knowledge traced by our Poincaré disk, exploitation occurs rapidly when knowledge moves longitudinally. Exploration, by contrast, occurs intermittently as inventors establish connections that cross Poincaré latitudes. Although the immediate advantage of exploitation diminishes in technological influence over time, it primes components for seamless integration into horizontal innovations. This "invisible impact" calls into question the conventional belief that deep recombination is solely associated with exploitation, and suggests the importance for innovative firms to maintain a portfolio of explorations and exploitations that can fuel each other over time.

3.7.2 Methodological Contributions and Other Applications

Over the past decade, computational tools have become increasingly used for building and testing theory in organizations research. As these studies have matured, three paradigms have emerged. The first uses algorithms constructively as "*cranes*". New computational tools enable analysts to structure previously unstructured data for statistical analysis that would otherwise be difficult or impossible for humans to synthesize. A recent example in organization studies uses topic modeling and convolutional neural networks (CNNs) to extract CEO oral communication styles from video interviews (Choudhury et al. 2019). The second approach uses algorithms as "*detectors*". Machine learning enables discovery of unexpected correlations and boundary conditions that cannot be easily surfaced with traditional regression (Shrestha et al. 2021). The third paradigm, less commonly seen in organizational studies but widely used in social physics, uses computational tools as "*cameras*". This approach builds a knowledge map by encoding or embedding structured and/or unstructured data, then tracing social mechanisms through measurement and visualization of objects on this landscape (Liu et al. 2021; F. Shi et al. 2017). When skillfully applied, this approach complements causal inference by cultivating intuitive understanding of social processes that exclude alternative explanations. Our work uses computational tools both as "*cranes*" and "*cameras*". Through this novel encoding method, we examine traditional hypotheses regarding the influence of depth and breadth on combinatorial invention and corporate innovation, evaluating and extending theories of organizational learning.

Representation learning has the potential to accelerate the building of organizational and social theory by offering concise and elegant mathematical expression for theoretical constructs, such as distance, parallelism, and hierarchy (Aceves and Evans n.d.). The hierarchical nature of Poincaré embedding is the reason we selected it here, but technological development is not the only one where explicitly incorporating hierarchy lends leverage. It can be advantageous for any organizational topic where hierarchy plays a prominent role. Consider the following examples. Since Weber, bureaucracy and other hierarchical organizational structures have been central to understanding modern organizations (Lawrence and Poliquin n.d.; Lee 2022; Valentine 2018). Organizational and social networks typically manifest a center and periphery (Yang, Keller, and Zheng 2016). These linkages influence the power, resources and knowledge available to firms (Barkey and Godart 2013; Jia, Lewis, and Negro 2022). Within industries, firms differ in their status, reputation, and history. This heterogeneity traces a hierarchy and shapes resource allocations and opportunities (Kim and Kim 2022; Wuebker, Hampl, and Wüstenhagen 2015).

Future research can expand on our use of representation learning to characterize and trace unfolding innovation by (1) incorporating emerging data from sources of detailed technology structure including product fact-sheets, patent claims, and patent diagrams and images; and (2) examining the effects of depth and breadth of technological search on firm-level outcomes. Our study paves the way for new research that encodes data into powerful representations directly available for organizational theorization and empirical evaluation (Aceves and Evans n.d.).

CHAPTER 4 SUBJECTIVE PERSPECTIVES WITHIN LEARNED REPRESENTATIONS PREDICT HIGH-IMPACT INNOVATION

4.1 Main

From a societal perspective, innovations have become routinely characterized in terms of endogenous growth(Marx 2004; Romer 1990; Schumpeter 2013)—natural outcomes of directed economic investment and structured social arrangement. Robust legal frameworks(Moser 2005; Moser and Voena 2012), supportive institutions(Alvarado-Vargas, Callaway, and Ariss 2017), diverse and egalitarian team structures(de Vaan, Stark, et al. 2015; Wu et al. 2019; Xu, Wu, and Evans 2022b), and efficient information flows(Burt 1995; Podolny 2001) all boost creativity and collaboration to promote the quantity and quality of new knowledge and the inventions that flow from it. In parallel, a burgeoning psychological literature has attended to individual determinants of creativity, from background experiences to behaviors that cultivate innovation. Nevertheless, little is known about how situated innovators perceive inventive opportunities, to which ideas and opportunities they attend, and what they ultimately combine to generate advancing and widely appreciated innovations. This investigation has largely been limited by the lack of large-scale data on situated innovator experiences and methods available for modeling the complex, high-dimensional distribution of individual perceptions and opportunities.

The configuration of digital trace and archival data, combined with emergent machine learning methods, enables us to examine the relationship between situated, subjective positions of individual innovators and how they condition what innovators can sense and combine in future creative activities. Moreover, it allows us to explore how innovator subjectivities combine in more and less creative and successful collaborations that advance across a range of cultural domains. Here we measure situated subjectivity as a function of individuals' prior work in five domains of innovation: scientific discoveries, patented inventions, entrepreneurial ventures, movie screenplays, and Wikipedia articles. We first use dynamic word embeddings to build five conceptual spaces(Aceves and Evans 2023) for scientific, technological, business, cultural, and Wikipedia domains, respectively. Then we situate millions of innovators across those spaces based on their prior creative outputs, allowing us to characterize their positions and subjective perceptions with respect to each idea in terms of its geometric angle in conceptual space.

This allows us to observe how innovator-idea relationships become established and evolve over time. It also enables us to investigate how diverse subjectivities combine within creative collaborations to generate more or less useful and successful innovations in terms of attention and assessment: paper and patent citations, movie ratings, startup fundraising success, and Wikipedia article quality. Our statistical analysis demonstrates that the geometric perspective from innovators' positions within conceptual space consistently predicts their creative movements. Confluences of ideas generate higher innovator uptake and recombination. Innovators are more likely to adopt visible ideas, determined by their geometric angle of perception, and abandon less visible ideas. And when innovators collaborate, greater diversity of angles with respect to collaborative outputs, and a decreasing distance between collaborators, increases output success.

4.2 Idea Movements in Conceptual Space Predict Innovative Opportunities

We represent the shifting landscape of ideas with dynamic word embeddings. Word embeddings produce dense, continuous models that represent words as high-dimensional vectors to enable the measurement of precise semantic distances between them(Joulin et al. 2016;

Mikolov, Chen, Corrado, and Dean 2013; Pennington et al. 2014b). Distance between word vectors in embedding space proxies for semantic relationships underlying referenced concepts(Aceves and Evans 2023). Substantial research has demonstrated the capacity of embedding models to robustly replicate even subtle semantic relationships, including analogies and metaphor(Mikolov, Chen, Corrado, and Dean 2013) that trace veridical shifts in society and culture(Caliskan et al. 2017; Garg et al. 2018; Kozlowski et al. 2019). As such, embeddings represent a powerful instrument from which to measure changing cultural associations(Arseniev-Koehler and Foster 2020; Evans and Aceves 2016; Nelson 2021).

In this work, we adopt dynamic word embeddings to naturally account for changes in concept relationships over historical time and allow comparison between periods(Yao et al. 2018a). We use five large-scale corpora to represent the space of concepts adjacent to cultural production within each of our innovation datasets. We use (1) Microsoft Academic Graph to build a dynamic scientific space within which we project articles and conference papers; (2) the US Patent & Trademark database to build an evolving technology space within which we dynamically project patented inventions; (3) ProQuest business news to build a changing space of business discourse that forms the background for new venture creation, within which we project Crunchbase descriptions of new companies; (4) the GoogleBooks' fiction corpus to build a dynamic literary space that forms the background for movie screenplay construction, within which we project IMDB film summaries; and (5) English-language Wikipedia to create an encyclopedic knowledge space within which we project individual Wikipedia articles.

In each temporal word embedding space, we apply density-peak clustering(d'Errico et al. 2021) to track the location and movement of each semantic cluster. If a subspace has more in-flowing, convergent clusters and less out-flowing, divergent clusters, we hypothesize the area will support more inventions as it supports more creative conflict. In Appendix, we show that in all five domains, the subspaces in which fresh ideas flow host a substantially above-average number of new innovations. The most prosperous subfields appear on the frontier of current knowledge, as quantitatively identified at the boundaries of the space.

4.3 Findings

When we position scholars within conceptual landscapes as a function of their prior work, their subjective perspective is highly predictive of their likelihood to adopt new concepts within future work. We assess the subjective salience of ideas as a function of increasing idea proximity and the degree to which idea movement is visible when projected onto scholars' personal horizons. Together, a receptive subjective position increases the likelihood that an innovator adopts a new idea as it moves closer, with the rate of increase ranging from 2.9% in movies to 153.8% in scientific publications (see Appendix).

Subjective position is even more important in predicting the success of creative collaborations. While success looks different in different domains—citations for papers and patents, funding for companies, ratings for movies, and evaluations for Wikipedia articles—the effect is the same. Greater diversity in the angle between collaborators and their creative products, which we call *perspective diversity*, positively and significantly anticipates the likelihood of creative success in all areas for all time periods investigated. Diverse perspectives are counterbalanced with shared experience. Less distance between collaborator background experiences, which we call *background diversity*, may increase collaborators' ability to effectively communicate with one another and consistently lead to creative success. These two concepts are illustrated in Figure 4.1. As demonstrated in Figure 4.2, greater perspective diversity and smaller background diversity is significant and substantial in anticipating performance success across contexts. These effects persist when we consider knowledge domain, year, team size, and the achievement of the same innovator when changing teams (see Appendix).

We evaluate a natural experiment associated with Wikipedia contributions during periods in which editors self-selected to edit specific content versus during campaigns in which banners broadly recruited contributors from across Wikipedia and the web to edit specific subjects. Results suggest a causal effect of perspective and background diversity (see Appendix). We further attempted to overcome potential concern about selection or success bias by simulating creative collaborations between Large Language Model (LLM) agents designed with distinct configurations of perspective and background diversity(Guo et al. 2024; Tran et al. 2025; Wu et al. 2024). Specifically, we constructed 720 teams of three agents each, simulating specific human scientists in all four binarized combinations of perspective and background diversity (i.e., high-high, high-low, low-high, low-low). Results in Figure 4.3 demonstrate that, as with our observational data and Wikipedia natural experiment, simulated scientific collaborations between agents with high perspective and low background diversity significantly and substantially performed better in originality, quality, clarity, and significance, as assessed by advanced reasoning models, and also expert human scientists. These results can be replicated using large language models of different architectures and parameter sizes (See Appendix).

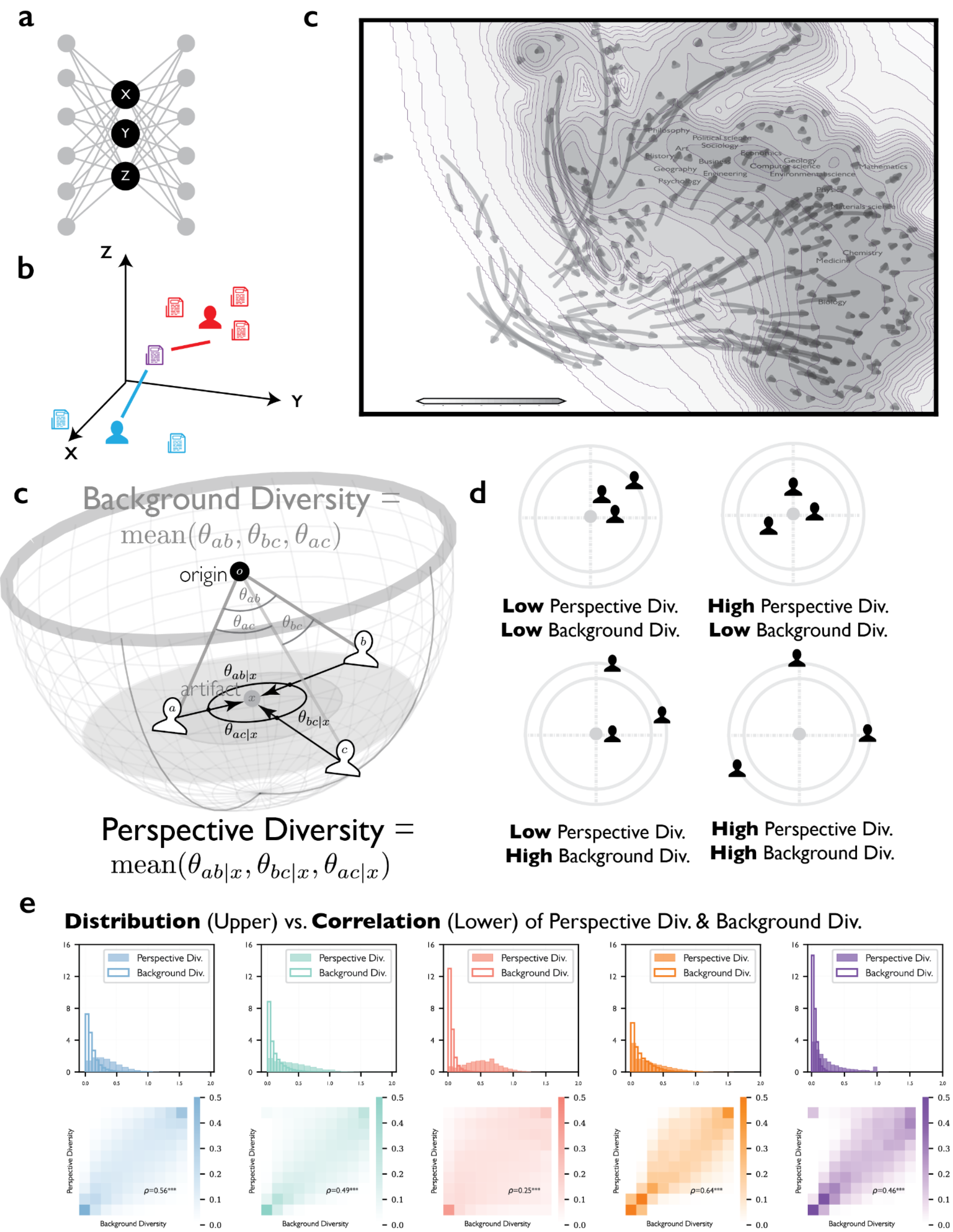

Figure 4.1: Conceptual space construction and dimensions of subjectivity. a, Schematic diagram of the construction of domain-specific conceptual spaces. b, Graph of idea movements across the conceptual space of science. c, Definition of perspective diversity ($\max(\theta_p)$) versus ($\max(\theta_b)$) background diversity. d, Configurations of perspective and background diversity. e, Statistical relationships between the distributions of perspective and background diversity by case.

