AMERICAN SOCIOLOGICAL ASSOCIATION

American Sociological Review
2015, Vol. 80(5) 875–908

DOI: 10.1177/0003122415601618
http://asr.sagepub.com

SAGE

Tradition and Innovation in Scientists' Research Strategies

Jacob G. Foster,[a] Andrey Rzhetsky,[b] and James A. Evans[b]

Abstract

What factors affect a scientist's choice of research problem? Qualitative research in the history and sociology of science suggests that this choice is patterned by an "essential tension" between productive tradition and risky innovation. We examine this tension through Bourdieu's field theory of science, and we explore it empirically by analyzing millions of biomedical abstracts from MEDLINE. We represent the evolving state of chemical knowledge with networks extracted from these abstracts. We then develop a typology of *research strategies* on these networks. Scientists can introduce novel chemicals and chemical relationships (innovation) or delve deeper into known ones (tradition). They can consolidate knowledge clusters or bridge them. The aggregate distribution of published strategies remains remarkably stable. High-risk innovation strategies are rare and reflect a growing focus on established knowledge. An innovative publication is more likely to achieve high impact than a conservative one, but the additional reward does not compensate for the risk of failing to publish. By studying prizewinners in biomedicine and chemistry, we show that occasional gambles for extraordinary impact are a compelling explanation for observed levels of risky innovation. Our analysis of the essential tension identifies institutional forces that sustain tradition and suggests policy interventions to foster innovation.

Keywords

sociology of science, tradition, innovation, networks, generative models, biomedicine, citations, awards, field theory

Why do scientists pursue a particular research problem? These decisions are consequential, not just for scientists but for science itself. Indeed, problem choice becomes more important as knowledge grows in scale and complexity (Evans and Foster 2011; Foote 2007). In an expanding universe of *possible* research questions, a topic that attracts intense investigation is separated from neglected or abandoned topics by more than just the contours of nature. Scientists' choices matter: in aggregate, patterned choices give scientific knowledge its shape and guide its future evolution.

Research choice is central to many classic investigations in the sociology of science

[a]University of California-Los Angeles
[b]University of Chicago

Corresponding Authors:
Jacob Foster, Department of Sociology, University of California-Los Angeles, 264 Haines Hall, 375 Portola Plaza, Los Angeles, CA 90095; James Evans, Department of Sociology, University of Chicago, 1126 East 59th Street, Chicago, IL 60637
E-mail: foster@soc.ucla.edu; jevans@uchicago.edu

(Busch, Lacy, and Sachs 1983; Diamond 1994; Gieryn 1978; Merton 1938; Zuckerman 1978). While it has received sustained attention in qualitative studies, it has been neglected in recent large-scale quantitative analyses (Ding, Murray, and Stuart 2006; Guimerà et al. 2005; Murray and Stern 2007; Shwed and Bearman 2010; Stuart and Ding 2006; Wuchty, Jones, and Uzzi 2007). In this article, we examine scientific choice quantitatively and at scale, using published claims in contemporary biomedicine to make inferences about underlying choices and dispositions.

As the qualitative and historical literature notes, many factors influence a scientist's choice of research problem. Some factors are particular to each scientist, and range from past interests and training (Kevles 1978; Malmgren, Ottino, and Amaral 2010) to serendipitous yet consequential encounters (Merton and Barber 2011) with new collaborators, expertise, or information (Evans 2010a; Shi, Foster, and Evans 2015). Other factors are shared across groups of scientists. Local factors reflect institutional context (Abbott 1999) or disciplinary culture (Abbott 2001; Knorr-Cetina 1999; Shapin 1995). For example, scientists in a large department may be unconstrained because they lack a common paradigm; those in an emerging subfield may be unconstrained because they have few accumulated assumptions. At the meso-scale lie factors that influence everyone in a particular discipline or field (Bourdieu 1975), like longstanding investments in particular methods and theories. The characteristic trajectory of a scientific career also shapes research choice. Graduate students, assistant professors, and eminent scientists select problems in profoundly different professional contexts, venturing risks with varied stocks of accumulated scientific capital (Bourdieu 1975) and varied opportunities to reap the reward (Merton 1968). At the largest scale are factors affecting every contemporary scientist, like the valorization of original contributions.[1]