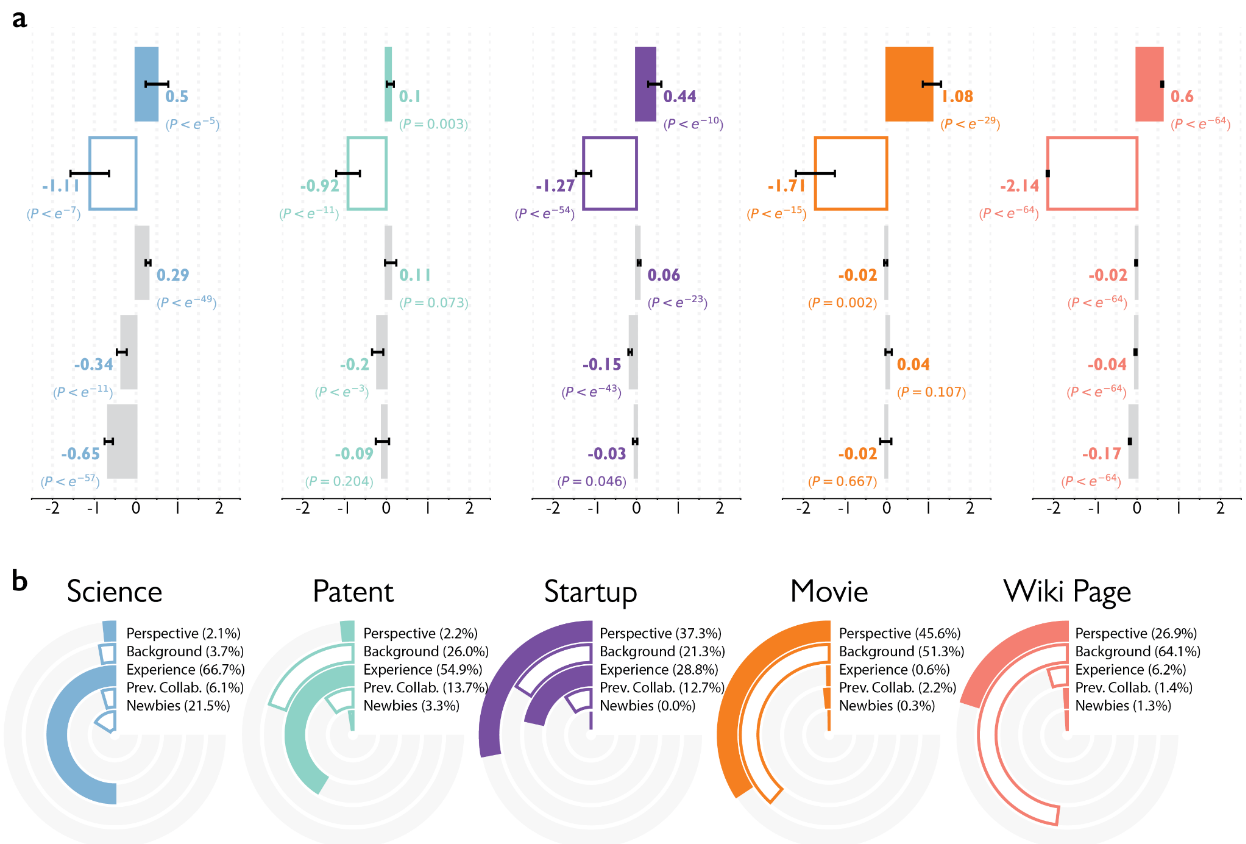

Figure 4.2: Comparison of perspective features in their contribution to the prediction of collaboration success. a, Variable-level dominance analysis demonstrating the relative importance of each predictor by comparing the incremental contribution of each variable to the model's predictive power (R^2) for domain-specific success in science, patents, start-ups, movies, and Wikipedia pages. b, Comparison of the influence of normalized variable coefficients in predicting domain-specific success.

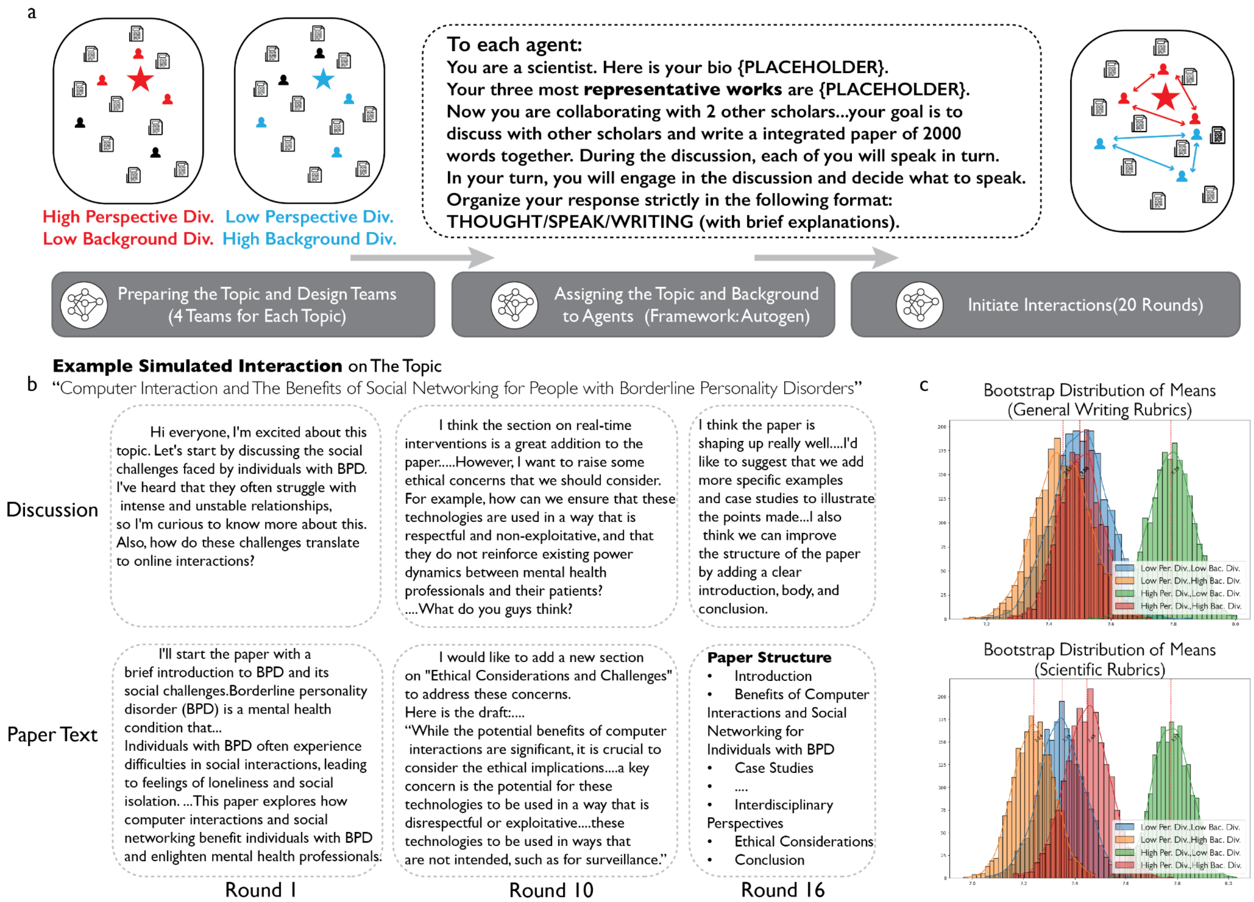

Figure 4.3: Design and Findings from the LLM Experiment. a, Key steps for assembling virtual teams from a pool of 20,000 scientists, whose profiles were obtained from the Microsoft Academic Graph. Teams were identified through a stochastic process until four were selected for each candidate topic. Information on both the topic and team members was then input into the AutoGen framework to initiate a simulated collaboration lasting 20 rounds. Each team generated a discussion transcript and produced a 2,000-word short paper on the given topic as their final output. b, Examples of discussions and paper drafts from rounds 1, 10, and 16, illustrating the step-by-step improvement of the final paper through revision and edits. c, The association between team composition and performance, assessed using general writing quality (based on AACU VALUE Rubrics) and the strength of scientific argumentation (based on NeurIPS 2024 reviewer guidelines). Among the four team compositions, those with high perspective diversity but low background diversity achieved the highest performance across both criteria.

We then explore mechanisms underlying these findings. To gain insight into team collaboration processes, we analyze self-reported contribution statements from 89,575 papers published in the multi-disciplinary journals *Science*, *Nature*, *PNAS*, and *PLOS ONE*. We examine how an individual's contribution to team diversity is associated with their role in the collaboration. Results show that when individuals bring new perspectives to a team, they are more likely to take on central roles, such as "generating ideas" and "drafting the manuscript". In contrast, those who contribute to background diversity without introducing new perspectives are more likely to take on only marginal tasks, such as "editing the manuscript". This pattern remains consistent even after controlling for author fixed effects, suggesting that the same individual may be assigned different responsibilities across teams, depending on the unique value they contribute to each collaboration (see Appendix).

Following this logic, teams with high perspective diversity may benefit from more active and engaged participation by each member, leading to better integration of individual expertise. We test this hypothesis using a knowledge integration index as a mediator of success. We construct this index from taxonomies uniquely designed for each of the five datasets. Fixed-effects models show teams that generate successful outcomes, where the team benefits from diverse perspectives without being weighed down by the confusion of excessive heterogeneity in language or background, are more likely to focus on shared expertise from members' prior work. Successful teams effectively mobilize their overlapping understanding in the new, collaborative project, which is less likely to stray into new domains that none, one or only a few members have previously explored (see Figure 4.4). This team design principle enables collaborators to bring together plumes of diverse experience, in a way that sparks inspiration through convergence and conflict to ultimately drive important innovations.

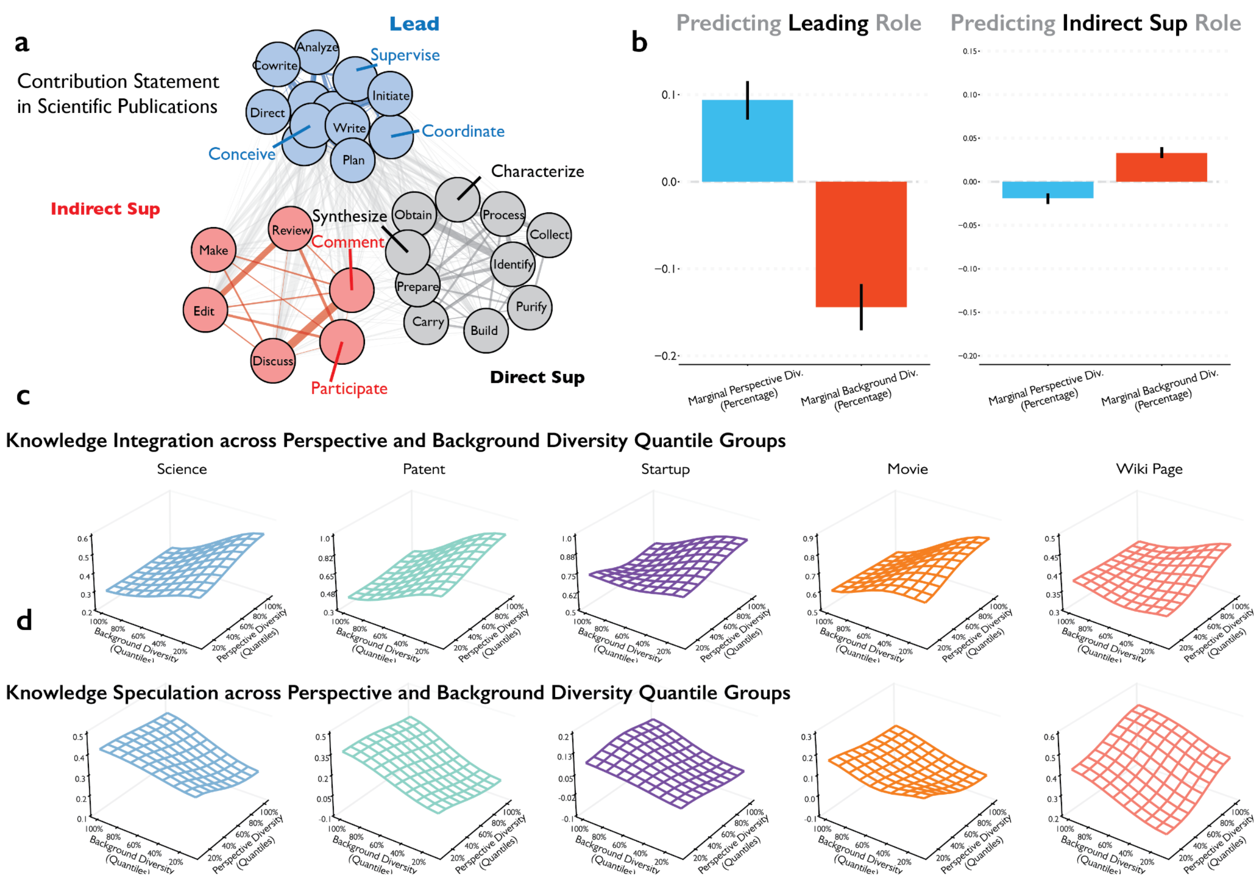

Figure 4.4: Mechanisms underlying the effects of diverse perspectives. a, Models predicting scientists' roles based on their contributions to perspective and background diversity within teams. Contributing to perspective diversity is linked to a higher likelihood of assuming central roles, whereas contributing to background diversity is associated with more marginal roles. b, The relationship between perspective and background diversity and the level of knowledge integration in the focal project, reflecting whether team members effectively recognize and leverage the intellectual overlap in their prior experiences. c, Relationship between perspective and background diversity and the likelihood of knowledge speculation, defined as a binary indicator of whether the focal project mobilizes a knowledge component that has not appeared in any team member's prior work.

4.4 Discussion

Our findings reveal that innovation success fundamentally depends on the interaction between team members' subjective perspectives and their shared background experiences. Across scientific discovery, technological invention, screenplay construction, venture creation, and Wikipedia article development, we consistently observe that teams with high perspective diversity combined with low background diversity achieve significantly higher performance outcomes. This pattern holds true in both observational data and controlled LLM simulations, suggesting a causal relationship between these diversity dimensions and innovation success. These results clarify a longstanding tension in team assembly research by demonstrating that effective collaboration requires both cognitive variety in the form of diverse perspectives and also communicative common ground through shared backgrounds, enabling teams to simultaneously generate novel combinations while maintaining effective coordination.

These findings hold profound implications for science policy and team assembly processes. Funding agencies, research institutions, and innovation-driven organizations should consider measuring and convening teams that exhibit this optimal configuration of diverse experience—high perspective diversity paired with low background diversity—rather than relying on traditional demographic or disciplinary diversity metrics alone. Our conceptual framework provides a quantitative method to identify potential collaborators who occupy distinct vantage points within shared knowledge domains, allowing for more strategic team formation. Policy interventions could focus on creating environments that foster connections between individuals with complementary perspective diversity rather than random or homogeneous groupings, potentially through targeted networking opportunities, collaborative platforms, or funding mechanisms that explicitly reward teams demonstrating subjective diversity and shared experience. This approach could significantly enhance innovation capacity across domains while maximizing the effectiveness of limited research resources.

4.5 Data and Methods

4.5.1 Datasets

Dataset of Scientific Publications. For scientific publications, we utilized a snapshot of Microsoft Academic Graph (MAG). accessed in December 2021. The dataset comprises 269,813,751 entries, including scientific papers, repositories, patents, and datasets. Due to data sparsity before 1980 and after 2015, we focused on journal articles published between 1981 and 2015, and retained only those articles with valid abstracts and at least two authors who had publishing records covering the previous five years to ensure meaningful diversity based on their past experiences. The final compiled dataset includes 20,405,559 valid papers. The dataset initially contained 278,614,730 individual profiles. We used density-based clustering(d'Errico et al. 2021; Wang et al. 2024) to identify multiple profiles linked to the same normalized author name, indexing these unique profile clusters with new author IDs. This additional disambiguation leaves 254,365,736 unique individuals in the whole dataset, among whom 16,878,569 are related to our paper sample. In mechanism analysis, we link MAG data to a dataset of increasingly mandatory author contribution statements from four journals, including 13 years of *Proceedings of the National Academy of Sciences* (*PNAS*; 18,354 statements from 2003–2015), 15 years of *Nature* (9,364 statements from 2006–2020), 3 years of *Science* (1,176 statements from 2018–2020), and 9 years of *PLOS ONE* (60,681 statements from 2006–2014). We classify the top 50 research activities within author contributions into categories including leadership (e.g., "conceived" of the study), direct support (e.g., "analyzed" data) and indirect support (e.g., “commented on” the manuscript) following the protocols established by Xu et al(Xu et al. 2022b). These activities account for 95.37% of all activities at the individual level.

Datasets of Patent Inventions. For technological inventions, we used the United States Patent and Trademark Office (USPTO) dataset. As of the data collection date, it contains a total of 8,517,464 patents from 1976 to the present. It provides content detail about the technology and its proposed use, such as patent abstracts, and metadata on inventors and assignees, such as name and gender of inventors, and patent ownership and transfers between organizations. We define a patent team as multiple inventors listed on the same granted patent, and retained only those patents with valid abstract and at least two inventors who had invention records in the previous five years, to ensure meaningful diversity measures and facilitate comparison with scientific publications. In the final dataset, we retained 2,841,066 valid patents with 1,844,890 related inventors.

Dataset of Startups and Venture Capital Firms. For business design, we used Crunchbase, a crowd-sourced platform that records extensive information on startup companies and their funding sources. Crunchbase data provides detailed information on both new ventures and investors, such as company name, address, industry, business descriptions, funding histories, and employee composition. We focused on startups based in the U.S. or with U.S. headquarters, established between January 1, 1950, and December 31, 2021. In total, this dataset contains 298,915 new ventures, 174,926 funding records, and 41,170 unique investors. Of these startups, 81,765 have received funding, and 198,226 have valid business descriptions. We retained only startups with (1) valid business descriptions and (2) at least two investors with investor records from the past five years to ensure meaningful diversity measures. The final dataset includes 10,181 valid startups and 11,572 related investors.

Dataset of Internet Movie Database. For movie creation, we used the Internet Movie Database (IMDb, www.imdb.com), which provides detailed records for each film, including movie titles, languages, regions, genres, crew information, and ratings. The official bulk data provided by IMDb is supplemented with additional information from the IMDb API (https://imdb-api.com/), which offers more in-depth information, including extended plot summaries, awards, and box office figures. The dataset we used, downloaded in February 2023, includes 261,302 movies, featuring 107,743 directors, 175,952 screenwriters, and 1,218,612 actors and actresses. We retained only those movies with valid plot summaries and at least two screenplay writers who have produced movie plots in the past five years to ensure meaningful diversity measures. The final dataset includes 29,797 valid movies and 28,556 related screenplay writers.

Dataset of Wikipedia Pages. For Wikipedia data, we used the bulk download service provided by Wikimedia database dumps. We downloaded all English Wikipedia pages in December 2023, which included 17,485,337 pages, and 9,280,838 registered users. To assess editor team structure, minor revision records that involve fewer than 100 English characters are excluded. We use each yearly snapshot as the analytical level, and retain only those with at least two editors who had Wikipedia revision records in the previous two years to ensure meaningful diversity measures. The final dataset includes 5,058,522 first-year creation records of pages, 89,261,730 page-year records as longitudinal data, and 20,501,961 users.

4.5.2 Modeling the Concept Spaces for Invention Emergence

Estimation. To construct conceptual spaces across multiple time periods, we use a dynamic embedding algorithm(Yao et al. 2018b) specifically designed for longitudinal corpora. This approach assumes and enforces a smooth transition between word vectors and their associated meanings in adjacent time periods and trains all periods jointly. Specifically, it alters the objective for the joint word embedding to the following:

$$\min_{U(1),\ldots,U(T)} \frac{1}{2}\sum_{t=1}^{T}\|Y(t) = U(t)U(t)^{T}\|_F^2 + \frac{\lambda}{2}\sum_{t=1}^{T}\|U(t)\|_F^2 + \frac{\tau}{2}\sum_{t=2}^{T}\|U(t-1) = U(t)\|_F^2 \tag{1}$$

In addition to the standard word embedding objective to align the empirical word proximities *Y*(*t*) and the modeled proximities *U*(*t*), dynamic word embeddings add two additional terms that: 1) allow the algorithm to minimize unnecessary word vector loadings (the L_2 norm of *U*(*t*)), and 2) shrink the distance between time-adjacent embeddings (the L_2 norm of the difference between *U*(*t-1*) and *U*(*t*)).[12] In this way, the jointly trained algorithm keeps adjacent semantic spaces close in structure. This approach allows us to construct comparable and historically accurate representations of conceptual structures for each time period.

For each case, we set a frequency threshold of 150 for word inclusion, learn 50-dimensional representations for each word using a window size of 5 and perform 10 iterations. For the science, technology, commerce, and culture cases, we use a 5-year time window, while for Wikipedia, we use a 2-year time window to capture the fast-paced changes and frequent updates typical of (hyper)active online platforms. To address instability in the IMDB and Crunchbase data due to their smaller volume, we use Google *n*-grams to construct a cultural space defined by fiction and nonfiction, within which we project the IMDb screenplay summaries, and use 118 financial newspapers (e.g., *Wall Street Journal*, *Financial Times*, *Bloomberg Businessweek*) from ProQuest to construct a commercial space, within which we projected the Crunchbase venture descriptions.

[12] This dynamic word embedding approach relies on the formal equivalence between the standard negative sampling approach to word embedding optimization used by `word2vec` and the low-rank factorization of a pointwise mutual information (PMI) matrix of word vectors (Levy and Goldberg, n.d., 2014), and uses the PMI as input to the objective above.

Validation. We validate our dynamic semantic space with a predictive task. We expect that within space, the sub areas where elements flow in and interact are precisely where innovations are most likely to emerge, consistent with theories of recombinant innovation(Cao et al. 2022; Stark 2011b; Xiao et al. 2022). For each case, we randomly select 5,000 points in the space and track the average "in-flow" of local concepts, defined by the increase in cosine similarity between points and nearby concepts between two time periods. We then record the number of innovations (e.g., scientific papers, technology patents, screenplays, company descriptions, Wikipedia pages) that appear within a specific range in the second period. Correlations between local concept convergence and innovation bursts are always positively and significantly correlated across contexts, as illustrated in Figures S1 to S5. We also conducted a Dumitrescu & Hurlin (2012) Granger Causality Test to examine the possibility of a causal relationship between the "in-flow" effect in semantic space and innovation emergence. This analysis was performed on a panel of 2,000 randomly selected subspaces. Results, presented in Table S1, indicate that spatial dynamics are a stronger driver of innovation than the reverse.

We also conducted a second validation test at the level of individual innovators. We hypothesize that innovators are more likely to adopt a concept when (1) the concept moves closer to their position over time, and (2) this movement is more visible to them. The same logic applies in reverse for concept abandonment. To test these hypotheses, we run a series of fixed effects regressions. Within each dataset, we examine the geometric relationships between innovators and concepts by measuring: (1) the change in cosine similarity between the innovator and the concept over time, and (2) the "visual angle" of concept movement, defined by the angle between its starting and ending points. These two variables, along with their interaction, are used to predict whether an innovator adopts a new concept (not used in period t) in t+1, or abandons an old concept (used in t) in t+1. We control for author and period fixed effects. As shown in Figure S6, the results support our hypotheses: innovators are more likely to adopt concepts that move closer to them, especially when such movement is more visible from their perspective.

In order to rule out the potential for reverse causality and guarantee robust estimation, in addition to estimating the position of patents and inventors within the patent space, we also project them within the adjacent scientific space, which reveals a similar pattern of results (see Supplement Materials).

4.5.3 Quantifying Diversity

We calculate an innovator's experience vector (V_i as the average of word vectors from their previous engagements—publishing science, inventing technologies, financing businesses, writing movies, and drafting Wikipedia pages. Based on these experience vectors, we construct perspective vectors for member i within team p as $V_{p,i} = V_{task} - V_i$, which reflect the relationship between an innovator's previous experience and the focal artifact they are constructing together (Allen and Hospedales 2019; Chen, Peterson, and Griffiths 2017). Given a set of experience vectors in a team $\{V1, V2, V3, ..., Vn\}$, the team will have a corresponding set of perspective vectors $\{V_{p,1}, V_{p,2}, V_{p,3}, ..., V_{p,n}\}$. The team's background diversity can be measured as

$$BD = \frac{1}{n \times (n-1)} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} CosineDistance(V_i, V_j)$$

, while perspective diversity can be defined as

$$PD = \frac{1}{n \times (n-1)} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} CosineDistance(V_{p,i}, V_{p,j})$$

.

The correlation between background diversity and perspective diversity across all five datasets is shown in Figure 1 in the main text. The coefficients, all positive and highly significant, range from 0.25 to 0.66. Variance Inflation Factor (VIF) tests have been conducted after each regression to confirm no issues with collinearity have been identified in our models.

In our analysis of mechanisms, we also estimate whether each team member's diversity contribution increases their probability of taking on distinctive roles. To measure the marginal diversity contribution for each individual, we construct marginal diversity contributions of each member, as the proportion of the difference between the full team's diversities and those of the counterfactual team where the focal member is absent, capturing the proportion of diversity that a focal individual contributes to the original team:

Marginal Contribution to Background Diversity $$MBD_a = \frac{(BD_{full} - BD_{-a})}{BD_{full}}$$

Marginal Contribution to Perspective Diversity $$MPD_a = \frac{(PD_{full} - PD_{-a})}{PD_{full}}$$

4.5.4 Quantifying Team Performance

We evaluate team performance within each case using domain-specific criteria, as extracted from previous studies. For science and technology teams, we use five-year forward citations as a proxy for impact(Price 1969; Wu et al. 2019). For startups, measure performance with the total number of funding rounds from venture capital firms(Hochberg, Ljungqvist, and Lu 2007; Ter Wal et al. 2016). Additionally, we assess long-term success using milestone events, such as initial public offerings (IPOs) or high-value acquisitions. For movies, we evaluate quality using average IMDb ratings, reflecting judgments from popular audiences, and complement it with a professional evaluation for whether the movie won an award in various movie festivals (1=YES, 0=NO). For Wikipedia pages, we use the page-quality assessment algorithm available through the official Wikipedia API, trained on expert ratings regarding article content and is blind to the process of article creation. This score has been frequently used in scientific research on open source knowledge production(Shi et al. 2019).

4.5.5 Quantifying Knowledge Integration and Knowledge Speculation

To measure knowledge integration within each team project, we utilize knowledge taxonomies, which represents each project's knowledge structure, according to a long tradition of innovation studies(Cao et al. 2022; Xiao et al. 2022). We measure whether all team members participate in decision-making and contribute effectively, through observing whether they have selected modules that align with their common and collective experiences. We constructed two measures to capture this information. The first variable, *Knowledge Integration,* measures to what extent the knowledge modules overlap with team members' previous knowledge modules on average. The second variable, *Speculation,* calculates the proportion of knowledge modules fresh to all team members. We constructed the two measures with equations:

$$Integration = \frac{1}{N_c \times N_p} \sum_{c=1}^{N_c} \sum_{p=1}^{N_p} (I_{c,p} == 1)$$

$$Speculation = \frac{1}{N_c} \sum_{c=1}^{N_c} (\sum_{p=1}^{N_p} I_{c,p} == 0)$$

where N_c represents the total number of relevant knowledge modules in the innovation product, N_p represents the number of individuals in the team, and $I_{c,p}$ is a dummy variable indicating whether innovator *p* has previously participated in category *c*, previous to this team project (1=Yes, 0=No). The operationalization details of these measures are illustrated in Figure S7.

4.5.6 Estimating the Relationship between Multi-Dimensional Diversity and Team Performance

We perform fixed-effect regressions to control for confounders including team size, year, and main subfield. Models with additional fixed effects, such as individual fixed effects, are included in the Appendix. The basic form of the regression equation is specified as:
$TeamPerformance_i = \beta 0 + \beta 1 \times PD_i + \beta 2 \times BD_i + Controls_i + FE_{teamsize} + FE_{year} + FE_{subfield} + \epsilon$,
$Controls_i$ represents controls on the team level and is case-specific. For science publications, subfields are defined by disciplines (level 1—292 in total) within MAG. For patents, control variables include (1) Number of claims; (2) Number of other patents in the same family; (3) Number of citations to non-patents (i.e., scientific documents), and subfields are defined by Cooperative Patent Classification ("CPC") categories. For movies, control variables include (1) Dummy for language; (2) Dummy for country, and subfields are defined by genres, as labeled in the IMDb dataset. For startups, control variables include (1) Dummy for headquarter location; (2) Dummy for founder's gender (0=Male, 1=Female) and (3) Dummy for the founder's race (0=White, 1=Minority), and subfields are defined by industry codes in Crunchbase[13]. For Wikipedia, we did not control for subfield fixed effects in Wikipedia pages given the mutually exclusive classification system, but control for page ID fixed effects to account for variations in content across different versions of the same article, addressing potential confounders related to the specific nature of knowledge.

Performance measures in our five cases vary in format. For zero-inflated count data (such as science citations, patent citations, and startup funding), we estimate negative binomial

[13] As listed in : https://support.crunchbase.com/hc/en-us/articles/360043146954-What-Industries-are-included-in-Crunchbase

regressions. For binary outcomes (such as the final success of startups and award-winning status among movies), we use logistic regressions. For continuous outcomes (such as average movie ratings and quality scores for Wikipedia pages), we apply OLS models.

Results for the main models and mechanism analyses are presented in Tables S2 to S8, while robustness checks are reported in Tables S9 to S11, and Figures S8 to S13.

4.5.7 The LLM Simulation Experiment

To test the causal relationship we identified, we design a virtual experiment based on a methodological framework of multi-agent interactions using large language models (LLMs). We begin by selecting 50 key topics from the MAG dataset and construct four virtual teams for each topic, following the design principles outlined in this paper. These teams are assembled from a pool of 20,000 real-world scientists, whose profiles are also drawn from MAG. We then input both the topic information and team member backgrounds into LLMs to initiate simulated discussions, prompting the agents to collaboratively generate a shared output. We use three models—Mistral-Nemo-Instruct-2407, Gemma-3-27b-it, QwQ-32b. As of April 2025, when this analysis was conducted, these models were recognized as some of the top-performing models available. With all other potential confounders held constant and key predictors randomly assigned, the relationship between team design and the quality of collaborative outcomes can be interpreted as causal.