These many factors can be systematically assembled using Bourdieu's (1975, 2004) field theory of science. In this view, scientists occupy more or less dominant positions in a specific scientific field, depending on the amount and types of scientifically recognized capital they have accumulated. Capital could be economic (e.g., grant money) or socially embedded (e.g., relationships with important scientists and patrons). It could be cultural, involving familiarity with the institutions, mores, and collective imagination of a field. Alternatively, scientists' capital could be technical, residing in skills, expertise, and intuition. Scientists "take a position" by pursuing *particular* research problems selected from the space of all those possible (Bourdieu 1975, 2004). These concrete actions are guided by the interplay between scientists' positions in the field and their *habitus*: acquired systems of tastes, dispositions, and expectations. At stake are recognition by fellow scientists, other currencies for which recognition can be traded, and an improved position in the field.

Although long neglected by sociologists of science, Bourdieu's distinctive approach has recently experienced a renaissance (Albert and Kleinman 2011; Camic 2011, 2013; Kim 2009; Panofsky 2011). This re-appropriation largely rests on a move to relax the autonomy of a given scientific domain and explore both its interchange with other domains and its hierarchical embedding in other fields.[2] For example, Bourdieu's approach has illuminated the idiosyncratic history of behavioral genetics (Panofsky 2011) and early-twentieth-century economics (Camic 2011). We adopt Bourdieu for a different reason: his field theory provides an organizing framework for the wealth of data now available about the outcome and consequences of scientific choice.

Using contemporary biomedicine as an example, we analyze the large-scale pattern of research claims and provide a strategic, dispositional account of scientific choice, drawing on the rich, published record of successful research choices and rewards. We focus here on the tension created by conflicting professional demands, which map to distinct strategies for accumulating recognition

and resources. To remain in the research game requires productivity. Scientists typically achieve this by incremental contributions to established research directions. This may yield enough recognition to maintain a (relatively low) position. Achieving high status, by contrast, requires original and transformative contributions, often obtained by pursuing risky new directions (Merton 1957).

These conflicting demands create a tension between two broad strategies: productive *tradition* and risky *innovation* (Kuhn [1959] 1977).[3] When following a conservative strategy and adhering to a research tradition in their domain, scientists achieve publication with high probability: they remain visibly productive, but forgo opportunities for originality. When following a risk-taking strategy, scientists fail more frequently: they may appear unproductive for long periods, like the seven years Andrew Wiles spent proving Fermat's Last Theorem or the decade Frederick Sanger invested in developing the "Sanger method" of DNA sequencing.[4] If a risky project succeeds, however, it may have a profound impact, generating substantial new knowledge and winning broad acclaim (Kuhn 1962). This strategic tension is repeatedly articulated as a dichotomy: in the sociology of science, as reliable "succession" versus risky "subversion" (Bourdieu 1975) or "relevance" versus "originality" (Whitley 2000); in the philosophy of science, as "conformity" versus "dissent" or "discipline" versus "rebellion" (Polanyi 1969); and in the study of innovation, as "exploitation" versus "exploration" (March 1991).[5] Recent theoretical work supports this broad picture by highlighting the distinctive contributions (Weisberg and Muldoon 2009) and rewards (Kleinberg and Oren 2011) associated with traditional versus innovative strategies.

In this article, we explore the essential tension at scale, studying 6.5 million abstracts in biomedicine. We extend and elaborate Kuhn's account of the essential tension using Bourdieu's (1975, 1990, 2004) field theory. We argue that *tradition* and *innovation* label distinct regions in the space of possible research claims or "position-takings." Scientists anticipate a certain risk and reward profile from each region. Through their choice of research problem, scientists invest in a particular mixture of tradition and innovation.