For each virtual team—defined by its assigned topic, selected members, and designated LLM—we simulate collaboration using the AutoGen toolkit. Each team engages in 20 rounds of discussion and editing, during which all agents participate in conversation and jointly draft, revise, and polish a short paper. After collecting all papers produced through these simulations, we evaluate their quality using GPT-4o (accessed in February 2025) (Kim, Evans, and Schein 2024). Two assessment criteria are employed: (1) the Association of American Colleges and Universities (AAC&U) VALUE Rubrics for written communication, to assess overall writing quality; and (2) the NeurIPS 2024 reviewer guidelines, to assess the strength of the scientific argument. We set the temperature of GPT-4o to 0.3 to ensure consistency while allowing for slight variation in responses. Finally, we statistically test whether the quality scores differ significantly across team compositions.

Details on the model settings—including prompts, examples of team interactions, and LLM-generated short papers—are provided in the Appendix, with Figure S15 visually illustrating the methodological framework.

CHAPTER 5 CONCLUSION: A TYPOLOGY OF INNOVATION SYSTEMS

How do innovations emerge? Across my three essays, I investigate the emergence of innovation in a range of social domains. This research demonstrates that the theoretical and conceptual toolkit of complexity, recombination, evolution, and creative destruction, despite their individual widespread use, each approach presents subtle dimensions that can steer systems in fundamentally different directions. Advancing innovation theory requires identifying these dimensions and developing a framework to guide future inquiry. This is the central aim of the project. Although space is limited and does not allow for a comprehensive exploration of all possible dimensions, this study identifies several important factors that merit a few further discussions.

In the remainder of this chapter, I examine each of these dimensions, explain their significance, and, based on them, propose a new typology of innovation systems.

5.1 The First Dimension: Risk Appetite

It may seem unusual to suggest that a domain or industry has a "risk appetite." Typically, this term describes individuals or hedge funds—we have all heard cautionary tales about reckless traders investing in highly volatile stocks and risking their whole portfolio for a substantial return. The idea of risk appetite can be essential for innovation analyses, however, because innovation, by its definition, involves balancing uncertainty with potential gains. Choosing a radical strategy to explore unfamiliar territory presents both opportunities and dangers, a pattern supported by decades of evidence on exploration and exploitation (March 1991; Raisch and Birkinshaw 2008).

This trade-off can also be described in statistical terms, as the two types of errors in hypothesis testing. Any innovation decision is essentially a gamble, testing whether the innovators' ideas and tastes (prediction) align with the realities of nature and the needs and preferences of the population (true value) at the time. As in all hypothesis testing, Type I and Type II errors cannot be minimized simultaneously. Risk-averse individuals may conserve resources but miss high-potential opportunities (false negatives), whereas risk-seeking individuals are more likely to seize those opportunities, often at the cost of frequent failures (false positives).

While all innovations involve both types of risk, the practical costs associated with each can differ significantly across systems. Innovation in some systems, when unsuccessful, result in substantial costs due to false positives (failures), whereas false negatives (missed opportunities) may have little immediate impact. A typical example comes from political institutions and legal reforms. When reforms fail to achieve their intended goals or strike a balance among competing interests, they can escalate into revolutions or wars. In contrast, maintaining the status quo may stoke tensions, but it often preserves the possibility of future resolution. Military strategy provides another example. Both the Alpine crossing and the Blitzkrieg are famous military innovations, known for their boldness and early successes. However, in both cases, failures can not only bring down a campaign, but also the empires behind them.

On the other hand, some innovation systems are designed to tolerate or even encourage false positives. Scientists and inventors may fail hundreds or even thousands of times; but a single breakthrough can secure their place in history. This was how Edison invented the light bulb—and it also reflects the guiding philosophy behind many modern efforts, such as AI-driven automation in science. In these systems, false positives are not just tolerated but encouraged.

Machines are valuable precisely because they help absorb and manage the inherent risks associated with false positives.

From this perspective, the structure of risk, embedded within different innovation systems, shapes the optimal path of evolution. While the mechanism of recombination remains fundamental, each system carries its own inertia, influencing which combinations are prioritized or avoided. In the first paper of this thesis, we find that startups are more likely to succeed when engaging in higher-order recombinations, precisely because many have already secured substantial investment at this stage, and the system is generally more risk-averse than science and technology. This idea has implications not only for innovation systems themselves, but also for individuals working within them. It determines which behaviors are rewarded or penalized. Becoming socialized into an innovation system—and thus a professional innovator—partly means internalizing its standards and prevailing attitudes toward risk. For example, becoming a mature politician or business leader often requires greater caution and conservatism than is expected of most artists.

5.2 The Second Dimension: Prevalence of Precedents

The first dimension concerns the demand for two types of errors across domains, while the second dimension focuses, by contrast, on their supply. The resources required, along with the difficulty of making such errors, vary across social spheres, largely depending on how knowledge and memory are produced, stored, and communicated.

On one end of this continuum, we find science, business strategies, and legislation. These domains provide nearly unlimited access to open knowledge, including billions of online research publications, business school case studies, and legal precedents. Making a false positive error may involve little more than piecing together fragments from several unremarkable journal articles. By contrast, as long as innovators maintain basic productivity, false negatives tend to be less frequent and more easily corrected. For example, passing on a fast-growing startup may leave such a strong impression on an investor that it becomes a story she retells throughout her career. Such mistakes are sometimes followed by additional investments, serving as a remedy for both psychological reassurance and financial correction.

In such a knowledge system, generating successful innovation is like finding a needle in a haystack. This is also where the recombination perspective is most effectively applied: existing knowledge structures offer vast possibilities for combination and creation, but the surrounding noise is so overwhelming that innovators become easily distracted or lost.

On the other end of the continuum, we find religion and politics, where false negatives are more common than false positives. Resistance to false positives operates on two levels. First, these systems often lack sufficient historical or comparative experience applicable to specific contexts, making true positives rare and difficult to serve as creative resources. Deng Xiaoping famously noted during the Reform and Opening-up period, one should "cross the river by feeling the stones" —but it always requires courage and considerable luck. Second, such systems tend to have strong internal inertia toward stability. While shifting artistic styles may be seen as signs of creativity, unstable religious doctrines threaten the very foundation of belief. As a result, false negatives often dominate these domains. As a result, priests in 21st-century China are still trained in the quadrivium, much like their counterparts in medieval monasteries.

In the second chapter of this thesis, we find that recombination gains local recognition more quickly, partly because relevant precedents within technological systems are more readily available. The mature supply of such precedents accelerates combinatorial innovation at the local level and alters the pace of knowledge accumulation in different directions. Although this finding is based on a single knowledge system, it has broad implications for heterogeneous forms of innovation.

As one might expect, the likelihood and consequences of these two types of errors are often correlated. When false positives pose a greater risk of systemic collapse, societies are more likely to suppress them as a form of self-protection. However, these two dimensions are not orthogonal and remain theoretically distinct. In the following section, we introduce a 2×2 typology based on these dimensions. Based on this typology, we examine four types of innovation systems and explore their potential implications for the study of innovation science.

5.3 Scientist, Legislator, Warrior, and Artist

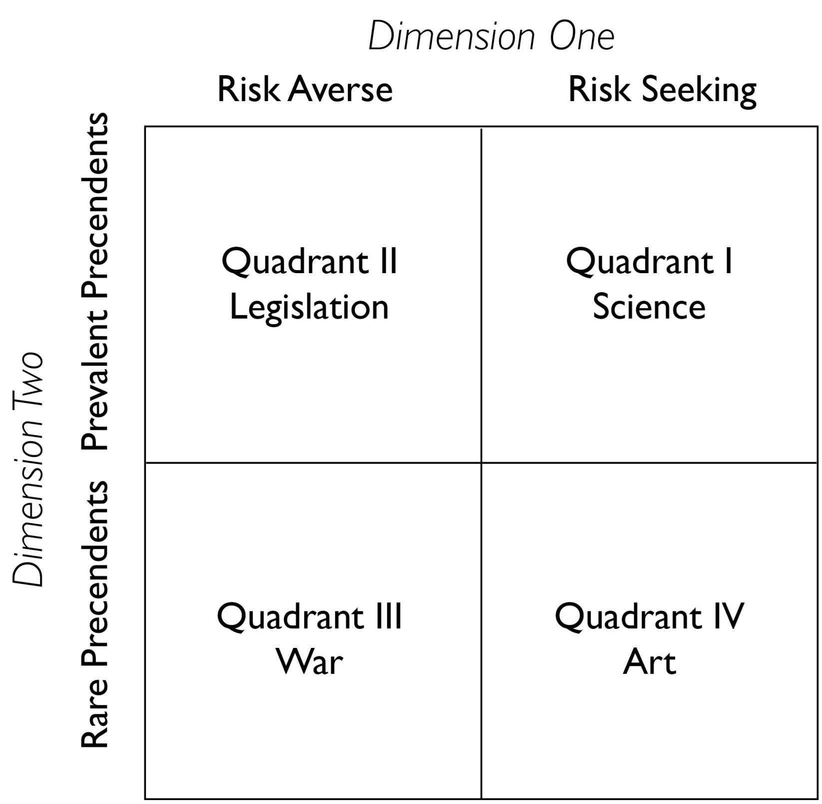

Figure 5.1: A typology of innovation systems

Four quadrants form the basis of this typology. Within each quadrant are multiple innovation systems, each characterized by distinct risk and resource preferences.

Quadrant I includes innovation systems that systematically support and maximize false positives, as such errors are both encouraged and relatively harmless. This is where science and technology are primarily situated, and where recombination occurs most actively. Innovators can readily access the intellectual resources they need from a vast pool of open knowledge and make progress through trial and error. This is also one of the most extensively studied domains in innovation research, given its abundance of data and the minimal risk associated with data use. Both the Science of Science within data science and innovation policy within management are predominantly rooted in this quadrant.

Quadrant II includes innovation systems that draw on rich precedents but are more risk-averse such as the legal system (especially the judiciary in common law systems). Unlike science and technology, which advance continuously under the banner of "endless frontiers," innovation in these domains occurs intermittently and only after careful scrutiny, deliberation, and compromise relative to system-level values. This is also where the recombination mechanism may function well: innovators draw on well-established precedents as foundational resources for building new frameworks. Given the abundance of available data and the domain's practical relevance, this quadrant holds significant promise for empirical research in the social sciences.

Beyond these two quadrants, we enter the realm of the unknown—the dark nights beyond the campfires of civilization, the domain of Dionysus, where even the most rational minds turn to oracles.

Quadrants III and IV represent cases in which novel ideas and actions emerge, yet their origins are harder to trace. Quadrant III includes instances such as war, where extraordinary moments in military history often occur only once, and strategies are rarely repeatable. Even a change in weather can determine the outcome of a battle. Art, located in Quadrant IV, differs mainly in that its failures carry fewer immediate consequences. Yet its resource structure closely resembles that of war: precedents are scarce, and genuine inspiration is difficult to replicate. This may help explain why many samurai were drawn to poetry—in both realms, they sought meaning in the singular rather than the systematic[14].

It is this theoretical ambition to make sense of the wilderness that led Norbert Elias to write *Mozart: Portrait of a Genius* (Elias 1993). Around the same time, we see *The Rise of the Medici (1993)*, one of Padgett's many works that attempt to explain revolutions and miracles. In these domains, social science theories still remain relatively scarce. What we know is far outweighed by what we do not—and much remains to be discovered.

5.4 Do Not Go Gentle into That Good Night

One of the greatest anxieties of our time stems from the perception that innovation is slowing down. Science and technology are becoming less disruptive (Park et al. 2023). New ideas are less likely to get attention (Chu and Evans 2021). Citations flow to well-known researchers, making it more difficult for outsiders to challenge field leaders (Zivin et al. 2019). And research spanning disciplines is slower and involves greater risk (Leahey et al. 2017; Teodoridis et al. 2019).

At the same time, we continue to witness the emergence of new developments, though these changes are not always benign for humanity. Artificial intelligence, once a technical term

[14] The examples presented here are not absolute. There are also borderline cases, such as popular art (e.g., commercial films) and technologies like electricity that fundamentally disrupted the industrial landscape. The arguments aim to describe the risk structure and resource conditions of innovation systems within the socio-economic context of their time, and should be understood as an ideal type, applying to the majority of cases rather than all.

in computer science or a speculative theme in fiction, now threatens to replace human labor across industries. Starlinks are used in wars. In Asia, after being detained by the Chinese government for unethical research, Jiankui He has chosen to resume his controversial gene editing work.

There are many possible explanations for this apparent contradiction, and one of them is the shrinking of institutionalized innovation, or what might be described as the deinstitutionalization of innovation. In other words, Quadrants I and II—the domains where we have accumulated years of experience—are becoming increasingly conservative. Meanwhile, in Quadrants III and IV, novel developments continue to emerge but fall beyond common view. Whether one accepts this claim or not, it is hard to deny that these occasional sparks—capable of burning an old world—often lie beyond the scope of our current theories.

I conclude the discussion here and leave the many important questions about the emergence and diffusion of innovation for future exploration. The goal of this thesis, as with my other projects, is to gather more accurate information, integrate existing insights with empirical evidence, represent social dynamics with clarity and precision, and reflect on both the scope and the limits of my arguments. I hope I have achieved at least part of this goal and, to some extent, contributed to the advancement of innovation science.

REFERENCES

Aceves, Pedro, and James Evans. n.d. “Mobilizing Conceptual Spaces: How Word Embedding Models Can Inform Measurement and Theory within Organization Science.”

Aceves, Pedro, and James A. Evans. 2023. “Mobilizing Conceptual Spaces: How Word Embedding Models Can Inform Measurement and Theory Within Organization Science.” *Organization Science*. doi: 10.1287/orsc.2023.1686.

Aghion, Philippe, Nick Bloom, Richard Blundell, Rachel Griffith, and Peter Howitt. 2005. “Competition and Innovation: An Inverted-U Relationship.” *The Quarterly Journal of Economics* 120(2):701–28.

Aharonson, Barak S., and Melissa A. Schilling. 2016. “Mapping the Technological Landscape: Measuring Technology Distance, Technological Footprints, and Technology Evolution.” *Research Policy* 45(1):81–96.

Ahmadpoor, Mohammad, and Benjamin F. Jones. 2019. “Decoding Team and Individual Impact in Science and Invention.” *Proceedings of the National Academy of Sciences of the United States of America* 116(28):13885–90.

Akcigit, U., and S. T. Ates. 2019. “What Happened to US Business Dynamism?”

Akcigit, Ufuk, and Sina T. Ates. 2021. “Ten Facts on Declining Business Dynamism and Lessons from Endogenous Growth Theory.” *American Economic Journal: Macroeconomics* 13(1):257–98.

Akcigit, Ufuk, Ms Wenjie Chen, Mr Federico J. Diez, Mr Romain A. Duval, Philipp Engler, Jiayue Fan, Chiara Maggi, Ms Marina Mendes Tavares, Mr Daniel A. Schwarz, Mr Ippei Shibata, and Carolina Villegas-Sánchez. 2021. *Rising Corporate Market Power: Emerging Policy Issues*. International Monetary Fund.

Alcácer, Juan, and Michelle Gittelman. 2006. “Patent Citations as a Measure of Knowledge Flows: The Influence of Examiner Citations.” *The Review of Economics and Statistics* 88(4):774–79.

Allen, Carl, and Timothy Hospedales. 2019. ““Analogies Explained: Towards Understanding Word Embeddings.””

Alvarado-Vargas, Marcelo J., Stephen K. Callaway, and Sonny Ariss. 2017. “Explaining Innovation Outputs by Different Types of R&D Inputs: Evidence from US Universities.” *JOURNAL OF STRATEGY AND MANAGEMENT* 10(3):326–41.

Anderson, Benedict. 2016. *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Revised edition. London New York: Verso.

Anderson, Philip. 1999. “Perspective: Complexity Theory and Organization Science.”

Organization Science 10(3):216–32.

Anon. 1966. *Some Methods for Classification and Analysis of Multivariate Observations*. Defense Technical Information Center.

Anon. 2013. “Capital.” Pp. 32–35 in *Karl Marx*. Routledge.

Aral, Sinan, and Paramveer S. Dhillon. 2023. “What (Exactly) Is Novelty in Networks? Unpacking the Vision Advantages of Brokers, Bridges, and Weak Ties.” *Management Science* 69(2):1092–1115.

Archibugi, Daniele, Andrea Filippetti, and Marion Frenz. 2013. “Economic Crisis and Innovation: Is Destruction Prevailing over Accumulation?” *Research Policy* 42(2):303–14.

Arendt, Hannah. 1970. *The Human Condition*. Chicago, IL: University of Chicago Press.

Argote, Linda, and Jerry M. Guo. 2016. “Routines and Transactive Memory Systems: Creating, Coordinating, Retaining, and Transferring Knowledge in Organizations.” *Research in Organizational Behavior* 36:65–84.

Aristotle. 1999. *The Metaphysics*. New Ed edition. London: Penguin Classics.

Arora, Li, Liang, Ma, and Risteski. n.d. “Linear Algebraic Structure of Word Senses, with Applications to Polysemy.” *International Journal of Small Craft Technology*. doi: 10.1162/tacl_a_00034/43451.

Arora, Sanjeev, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2018. “Linear Algebraic Structure of Word Senses, with Applications to Polysemy.” *Transactions of the Association for Computational Linguistics* 6:483–95.

Arseniev-Koehler, Alina, Susan D. Cochran, Vickie M. Mays, Kai-Wei Chang, and Jacob G. Foster. 2022. “Integrating Topic Modeling and Word Embedding to Characterize Violent Deaths.” *Proceedings of the National Academy of Sciences of the United States of America* 119(10):e2108801119.

Arseniev-Koehler, Alina, and Jacob G. Foster. 2020. “Sociolinguistic Properties of Word Embeddings.”

Arthur, W. Brian. 1995. “Complexity in Economic and Financial Markets.” *Complexity* 1(1):20–25.

Arthur, W. Brian. 2009. *The Nature of Technology: What It Is and How It Evolves*. Reprint edition. Free Press.

Arthur, W. Brian, and Wolfgang Polak. 2006. “The Evolution of Technology within a Simple Computer Model.” *Complexity* 11(5):23–31.

Augustine, Saint. 2003. *City of God*. Penguin UK.

Baevski, Alexei, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 17--23 Jul 2022. “data2vec: A General Framework for Self-Supervised Learning in Speech, Vision and Language.” Pp. 1298–1312 in *Proceedings of the 39th International Conference on Machine Learning*. Vol. 162, *Proceedings of Machine Learning Research*, edited by K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato. PMLR.

Barbot, Baptiste. 2022. “Intra-Individual Variability in Creativity: Nature, Measurement, and Prospects.” *European Psychologist*. doi: 10.1027/1016-9040/a000470.

Barkey, Karen, and Frederic C. Godart. 2013. “Empires, Federated Arrangements, and Kingdoms: Using Political Models of Governance to Understand Firms’ Creative Performance.” *ORGANIZATION STUDIES* 34(1):79–104.

Barlow, Matthew A., J. Cameron Verhaal, and Ryan W. Angus. 2019. “Optimal Distinctiveness, Strategic Categorization, and Product Market Entry on the Google Play App Platform.” *Strategic Management Journal*.

Barns, Sarah. 2020. “The Uberisation of Everything.” Pp. 79–98 in *Platform Urbanism: Negotiating Platform Ecosystems in Connected Cities*, edited by S. Barns. Singapore: Springer Singapore.

Battiston, Stefano, J. Doyne Farmer, Andreas Flache, Diego Garlaschelli, Andrew G. Haldane, Hans Heesterbeek, Cars Hommes, Carlo Jaeger, Robert May, and Marten Scheffer. 2016. “Complexity Theory and Financial Regulation.” *Science* 351(6275):818–19.

Becker, Markus C., and Francesco Zirpoli. 2008. “Applying Organizational Routines in Analyzing the Behavior of Organizations.” *Journal of Economic Behavior & Organization* 66(1):128–48.

Bell, Geoffrey G. 2005. “Clusters, Networks, and Firm Innovativeness.” *Strategic Management Journal* 26(3):287–95.

Berlin, Isaiah. 2013. *The Hedgehog and the Fox: An Essay on Tolstoy’s View of History - Second Edition*. Princeton University Press.

Bernstein, Jeffrey I., and M. Ishaq Nadiri. 1989. “Research and Development and Intra-Industry Spillovers: An Empirical Application of Dynamic Duality.” *The Review of Economic Studies* 56(2):249–67.

Bikard, Michaël. 2020. “Idea Twins: Simultaneous Discoveries as a Research Tool.” *Strategic Management Journal* 41(8):1528–43.

van Binsbergen, Jules H., Svetlana Bryzgalova, Mayukh Mukhopadhyay, and Varun Sharma. n.d. “(Almost) 200 Years of News-Based Economic Sentiment.” *SSRN Electronic Journal*.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. “Latent Dirichlet Allocation.” *Journal of Machine Learning Research: JMLR* 3(Jan):993–1022.

Bloom, Nicholas, Charles Jones, John Van Reenen, and Michael Webb. 2017. "Are Ideas Getting Harder to Find?"

Bloom, Nicholas, Mark Schankerman, and John Van Reenen. 2013. "Identifying Technology Spillovers and Product Market Rivalry." *Econometrica: Journal of the Econometric Society* 81(4):1347–93.

Blossfeld, Hans-Peter, Katrin Golsch, and Gotz Rohwer. 2007. "Event History Analysis With Stata."

Boh, Wai Fong, Roberto Evaristo, and Andrew Ouderkirk. 2014. "Balancing Breadth and Depth of Expertise for Innovation: A 3M Story." *Research Policy* 43(2):349–66.

Brahm, Francisco, Anne Parmigiani, and Jorge Tarzijan. 2021. "Can Firms Be Both Broad and Deep? Exploring Interdependencies Between Horizontal and Vertical Firm Scope." *Journal of Management* 47(5):1219–54.

Brian Arthur, W. 2009. *The Nature of Technology: What It Is and How It Evolves*. Simon and Schuster.

Burt, Ronald S. 1995. *Structural Holes: The Social Structure of Competition*. 1st Paperback Edition. Cambridge, Mass.: Harvard University Press.

Burt, Ronald S. 2004. "Structural Holes and Good Ideas." *The American Journal of Sociology* 110(2):349–99.

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. 2017. "Semantics Derived Automatically from Language Corpora Contain Human-like Biases." *Science* 356(6334):183–86.

Callon, Michel. 1986. "ÉLÉMENTS POUR UNE SOCIOLOGIE DE LA TRADUCTION: La Domestication Des Coquilles Saint-Jacques et Des Marins-Pêcheurs Dans La Baie de Saint-Brieuc." *L'Année Sociologique (1940/1948-)* 36:169–208.

Callon, Michel, Arie Rip, and John Law. 1986. *Mapping the Dynamics of Science and Technology: Sociology of Science in the Real World*. Springer.

Cao, Likun, Ziwen Chen, and James Evans. 2022. "Destructive Creation, Creative Destruction, and the Paradox of Innovation Science." *Sociology Compass* 16(11):e13043.

Capelleras, Joan-Lluis, and Francis J. Greene. 2008. "The Determinants and Growth Implications of Venture Creation Speed." *Entrepreneurship and Regional Development* 20(4):317–43.

Carlson, Natalie A. 2023. "Differentiation in Microenterprises." *Strategic Management Journal* 44(5):1141–67.

Chae, Bongsug, and David L. Olson. 2021. "Discovering Latent Topics of Digital Technologies From Venture Activities Using Structural Topic Modeling." *IEEE Transactions on*

Computational Social Systems 8(6):1438–49.

Chamberlain, Benjamin Paul, James Clough, and Marc Peter Deisenroth. 2017. "Neural Embeddings of Graphs in Hyperbolic Space."

Chami, Ines, Rex Ying, Christopher Ré, and Jure Leskovec. 2019. "Hyperbolic Graph Convolutional Neural Networks." *Advances in Neural Information Processing Systems* 32:4869–80.

Chang, Sea Jin. 2004. "Venture Capital Financing, Strategic Alliances, and the Initial Public Offerings of Internet Startups." *Journal of Business Venturing* 19(5):721–41.

Chang, Shann-Bin, Kuei-Kuei Lai, and Shu-Min Chang. 2009. "Exploring Technology Diffusion and Classification of Business Methods: Using the Patent Citation Network." *Technological Forecasting and Social Change* 76(1):107–17.

Chattopadhyay, Shinjinee, and Janet Bercovitz. 2020. "When One Door Closes, Another Door Opens … for Some: Evidence from the Post-TRIPS Indian Pharmaceutical Industry." *Strategic Management Journal* 41(6):988–1022.

Chen, Dawn, Joshua C. Peterson, and Thomas L. Griffiths. 2017. ""Evaluating Vector-Space Models of Analogy.""

Choi, Jaeho, Anoop Menon, and Haris Tabakovic. 2021. "Using Machine Learning to Revisit the Diversification–performance Relationship." *Strategic Management Journal* 42(9):1632–61.

Choudhury, Prithwiraj, Ryan T. Allen, and Michael G. Endres. 2021. "Machine Learning for Pattern Discovery in Management Research." *Strategic Management Journal* 42(1):30–57.

Choudhury, Prithwiraj, and Martine R. Haas. 2018. "Scope versus Speed: Team Diversity, Leader Experience, and Patenting Outcomes for Firms." *Strategic Management Journal* 39(4):977–1002.

Choudhury, Prithwiraj, Dan Wang, Natalie A. Carlson, and Tarun Khanna. 2019. "Machine Learning Approaches to Facial and Text Analysis: Discovering CEO Oral Communication Styles." *Strategic Management Journal* 40(11):1705–32.

Christensen, Clayton M., Rory McDonald, Elizabeth J. Altman, and Jonathan E. Palmer. 2018. "Disruptive Innovation: An Intellectual History and Directions for Future Research." *Journal of Management Studies* 55(7):1043–78.

Chu, Johan S. G., and James A. Evans. 2021. "Slowed Canonical Progress in Large Fields of Science." *Proceedings of the National Academy of Sciences of the United States of America* 118(41). doi: 10.1073/pnas.2021636118.

Cohen, Michael D., James G. March, and Johan P. Olsen. 1972. "A Garbage Can Model of Organizational Choice." *Administrative Science Quarterly* 17(1):1–25.

Cohen, Wesley M., Richard R. Nelson, and John P. Walsh. 2000. "“Protecting Their Intellectual Assets: Appropriability Conditions and Why U.S. Manufacturing Firms Patent (or Not).”" National Bureau of Economic Research, 7552.

Cornelissen, Joep, and Magdalena Cholakova. 2021. "Profits Uber Everything? The Gig Economy and the Morality of Category Work." *Strategic Organization* 19(4):722–31.

Corrocher, Nicoletta, Franco Malerba, and Fabio Montobbio. 2007. "Schumpeterian Patterns of Innovative Activity in the ICT Field." *Research Policy* 36(3):418–32.

Crafts, Nicholas. 2021. "Artificial Intelligence as a General-Purpose Technology: An Historical Perspective." *OXFORD REVIEW OF ECONOMIC POLICY* 37(3):521–36.

Crisp, Richard J., and Rhiannon N. Turner. 2011. "Cognitive Adaptation to the Experience of Social and Cultural Diversity." *Psychological Bulletin* 137(2):242–66.

Csikszentmihalyi, Mihaly. 2009. *Creativity: Flow and the Psychology of Discovery and*. Harper Collins.

Davis, Jason P., and Kathleen M. Eisenhardt. 2011. "Rotating Leadership and Collaborative Innovation: Recombination Processes in Symbiotic Relationships." *Administrative Science Quarterly* 56(2):159–201.

De Dreu, Carsten K. W., and Bernard A. Nijstad. 2008. "Mental Set and Creative Thought in Social Conflict: Threat Rigidity versus Motivated Focus." *Journal of Personality and Social Psychology* 95(3):648–61.

De Dreu, C. K., and M. A. West. 2001. "Minority Dissent and Team Innovation: The Importance of Participation in Decision Making." *The Journal of Applied Psychology* 86(6):1191–1201.

De Vaan, Mathijs, David Stark, and Balazs Vedres. 2015. "Game Changer: The Topology of Creativity." *The American Journal of Sociology* 120(4):1144–94.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. *BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding*. edited by J. Burstein, C. Doran, and T. Solorio. Association for Computational Linguistics.

Dhanaraj, Charles, and Arvind Parkhe. 2006. "Orchestrating Innovation Networks." *Academy of Management Review* 31(3):659–69.

D'hondt, Eva, Suzan Verberne, Nelleke Oostdijk, Jean Beney, Cornelius Koster, and Lou Boves. 2014. "Dealing with Temporal Variation in Patent Categorization." *Information Retrieval* 17(5-6):520–44.

Diamond, Jared. 2013. *Guns, Germs and Steel: A Short History of Everybody for the Last 13,000 Years*. Random House.

Dowd, T. J. 2004. "Concentration and Diversity Revisited: Production Logics and the U.S.

Mainstream Recording Market, 1940-1990." *Social Forces* 82(4):1411–55.

Drover, Will, Lowell Busenitz, Sharon Matusik, David Townsend, Aaron Anglin, and Gary Dushnitsky. 2017. "A Review and Road Map of Entrepreneurial Equity Financing Research: Venture Capital, Corporate Venture Capital, Angel Investment, Crowdfunding, and Accelerators." *Journal of Management* 43(6):1820–53.

Dyson, Freeman J. 2015. *Birds and Frogs: Selected Papers of Freeman Dyson, 1990–2014*. World Scientific Publishing Company.

Eisenhardt, Kathleen M., and Jeffrey A. Martin. 2000. "Dynamic Capabilities: What Are They?" *Strategic Management Journal* 21(10-11):1105–21.

Elias, Norbert. 1993. *Mozart: Portrait of a Genius*. First Edition. Berkeley, CA: University of California Press.

d'Errico, Maria, Elena Facco, Alessandro Laio, and Alex Rodriguez. 2021. "Automatic Topography of High-Dimensional Data Sets by Non-Parametric Density Peak Clustering." *Information Sciences* 560:476–92.

Erwin, Douglas H., and David C. Krakauer. 2004a. "Insights into Innovation." *Science* 304(5674):1117–19.

Erwin, Douglas H., and David C. Krakauer. 2004b. "Insights into Innovation." *Science* 304(5674):1117–19.

Escolar, Emerson G., Yasuaki Hiraoka, Mitsuru Igami, and Yasin Ozcan. 2023. "Mapping Firms' Locations in Technological Space: A Topological Analysis of Patent Statistics." *Research Policy* 52(8):104821.

Eshima, Yoshihiro, and Brian S. Anderson. 2017. "Firm Growth, Adaptive Capability, and Entrepreneurial Orientation." *Strategic Management Journal* 38(3):770–79.

Ethiraj, Sendil K., and Daniel Levinthal. 2004. "Modularity and Innovation in Complex Systems." *Management Science* 50(2):159–73.

Ethiraj, Sendil K., Daniel Levinthal, and Rishi R. Roy. 2008. "The Dual Role of Modularity: Innovation and Imitation." *Management Science* 54(5):939–55.

Evans, James A., and Pedro Aceves. 2016. "Machine Translation: Mining Text for Social Theory." *Annual Review of Sociology* 42(1):21–50.

Farmer, J. Doyne, and John Geanakoplos. 2009. "The Virtues and Vices of Equilibrium and the Future of Financial Economics." *Complexity* 14(3):11–38.

Feldman, Maryann P., and Ji Woong Yoon. 2012. "An Empirical Test for General Purpose Technology: An Examination of the Cohen–Boyer rDNA Technology." *Industrial and Corporate Change* 21(2):249–75.

Fellbaum, Christiane. 2005. “Wordnet and Wordnets.” Pp. 2–665 in *Encyclopedia of Language and Linguistics*, edited by A. Barber. Elsevier.

Fink, Thomas M. A., and Ali Teimouri. 2019. ““The Mathematical Structure of Innovation.””

Fleming, Lee. 2001. “Recombinant Uncertainty in Technological Search.” *Management Science* 47(1):117–32.

Fleming, Lee, and Gianna G. Giudicati. 2018. “Recombination of Knowledge.” *The Palgrave Encyclopedia of Strategic Management* 1403–6.

Fleming, Lee, and Olav Sorenson. 2001. “Technology as a Complex Adaptive System: Evidence from Patent Data.” *Research Policy* 30(7):1019–39.

Fleming, Lee, and Olav Sorenson. 2004. “Science as a Map in Technological Search.” *Strategic Management Journal* 25(8-9):909–28.

Fontana, Roberto, Alessandro Nuvolari, and Bart Verspagen. 2008. “Mapping Technological Trajectories as Patent Citation Networks. An Application to Data Communication Standards.” *Economics of Innovation and New Technology* 18. doi: 10.1080/10438590801969073.

Forgeard, Marie J. C. 2013. “Perceiving Benefits after Adversity: The Relationship between Self-Reported Posttraumatic Growth and Creativity.” *Psychology of Aesthetics, Creativity, and the Arts* 7(3):245–64.

Forman, Paul. 1971. “Weimar Culture, Causality, and Quantum Theory, 1918-1927: Adaptation by German Physicists and Mathematicians to a Hostile Intellectual Environment.” *Historical Studies in the Physical Sciences* 3:1–115.

Freeman, Chris. 1995. “The ‘National System of Innovation’in Historical Perspective.” *Cambridge Journal of Economics* 19(1):5–24.

Funk, Russell J., and Jason Owen-Smith. 2017. “A Dynamic Network Measure of Technological Change.” *Management Science* 63(3):791–817.

Galunic, D. Charles, D. Charles Galunic, and Simon Rodan. 1998. “Resource Recombinations in the Firm: Knowledge Structures and the Potential for Schumpeterian Innovation.” *Strategic Management Journal* 19(12):1193–1201.

Gans, Joshua, Erin L. Scott, and Scott Stern. 2022. “Entrepreneurial Strategy: A Choice-Based Approach to Entrepreneurship Education.” Pp. 393–400 in *Annals of Entrepreneurship Education and Pedagogy – 2023*. Edward Elgar Publishing.

Garcia-Macia, Daniel, Chang-Tai Hsieh, and Peter Klenow. 2016. “How Destructive Is Innovation?”

Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, and James Zou. 2017. ““Word Embeddings

Quantify 100 Years of Gender and Ethnic Stereotypes.""

Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. "Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes." *Proceedings of the National Academy of Sciences* 115(16):E3635–44.

Ghosh, Anindya, Ram Ranganathan, and Lori Rosenkopf. 2016. "The Impact of Context and Model Choice on the Determinants of Strategic Alliance Formation: Evidence from a Staged Replication Study." *Strategic Management Journal* 37(11):2204–21.

Gocłowska, Małgorzata A., Richard J. Crisp, and Kirsty Labuschagne. 2013. "Can Counter-Stereotypes Boost Flexible Thinking?" *Group Processes & Intergroup Relations: GPIR* 16(2):217–31.

Goldberg, Amir, Sameer B. Srivastava, V. Govind Manian, William Monroe, and Christopher Potts. 2016. "Fitting In or Standing Out? The Tradeoffs of Structural and Cultural Embeddedness." *American Sociological Review* 81(6):1190–1222.

Goncalo, Jack A., Jennifer Chatman, Michelle Duguid, and Jessica Kennedy. 2015. *Creativity from Constraint? How Political Correctness Influences Creativity in Mixed-Sex Work Groups*.

Granovetter, M. S. 1973. "The Strength of Weak Ties." *The American Journal of Sociology*.

Gross, Daniel P., and Bhaven N. Sampat. 2020. ""Inventing the Endless Frontier: The Effects of the World War II Research Effort on Post-War Innovation."" National Bureau of Economic Research, 27375.

Gross, Daniel P., and Bhaven N. Sampat. 2022. "America, Jump-Started: World War II R&D and the Takeoff of the U.S. Innovation System."

Grover, Aditya, and Jure Leskovec. 2016. *node2vec: Scalable Feature Learning for Networks*. Association for Computing Machinery.

Guillard, Charlotte, Ralf Martin, Pierre Mohnen, Catherine Thomas, and Dennis Verhoeven. n.d. *Efficient Industrial Policy for Innovation: Standing on the Shoulders of Hidden Giants*. CEPDP1813. Centre for Economic Performance, LSE.

Guo, Taicheng, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. *Large Language Model Based Multi-Agents: A Survey of Progress and Challenges*. Vol. 9.

Guzman, Jorge, and Aishen Li. n.d. "Measuring Founding Strategy."

Hall, Bronwyn H., Adam B. Jaffe, and Manuel Trajtenberg. 2001. "The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools."

Hamilton, William L., Jure Leskovec, and Dan Jurafsky. 2016. ""Diachronic Word Embeddings

Reveal Statistical Laws of Semantic Change.””

Hamilton, William L., Rex Ying, and Jure Leskovec. 2017. *Inductive Representation Learning on Large Graphs*. Curran Associates Inc.

Hanlon, W. Walker, and W. Walker Hanlon. 2015. “Necessity Is the Mother of Invention: Input Supplies and Directed Technical Change.” *Econometrica* 83(1):67–100.

Hannan, Michael T., and John Freeman. 1977. “The Population Ecology of Organizations.” *The American Journal of Sociology* 82(5):929–64.

Hargadon, Andrew, and Robert I. Sutton. 1997. “Technology Brokering and Innovation in a Product Development Firm.” *Administrative Science Quarterly* 42(4):716.

Harrison, Joseph S., Matthew A. Josefy, Matias Kalm, and Ryan Krause. n.d. “Using Supervised Machine Learning to Scale Human-Coded Data: A Method and Dataset in the Board Leadership Context.” *Strategic Management Journal*. doi: 10.1002/smj.3480.

Heinlein, Robert A. 1947. “On the Writing of Speculative Fiction.” *Of Worlds Beyond,(Chicago, 1964)* 17.

Helfat, Constance E., and Kathleen M. Eisenhardt. 2004. “Inter-Temporal Economies of Scope, Organizational Modularity, and the Dynamics of Diversification.” *Strategic Management Journal* 25(13):1217–32.

Henderson, Rebecca M., and Kim B. Clark. 1990. “Architectural Innovation: The Reconfiguration of Existing Product Technologies and the Failure of Established Firms.” *Administrative Science Quarterly* 35(1):9–30.

Hess, Andrew M., and Frank T. Rothaermel. 2011. “When Are Assets Complementary? Star Scientists, Strategic Alliances, and Innovation in the Pharmaceutical Industry.” *Strategic Management Journal* 32(8):895–909.

Hinton, G. E., and R. R. Salakhutdinov. 2006. “Reducing the Dimensionality of Data with Neural Networks.” *Science* 313(5786):504–7.

Hochberg, Yael V., Alexander Ljungqvist, and Yang Lu. 2007. “Whom You Know Matters: Venture Capital Networks and Investment Performance.” *The Journal of Finance* 62(1):251–301.

Hofstra, Bas, Vivek V. Kulkarni, Sebastian Munoz-Najar Galvez, Bryan He, Dan Jurafsky, and Daniel A. McFarland. 2020. “The Diversity–Innovation Paradox in Science.” *Proceedings of the National Academy of Sciences* 117(17):9284–91.

Holland, John H. 1992. “Complex Adaptive Systems.” *Daedalus* 121(1):17–30.

Huenteler, Joern, Jan Ossenbrink, Tobias S. Schmidt, and Volker H. Hoffmann. 2016. “How a Product’s Design Hierarchy Shapes the Evolution of Technological knowledge—Evidence from Patent-Citation Networks in Wind Power." *Research Policy* 45(6):1195–1217.

Humphery-Jenner, Mark, and Jo-Ann Suchard. 2013. "Foreign VCs and Venture Success: Evidence from China." *Journal of Corporate Finance* 21:16–35.

Jia, Ruo, Demetrius Lewis, and Giacomo Negro. 2022. "Collaborations and Innovation in Partitioned Industries: An Analysis of US Feature Film Coproductions." *ORGANIZATION SCIENCE*. doi: 10.1287/orsc.2022.1600.

Jones, Benjamin F. 2009. "The Burden of Knowledge and the 'Death of the Renaissance Man': Is Innovation Getting Harder?" *The Review of Economic Studies* 76(1):283–317.

Jones, Benjamin F., and Lawrence H. Summers. 2020. *A Calculation of the Social Returns to Innovation*. w27863. National Bureau of Economic Research. doi: 10.3386/w27863.

Joulin, Armand, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. "“Bag of Tricks for Efficient Text Classification.”"

Joulin, Armand, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. *Bag of Tricks for Efficient Text Classification*. Association for Computational Linguistics.

Jung, Hyun Ju, and Jeongsik "jay" Lee. 2016. "The Quest for Originality: A New Typology of Knowledge Search and Breakthrough Inventions." *Academy of Management Journal* 59(5):1725–53.

Kang, Donghyun, Robert S. Danziger, Jalees Rehman, and James A. Evans. 2024. "Limited Diffusion of Scientific Knowledge Forecasts Collapse." *Nature Human Behaviour* 1–9.

Kaplan, Sarah, and Keyvan Vakili. 2015. "The Double-Edged Sword of Recombination in Breakthrough Innovation." *Strategic Management Journal* 36(10):1435–57.

Kaplan, Steven N., Berk A. Sensoy, and Per Strömberg. 2009. "Should Investors Bet on the Jockey or the Horse? Evidence from the Evolution of Firms from Early Business Plans to Public Companies." *The Journal of Finance* 64(1):75–115.

Kauffman, Stuart A. 1996. "Investigations: The Nature of Autonomous Agents and the Worlds They Mutually Create." Santa Fe Institute.

Kauffman, Stuart A. 2002. *Investigations*. Oxford University Press.

Ke, Qing, Emilio Ferrara, Filippo Radicchi, and Alessandro Flammini. 2015. "Defining and Identifying Sleeping Beauties in Science." *Proceedings of the National Academy of Sciences of the United States of America* 112(24):7426–31.

Kessler, Eric H., and Alok K. Chakrabarti. 1996. "Innovation Speed: A Conceptual Model of Context, Antecedents, and Outcomes." *AMRO* 21(4):1143–91.

Keyhani, Mohammad, Yuval Deutsch, Anoop Madhok, and Moren Lévesque. 2022.

“Exploration-Exploitation and Acquisition Likelihood in New Ventures.” *Small Business Economics* 58(3):1475–96.

Kidder, J. T. 1981. “The Soul of a New Machine, Little.”

Kim, Heeyon, and Bo Kyung Kim. 2022. “To Be in Vogue: How Mere Proximity to High-Status Neighbors Affects Aspirational Pricing in the US Fashion Industry.” *Strategic Management Journal* 43(6):1208–30.

Kim, Junsol, James Evans, and Aaron Schein. 2024. *Linear Representations of Political Perspective Emerge in Large Language Models*.

Kim, Kyungsoo, Isin Guler, and Samina Karim. n.d. “Who Gets Redeployed? Inventor Characteristics and Resource Redeployment Decisions.” *Strategic Management Journal* n/a(n/a). doi: 10.1002/smj.3661.

Kiss, Andreea N., and Pamela S. Barr. 2015. “New Venture Strategic Adaptation: The Interplay of Belief Structures and Industry Context.” *Strategic Management Journal* 36(8):1245–63.

Kogut, Bruce, and Udo Zander. 1992. “Knowledge of the Firm, Combinative Capabilities, and the Replication of Technology.” *Organization Science* 3(3):383–97.

Kotha, Suresh, Seowon Joseph Shin, and Greg Fisher. 2022. “Time to Unicorn Status: An Exploratory Examination of New Ventures with Extreme Valuations.” *Strategic Entrepreneurship Journal* 16(3):460–90.

Kovács, Balázs, Gianluca Carnabuci, and Filippo Carlo Wezel. 2021. “Categories, Attention, and the Impact of Inventions.” *Strategic Management Journal* 42(5):992–1023.

Kozlowski, Austin C., Matt Taddy, and James A. Evans. 2019. “The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings.” *American Sociological Review* 84(5):905–49.

Kreiss, Daniel, and Adam J. Saffer. 2017. “Networks and Innovation in the Production of Communication: Explaining Innovations in U.S. Electoral Campaigning From 2004 to 2012.” *The Journal of Communication* 67(4):521–44.

Kremp, Pierre-Antoine. 2010. “Innovation and Selection: Symphony Orchestras and the Construction of the Musical Canon in the United States (1879–1959).” *Social Forces; a Scientific Medium of Social Study and Interpretation* 88(3):1051–82.

Kumar, Pankaj, Xiaojin Liu, and Akbar Zaheer. 2022. “How Much Does the Firm’s Alliance Network Matter?” *Strategic Management Journal* 43(8):1433–68.

Kumar, Pankaj, and Akbar Zaheer. 2019. “Ego-Network Stability and Innovation in Alliances.” *Academy of Management Journal* 62(3):691–716.

de Laat, Kim. 2014. “Innovation and Diversity Redux: Analyzing Musical Form and Content in the American Recording Industry, 1990-2009." *Sociological Forum* 29(3):673–97.

Lamoreaux, Naomi R., and Kenneth L. Sokoloff. 1999. "“Inventive Activity and the Market for Technology in the United States, 1840-1920.”" National Bureau of Economic Research, 7107.

Latour, Bruno. 1987. *Science in Action: How to Follow Scientists and Engineers Through Society*. Harvard University Press.

Latour, Bruno, and Centre de Sociologie de L'Innovation Bruno Latour. 1999. *Pandora's Hope: Essays on the Reality of Science Studies*. Harvard University Press.

Latour, Bruno, and Steven Woolgar. 1979. "Laboratory Life: The Social Construction of Social Facts."

Lawrence, Megan, and Christopher Poliquin. n.d. "The Growth of Hierarchy in Organizations: Managing Knowledge Scope." *Strategic Management Journal* n/a(n/a). doi: 10.1002/smj.3539.

Leahey, Erin, Christine M. Beckman, and Taryn L. Stanko. 2017. "Prominent but Less Productive." *Administrative Science Quarterly* 62(1):105–39.

Le Couteur, Penny. 2003. *Napoleon's Buttons: How 17 Molecules Changed History*. Jeremy P. Tarcher/Putnam.

Lee, Saerom. 2022. "The Myth of the Flat Start-up: Reconsidering the Organizational Structure of Start-Ups." *Strategic Management Journal* 43(1):58–92.

Lee, Wayne E. 2016. *Waging War: Conflict, Culture, and Innovation in World History*. Oxford University Press.

Lena, Jennifer C., and Mark C. Pachucki. 2013. "The Sincerest Form of Flattery: Innovation, Repetition, and Status in an Art Movement." *Poetics* 41(3):236–64.

Levinthal, Daniel A. 1997. "Adaptation on Rugged Landscapes." *Management Science* 43(7):934–50.

Levinthal, Daniel A. 2021. *Evolutionary Processes and Organizational Adaptation: A Mendelian Perspective on Strategic Management*. Oxford University Press.

Levy, and Goldberg. n.d. "Neural Word Embedding as Implicit Matrix Factorization." *Advances in Neural Information Processing Systems*.

Lin, Yiling, Carl Benedikt Frey, and Lingfei Wu. 2023. "Remote Collaboration Fuses Fewer Breakthrough Ideas." *Nature* 623(7989):987–91.

Linzhuo, Li, Wu Lingfei, and Evans James. 2020. "Social Centralization and Semantic Collapse: Hyperbolic Embeddings of Networks and Text." *Poetics* 78:101428.

Lippman, Steven A., and Richard P. Rumelt. 2003. “A Bargaining Perspective on Resource Advantage.” *Strategic Management Journal* 24(11):1069–86.

Liu, Fagui, Lailei Zheng, and Jingzhong Zheng. 2020. “HieNN-DWE: A Hierarchical Neural Network with Dynamic Word Embeddings for Document Level Sentiment Classification.” *Neurocomputing* 403:21–32.

Liu, Lu, Nima Dehmamy, Jillian Chown, C. Lee Giles, and Dashun Wang. 2021. “Understanding the Onset of Hot Streaks across Artistic, Cultural, and Scientific Careers.” *Nature Communications* 12(1):5392.

Ljungqvist, Alexander, and Yael V. Hochberg. 2009. *Whom You Know Matters: Venture Capital Networks and Investment Performance*. SSRN.

Lobo, Jose, and Deborah Strumsky. 2019. “Sources of Inventive Novelty: Two Patent Classification Schemas, Same Story.” *Scientometrics* 120(1):19–37.

Lodh, Suman, and Maria Rosa Battaggion. 2015. “Technological Breadth and Depth of Knowledge in Innovation: The Role of Mergers and Acquisitions in Biotech.” *Industrial and Corporate Change* 24(2):383–415.

Lundvall, Bengt-Åke. 2010. *National Systems of Innovation: Toward a Theory of Innovation and Interactive Learning*. Anthem Press.

Malhotra, Abhishek, Huiting Zhang, Martin Beuse, and Tobias Schmidt. 2021. “How Do New Use Environments Influence a Technology’s Knowledge Trajectory? A Patent Citation Network Analysis of Lithium-Ion Battery Technology.” *Research Policy* 50(9):104318.

Mannucci, Pier Vittorio, and Kevyn Yong. 2018. “The Differential Impact of Knowledge Depth and Knowledge Breadth on Creativity Over Individual Careers.” *Academy of Management Journal. Academy of Management* 61(5):1741–63.

March, James G. 1991. “Exploration and Exploitation in Organizational Learning.” *Organization Science* 2(1):71–87.

Marx, Karl. 2004. *Capital: Volume I*. Penguin UK.

Marx, Matt, Joshua S. Gans, and David H. Hsu. 2014. “Dynamic Commercialization Strategies for Disruptive Technologies: Evidence from the Speech Recognition Industry.” *Management Science* 60(12):3103–23.

McDonald, Rory M., and Kathleen M. Eisenhardt. 2020. “Parallel Play: Startups, Nascent Markets, and Effective Business-Model Design.” *Administrative Science Quarterly* 65(2):483–523.

McKnight, Brent, and Charlene Zietsma. 2018. “Finding the Threshold: A Configurational Approach to Optimal Distinctiveness.” *Journal of Business Venturing* 33(4):493–512.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "“Efficient Estimation of Word Representations in Vector Space.”"

Mikolov, Tomas, Kai Chen, Greg Corrado, Jeffrey Dean, L. Sutskever, and G. Zweig. 2013. "word2vec." *URL Https://code. Google. com/p/word2vec* 22.

Milojević, Staša. 2015. "Quantifying the Cognitive Extent of Science." *Journal of Informetrics* 9(4):962–73.

Miron-Spektor, Ella, Francesca Gino, and Linda Argote. 2011. "Paradoxical Frames and Creative Sparks: Enhancing Individual Creativity through Conflict and Integration." *Organizational Behavior and Human Decision Processes* 116(2):229–40.

Mohr, John W., and Petko Bogdanov. 2013. "Introduction—Topic Models: What They Are and Why They Matter." *Poetics* 41(6):545–69.

Moorthy, Subba, and Douglas E. Polley. 2010. "Technological Knowledge Breadth and Depth: Performance Impacts." *Journal of Knowledge Management* 14(3):359–77.

Moser, Petra. 2005. "How Do Patent Laws Influence Innovation? Evidence from Nineteenth-Century World's Fairs." *The American Economic Review* 95(4):1214–36.

Moser, Petra, and Alessandra Voena. 2012. "Compulsory Licensing: Evidence from the Trading with the Enemy Act." *The American Economic Review* 102(1):396–427.

Mozumder, Sarwar Islam, Mark J. Rutherford, and Paul C. Lambert. 2017. "stpm2cr: A Flexible Parametric Competing Risks Model Using a Direct Likelihood Approach for the Cause-Specific Cumulative Incidence Function." *The Stata Journal* 17(2):462–89.

Mueller, Philipp E. M., Dimitrios Georgakakis, Peder Greve, Simon Peck, and Winfried Ruigrok. 2021. "The Curse of Extremes: Generalist Career Experience and CEO Initial Compensation." *Journal of Management* 47(8):1977–2007.

Mukherjee, Siddhartha. 2010. *The Emperor of All Maladies: A Biography of Cancer*. Simon and Schuster.

Müller, Kathrin. 2010. "Academic Spin-Off's Transfer speed—Analyzing the Time from Leaving University to Venture." *Research Policy* 39(2):189–99.

Murdock, Jaimie, Colin Allen, and Simon DeDeo. 2017. "Exploration and Exploitation of Victorian Science in Darwin's Reading Notebooks." *Cognition* 159:117–26.

Nahata, Rajarishi. 2008. "Venture Capital Reputation and Investment Performance." *Journal of Financial Economics* 90(2):127–51.

Nakamura, Hiroko, Shinji Suzuki, Ichiro Sakata, and Yuya Kajikawa. 2015. "Knowledge Combination Modeling: The Measurement of Knowledge Similarity between Different Technological Domains." *Technological Forecasting and Social Change* 94:187–201.

Nelson, Laura K. 2021. "Leveraging the Alignment between Machine Learning and Intersectionality: Using Word Embeddings to Measure Intersectional Experiences of the Nineteenth Century U.S. South." *Poetics* 88:101539.

Nelson, Richard R., and Nathan Rosenberg. 1993. "Technical Innovation and National Systems." *National Innovation Systems: A Comparative Analysis* 322.

Nelson, Richard R., and Sidney G. Winter. 1977. "In Search of a Useful Theory of Innovation." Pp. 215–45 in *Innovation, Economic Change and Technology Policies*. Birkhäuser Basel.

Nelson, Richard R., and Sidney G. Winter. 2002. "Evolutionary Theorizing in Economics." *The Journal of Economic Perspectives: A Journal of the American Economic Association* 16(2):23–46.

Nickel, Maximillian, and Douwe Kiela. 2017a. *Poincaré Embeddings for Learning Hierarchical Representations*. Vol. 30. Curran Associates, Inc.

Nickel, Maximillian, and Douwe Kiela. 2017b. "Poincaré Embeddings for Learning Hierarchical Representations." Pp. 6338–47 in *Advances in Neural Information Processing Systems 30*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc.

Ogburn, William Fielding. 1947. "How Technology Changes Society." *The Annals of the American Academy of Political and Social Science* 249(1):81–88.

Olsson, Ola, and Bruno S. Frey. 2002. "Entrepreneurship as Recombinant Growth." *Small Business Economics* 19(2):69–80.

Omezzine, Fakher, and Isabel Maria Bodas Freitas. 2022. "New Market Creation through Exaptation: The Role of the Founding Team's Prior Professional Experience." *Research Policy* 51(5):104494.

O'Quigley, John, Ronghui Xu, and Janez Stare. 2005. "Explained Randomness in Proportional Hazards Models." *Statistics in Medicine* 24(3):479–89.

Padgett, John F., and Christopher K. Ansell. 1993. "Robust Action and the Rise of the Medici, 1400-1434." *American Journal of Sociology* 98(6):1259–1319.

Padgett, John F., and Walter W. Powell. 2012. *The Emergence of Organizations and Markets*. Princeton University Press.

Page, Scott E. 2008. *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies - New Edition*. Princeton University Press.

Page, Scott E. 2019. *The Diversity Bonus: How Great Teams Pay Off in the Knowledge Economy*. Princeton University Press.

Pahnke, Emily Cox, Riitta Katila, and Kathleen M. Eisenhardt. 2015. "Who Takes You to the

Dance? How Partners' Institutional Logics Influence Innovation in Young Firms." *Administrative Science Quarterly* 60(4):596–633.

Papazoglou, Michalis E., and Yiannis E. Spanos. 2018. "Bridging Distant Technological Domains: A Longitudinal Study of the Determinants of Breadth of Innovation Diffusion." *Research Policy* 47(9):1713–28.

Park, Michael, Erin Leahey, and Russell J. Funk. 2023. "Papers and Patents Are Becoming Less Disruptive over Time." *Nature* 613(7942):138–44.

Paruchuri, Srikanth, and Snehal Awate. 2017. "Organizational Knowledge Networks and Local Search: The Role of Intra-Organizational Inventor Networks." *Strategic Management Journal* 38(3):657–75.

Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014a. "Glove: Global Vectors for Word Representation." Pp. 1532–43 in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*.

Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014b. *GloVe: Global Vectors for Word Representation*. Association for Computational Linguistics.

Pentland, Brian T., and Henry H. Rueter. 1994. "Organizational Routines as Grammars of Action." *Administrative Science Quarterly* 39(3):484–510.

Perozzi, Bryan, Rami Al-Rfou, and Steven Skiena. 2014. *DeepWalk: Online Learning of Social Representations*. Association for Computing Machinery.

Petralia, Sergio. 2020. "Mapping General Purpose Technologies with Patent Data." *Research Policy* 49(7):104013.

Phillips, Damon J., and David A. Owens. 2004. "Incumbents, Innovation, and Competence: The Emergence of Recorded Jazz, 1920 to 1929." *Poetics* 32(3-4):281–95.

Piketty, Thomas. 2014. "Capital in the Twenty-First Century."

Piketty, Thomas, and Gabriel Zucman. 2014. "Capital Is Back: Wealth-Income Ratios in Rich Countries 1700–2010 *." *The Quarterly Journal of Economics* 129(3):1255–1310.

Podolny, Joel M. 2001. "Networks as the Pipes and Prisms of the Market." *The American Journal of Sociology* 107(1):33–60.

Pontikes, Elizabeth G. 2018. "Category Strategy for Firm Advantage." *Strategy Science* 3(4):620–31.

Porter, Michael E. 1980. *Competitive Strategy: Techniques for Analyzing Industries and Competitors*. Free Press.

Postman, Neil, and Andrew Postman. 2005. *Amusing Ourselves to Death: Public Discourse in* *the Age of Show Business*. Anniversary edition. Harlow, England: Penguin Books.

Powell, Walter W., Douglas R. White, Kenneth W. Koput, and Jason Owen-Smith. 2005. “Network Dynamics and Field Evolution: The Growth of Interorganizational Collaboration in the Life Sciences.” *The American Journal of Sociology* 110(4):1132–1205.

Prencipe, Andrea. 2000. “Breadth and Depth of Technological Capabilities in CoPS: The Case of the Aircraft Engine Control System.” *Research Policy* 29(7):895–911.

Price, Derek John de Solla. 1969. *Citation Measures of Hard Science, Soft Science, Technology and Non-Science: Paper Presented at a Conference on Communication Among Scientists and Technologists. The Johns Hopkins University, October 28-30, 1969*.

Qin, Fei, Mike Wright, and Jian Gao. 2017. “Are ‘sea Turtles’ Slower? Returnee Entrepreneurs, Venture Resources and Speed of Entrepreneurial Entry.” *Journal of Business Venturing* 32(6):694–706.

Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. “Improving Language Understanding by Generative Pre-Training.”

Raisch, Sebastian, and Julian Birkinshaw. 2008. “Organizational Ambidexterity: Antecedents, Outcomes, and Moderators.” *Journal of Management* 34(3):375–409.

Ries, Eric. 2011. *The Lean Startup: How Today’s Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses*. Crown.

Risha, Zak, Yiling Lin, Erin Leahey, and Lingfei Wu. 2023. ““Replacing the Renaissance Man: Are Teams More than the Sum of Their Parts?””

Ritter, Simone M., Rodica Ioana Damian, Dean Keith Simonton, Rick B. van Baaren, Madelijn Strick, Jeroen Derks, and Ap Dijksterhuis. 2012. “Diversifying Experiences Enhance Cognitive Flexibility.” *Journal of Experimental Social Psychology* 48(4):961–64.

Roche, Maria P., Annamaria Conti, and Frank T. Rothaermel. 2020. “Different Founders, Different Venture Outcomes: A Comparative Analysis of Academic and Non-Academic Startups.” *Research Policy* 49(10):104062.

Romer, Paul M. 1990. “Endogenous Technological Change.” *The Journal of Political Economy* 98(5, Part 2):S71–102.

Royston, Patrick. 2006. “Explained Variation for Survival Models.” *The Stata Journal: Promoting Communications on Statistics and Stata* 6(1):83–96.

Salganik, Matthew J. 2017. *Bit by Bit: Social Research in the Digital Age*. Illustrated edition. Princeton, NJ: Princeton University Press.

Salganik, Matthew J. 2019. *Bit by Bit: Social Research in the Digital Age*. Princeton University Press.

Sampat, Bhaven N., and Arvids A. Ziedonis. 2005. “Patent Citations and the Economic Value of Patents.” Pp. 277–98 in *Handbook of Quantitative Science and Technology Research: The Use of Publication and Patent Statistics in Studies of S&T Systems*, edited by H. F. Moed, W. Glänzel, and U. Schmoch. Dordrecht: Springer Netherlands.

Savin, Ivan, Kristina Chukavina, and Andrey Pushkarev. 2022. “Topic-Based Classification and Identification of Global Trends for Startup Companies.” *Small Business Economics*.

Schumpeter, Joseph A. 2006. *Theorie der wirtschaftlichen Entwicklung*. Duncker & Humblot.

Schumpeter, Joseph A. 2008. *Capitalism, Socialism, and Democracy: Third Edition*. unknown edition. Harper Perennial Modern Classics.

Schumpeter, Joseph A. 2013. *Capitalism, Socialism and Democracy*. Routledge.

Sgourev, Stoyan V., Erik Aadland, and Giovanni Formilan. 2023. “Relations in Aesthetic Space: How Color Enables Market Positioning.” *Administrative Science Quarterly* 68(1):146–85.

Shapiro, Carl. 2000. “Navigating the Patent Thicket: Cross Licenses, Patent Pools, and Standard Setting.” *Innovation Policy and the Economy* 1:119–50.

Sharma, Abhishek, Dániel Czégel, Michael Lachmann, Christopher P. Kempes, Sara I. Walker, and Leroy Cronin. 2023. “Assembly Theory Explains and Quantifies Selection and Evolution.” *Nature* 622(7982):321–28.

Shi, Feng, and James Evans. 2019. ““Science and Technology Advance through Surprise.””

Shi, Feng, and James Evans. 2023. “Surprising Combinations of Research Contents and Contexts Are Related to Impact and Emerge with Scientific Outsiders from Distant Disciplines.” *Nature Communications* 14(1641). doi: 10.1038/s41467-023-36741-4.

Shi, Feng, Jacob G. Foster, and James A. Evans. 2015. “Weaving the Fabric of Science: Dynamic Network Models of Science’s Unfolding Structure.” *Social Networks* 43:73–85.

Shi, Feng, Yongren Shi, Fedor A. Dokshin, James A. Evans, and Michael W. Macy. 2017. “Millions of Online Book Co-Purchases Reveal Partisan Differences in the Consumption of Science.” *Nature Human Behaviour* 1(4):1–9.

Shi, Feng, Misha Teplitskiy, Eamon Duede, and James A. Evans. 2019. “The Wisdom of Polarized Crowds.” *Nature Human Behaviour* 3(4):329–36.

Shiller, Robert J., and Brendan Brown. 2019. “NARRATIVE ECONOMICS: HOW STORIES GO VIRAL AND DRIVE MAJOR ECONOMIC EVENTS.” *The Quarterly Journal of Austrian Economics* 22:620+.

Shi, Yongren, Yisook Lim, and Chan S. Suh. 2018. “Innovation or Deviation? The Relationship between Boundary Crossing and Audience Evaluation in the Music Field.” *PloS One* 13(10):e0203065.

Shi, Yuan, Olav Sorenson, and David M. Waguespack. 2017. “Temporal Issues in Replication: The Stability of Centrality-Based Advantage.” *Sociological Science* 4:107–22.

Shrestha, Yash Raj, Vivianna Fang He, Phanish Puranam, and Georg von Krogh. 2021. “Algorithm Supported Induction for Building Theory: How Can We Use Prediction Models to Theorize?” *Organization Science* 32(3):856–80.

Silva, Filipi Nascimento, Aditya Tandon, Diego Raphael Amancio, Alessandro Flammini, Filippo Menczer, Staša Milojević, and Santo Fortunato. 2020. “Recency Predicts Bursts in the Evolution of Author Citations.” *Quantitative Science Studies* 1(3):1298–1308.

Simon, Herbert A. 1991a. “The Architecture of Complexity.” *Facets of Systems Science* 457–76.

Simon, Herbert A. 1991b. “The Architecture of Complexity.” Pp. 457–76 in *Facets of Systems Science*, edited by G. J. Klir. Boston, MA: Springer US.

Singh, Harpreet, David Kryscynski, Xinxin Li, and Ram Gopal. 2016. “Pipes, Pools, and Filters: How Collaboration Networks Affect Innovative Performance.” *Strategic Management Journal* 37(8):1649–66.

Sorenson, Olav. 2018. “Social Networks and the Geography of Entrepreneurship.” *Small Business Economics* 51(3):527–37.

Sorenson, Olav, Jan W. Rivkin, and Lee Fleming. n.d. “Complexity, Networks and Knowledge Flow.” *SSRN Electronic Journal*.

Sourati, Jamshid, and James A. Evans. 2023. “Accelerating Science with Human-Aware Artificial Intelligence.” *Nature Human Behaviour*. doi: 10.1038/s41562-023-01648-z.

Stańczyk-Hugiet, Ewa. 2014. “Routines in the Process of Organizational Evolution.” *Management* 18(2):73–87.

Stark, David. 2009. “The Sense of Dissonance.”

Stark, David. 2011a. *The Sense of Dissonance*. Princeton University Press.

Stark, David. 2011b. *The Sense of Dissonance: Accounts of Worth in Economic Life*. Princeton, NJ: Princeton University Press.

Sternberg, Robert J., and Todd I. Lubart. 1995. *Defying the Crowd: Cultivating Creativity in a Culture of Conformity*.

Sterzi, Valerio. 2013. “Patent Quality and Ownership: An Analysis of UK Faculty Patenting.” *Research Policy* 42(2):564–76.

Stettner, Uriel, and Dovev Lavie. 2014. “Ambidexterity under Scrutiny: Exploration and Exploitation via Internal Organization, Alliances, and Acquisitions.” *Strategic Management Journal* 35(13):1903–29.

Stuart, Toby E., and Joel M. Podolny. 2007. “Local Search and the Evolution of Technological Capabilities.” *Strategic Management Journal* 17(S1):21–38.

Sutton, Robert I. 2002. *Weird Ideas That Work: 11 1/2 Practices for Promoting, Managing, and Sustaining Innovation*. Simon and Schuster.

Taeuscher, Karl, and Hannes Rothe. 2021. “Optimal Distinctiveness in Platform Markets: Leveraging Complementors as Legitimacy Buffers.” *Strategic Management Journal* 42(2):435–61.

Taylor, Alva, and Henrich R. Greve. 2006. “Superman or the Fantastic Four? Knowledge Combination And Experience in Innovative Teams.” *Academy of Management Journal* 49(4):723–40.

Teodoridis, Florenta, Michaël Bikard, and Keyvan Vakili. 2019. “Creativity at the Knowledge Frontier: The Impact of Specialization in Fast- and Slow-Paced Domains.” *Administrative Science Quarterly* 64(4):894–927.

Teodoridis, Florenta, Jino Lu, and Jeffrey L. Furman. 2022. ““Mapping the Knowledge Space: Exploiting Unassisted Machine Learning Tools.”” National Bureau of Economic Research, 30603.

Terjesen, Siri, and Pankaj C. Patel. 2017. “In Search of Process Innovations: The Role of Search Depth, Search Breadth, and the Industry Environment.” *Journal of Management* 43(5):1421–46.

Ter Wal, Anne L. J., Oliver Alexy, Jörn Block, and Philipp G. Sandner. 2016. “The Best of Both Worlds: The Benefits of Open-Specialized and Closed-Diverse Syndication Networks for New Ventures’ Success.” *Administrative Science Quarterly* 61(3):393–432.

Thomas, Kyla. 2019. “When Disney Came to Broadway: Assessing the Impact of Corporatization in an Art World.” *Poetics* 77:101384.

Trajtenberg, Manuel, Rebecca Henderson, and Adam Jaffe. 1997. “University Versus Corporate Patents: A Window On The Basicness Of Invention.” *Economics of Innovation and New Technology* 5(1):19–50.

Tran, Khanh-Tung, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O’Sullivan, and Hoang D. Nguyen. 2025. “Multi-Agent Collaboration Mechanisms: A Survey of LLMs.”

Uzzi, Brian, Satyam Mukherjee, Michael Stringer, and Ben Jones. 2013. “Atypical Combinations and Scientific Impact.” *Science* 342(6157):468–72.

de Vaan, Mathijs, David Stark, and Balazs Vedres. 2015. “Game Changer: The Topology of Creativity.” *The American Journal of Sociology* 120(4):1144–94.

de Vaan, Mathijs, Balazs Vedres, and David Stark. 2015. “Game Changer: The Topology of Creativity1.” *AJS; American Journal of Sociology* 120(4):1–51.

Valentine, Melissa A. 2018. “Renegotiating Spheres of Obligation: The Role of Hierarchy in Organizational Learning.” *Administrative Science Quarterly* 63(3):570–606.

Verhoeven, Dennis, Jurriën Bakker, and Reinhilde Veugelers. 2016. “Measuring Technological Novelty with Patent-Based Indicators.” *Research Policy* 45(3):707–23.

Verspagen, Bart. 2007. “Mapping Technological Trajectories as Patent Citation Networks: A Study on the History of Fuel Cell Research.” *Advances in Complex Systems. A Multidisciplinary Journal* 10(1):93–115.

Vicinanza, Paul, Amir Goldberg, and Sameer B. Srivastava. 2023. “A Deep-Learning Model of Prescient Ideas Demonstrates That They Emerge from the Periphery.” *PNAS Nexus* 2(1):gac275.

Vincent, Pascal, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. *Extracting and Composing Robust Features with Denoising Autoencoders*. Association for Computing Machinery.

Vries, Thomas A. de, Thomas A. de Vries, Frank Walter, Gerben S. Van der Vegt, and Peter J. M. Essens. 2014. “Antecedents of Individuals’ Interteam Coordination: Broad Functional Experiences as a Mixed Blessing.” *Academy of Management Journal* 57(5):1334–59.

Wang, Dan J., and Sarah A. Soule. 2016. “Tactical Innovation in Social Movements.” *American Sociological Review* 81(3):517–48.

Wang, Dashun, Chaoming Song, and Albert-László Barabási. 2013. “Quantifying Long-Term Scientific Impact.” *Science* 342(6154):127–32.

Wang, Jian. 2016. “Knowledge Creation in Collaboration Networks: Effects of Tie Configuration.” *Research Policy* 45(1):68–80.

Wang, Yizhang, Jiaxin Qian, Muhammad Hassan, Xinyu Zhang, Tao Zhang, Chao Yang, Xingxing Zhou, and Fengjin Jia. 2024. “Density Peak Clustering Algorithms: A Review on the Decade 2014–2023.” *Expert Systems with Applications* 238:121860.

Weingart, Scott B. n.d. “From Trees to Webs: Uprooting Knowledge through Visualization.”

Weitzman, Martin L. 1996. *Recombinant Growth*. Harvard Institute for International Development, Harvard University.

Weitzman, Martin L. 1998. “Recombinant Growth*.” *The Quarterly Journal of Economics* 113(2):331–60.

Williamson, Oliver, and Sidney Winter. 1993. *The Nature of the Firm: Origins, Evolution, and Development*. Oxford University Press.

Wimsatt, William C. 2014. “Entrenchment and Scaffolding: An Architecture for a Theory of Cultural Change.” *Developing Scaffolds in Evolution, Culture, and Cognition* 17:77–105.

Wuebker, Robert, Nina Hampl, and Rolf Wüstenhagen. 2015. “The Strength of Strong Ties in an Emerging Industry: Experimental Evidence of the Effects of Status Hierarchies and Personal Ties in Venture Capitalist Decision Making.” *Strategic Entrepreneurship Journal* 9(2):167–87.

Wu, Lingfei, Dashun Wang, and James A. Evans. 2019. “Large Teams Develop and Small Teams Disrupt Science and Technology.” *Nature* 566(7744):378 – +.

Wu, Qingyun, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang (eric) Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Ahmed Awadallah, Ryen W. White, Doug Burger, and Chi Wang. 2024. *AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation*.

Xiao, Ting, Mona Makhija, and Samina Karim. 2022. “A Knowledge Recombination Perspective of Innovation: Review and New Research Directions.” *Journal of Management* 48(6):1724–77.

Xu, Fengli, Lingfei Wu, and James Evans. 2022a. ““Flat Teams Drive Scientific Innovation.””

Xu, Fengli, Lingfei Wu, and James Evans. 2022b. “Flat Teams Drive Scientific Innovation.” *Proceedings of the National Academy of Sciences of the United States of America* 119(23):e2200927119.

Yang, Alex J. 2024. “On the Temporal Diversity of Knowledge in Science.” *Journal of Informetrics* 18(4):101594.

Yang, Song, Franziska B. Keller, and Lu Zheng. 2016. *Social Network Analysis: Methods and Examples*. 1 edition. Los Angeles: SAGE Publications, Inc.

Yao, Zijun, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. 2018a. “Dynamic Word Embeddings for Evolving Semantic Discovery.” Pp. 673–81 in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, *WSDM ’18*. New York, NY, USA: Association for Computing Machinery.

Yao, Zijun, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. 2018b. *Dynamic Word Embeddings for Evolving Semantic Discovery*. Association for Computing Machinery.

Zander, Udo, and Bruce Kogut. 1995. “Knowledge and the Speed of the Transfer and Imitation of Organizational Capabilities: An Empirical Test.” *Organization Science* 6(1):76–92.

Zhang, Yating, Adam Jatowt, Sourav S. Bhowmick, and Katsumi Tanaka. 2016. “The Past Is Not a Foreign Country: Detecting Semantically Similar Terms across Time.” *IEEE Transactions on Knowledge and Data Engineering* 28(10):2793–2807.

Zhao, Eric Yanfei, Masakazu Ishihara, P. Devereaux Jennings, and Michael Lounsbury. 2018. “Optimal Distinctiveness in the Console Video Game Industry: An Exemplar-Based Model of Proto-Category Evolution.” *Organization Science* 29(4):588–611.

Zheng, Hongwei, Weihua Li, and Dashun Wang. 2022. "Expertise Diversity of Teams Predicts Originality and Long-Term Impact in Science and Technology."

Zheng, Lu, Likun Cao, Jie Ren, Xibao Li, Ximing Yin, and Jin Chen. 2022. "How Venture Capital Firms Choose Syndication Partners: The Moderating Effects of Institutional Uncertainty and Investment Preference." *Management and Organization Review* 18(3):463–90.

Zhou, Kevin Zheng, and Caroline Bingxin Li. 2012. "How Knowledge Affects Radical Innovation: Knowledge Base, Market Knowledge Acquisition, and Internal Knowledge Sharing." *Strategic Management Journal* 33(9):1090–1102.

Zivin, Joshua Graff, Pierre Azoulay, and Christian Fons-Rosen. 2019. "Does Science Advance One Funeral at a Time?" *The American Economic Review* 109(8):2889–2920.

Zuckerman, Ezra W. 1999. "The Categorical Imperative: Securities Analysts and the Illegitimacy Discount." *The American Journal of Sociology* 104(5):1398–1438.

LINKS FOR ONLINE APPENDICES

The appendices for all chapters are shared via Google Drive through the link below.

https://docs.google.com/document/d/1w3zl56kov_t4ol3172NTH33_PjxADdTXpyK7PMvVcrI/edit?usp=sharing