Our empirical analysis follows six steps: (1) we show how networks can be used to map the evolving landscape of chemical knowledge in biomedicine; (2) we identify clusters of knowledge within that map and demonstrate their stability; (3) we define a simple structural typology of *research strategies* corresponding to tradition and innovation, and we show that the distribution of published strategies is remarkably stable; (4) we use a simple probabilistic model to measure the broad habits of perception, attention, and choice—the habitus—that constrain research activity and support stability; (5) we quantify the relationship between strategy, risks, and rewards (citations) using several regression models; and (6) we explore citation accumulation and awards as incentives that discipline scientists' choices and help structure the field. This matrix of incentives maintains the essential tension and is manifest in both the stable distribution of published research choices and the underlying habitus shaping them.

This article makes distinct contributions to the sociology of science, the study of networks, and quantitative methodology. It provides the first large-scale test of a key hypothesis in the sociology of science (the *essential tension*), drawing on a dataset of unusual size and quality to connect the structure of scientific content to multiple forms of reward, including citations and awards. We also make new connections between important although sometimes unfashionable streams of theory (Bourdieu 1975; Kuhn [1959] 1977), and extend Bourdieu's internal approach to the scientific field in new directions. Methodologically, we develop and deploy quantitative approaches for operationalizing key concepts in the sociology of science (e.g., position-taking, field, and habitus). Concretely, we demonstrate that contemporary network analytic methods (Rosvall and

Bergstrom 2008) can be extended beyond citations to map the structure of content (scientific and otherwise) at a very large scale. We also show how concepts from information theory (Cover and Thomas 1991) and techniques of probabilistic modeling (Cokol et al. 2005) can be used to quantify rich qualitative concepts like novelty and the scientific habitus. Above all, our core question—how do I choose what to study?—is of reflexive interest to all scholars and scientists.[6]

THE ESSENTIAL TENSION AND SCIENTIFIC HABITUS

Kuhn ([1959] 1977) introduced the notion of the essential tension at a conference motivated by Cold War concerns about declining originality, innovation, and scientific competitiveness in the United States (a persistent concern; see Cowen 2011). The conveners, all psychologists, had framed the conference around a dichotomy between *convergent* and *divergent* styles of thinking. Convergent thinking was conservative, oriented toward consensus and shared patterns of thought (Kuhn [1959] 1977). Divergent thinking, by contrast, was radical, characterized by "flexibility and open-mindedness" (Kuhn [1959] 1977:226). According to most of the conference speakers, divergent thought was essential for scientific progress, yet it was being stifled by the U.S. educational system.

Kuhn challenged the privileged role of divergent thinking in his remarks. In an argument prefiguring the better-known *Structure of Scientific Revolutions* (1962), Kuhn provided a functionalist account of the *interplay* between tradition (convergent thinking) and innovation (divergent thinking). Tradition focuses everyone on the same problems and methods and creates a well-defined community of practice. This particular constellation of problems and methods eventually exhausts itself, paving the way for revolution. In this way, convergent research within a tradition ushers in the next turn in an endless cycle of revolution and normal science (Kuhn [1959] 1977).[7] Kuhn ([1959] 1977:234) observed that "work within a well-defined and deeply ingrained tradition seems more productive of tradition-shattering novelties than work in which no similarly convergent standards are involved." This argument places Kuhn squarely within the functionalist paradigm of Merton's sociology of science; consider Merton's claim that original contributions are rewarded precisely because that is how knowledge grows (Guetzkow, Lamont, and Mallard 2004; Merton 1942).

For Kuhn ([1959] 1977:229), the primary mechanism that maintains tradition is education: the solution of textbook problems "that the profession has come to accept as paradigms." Although Kuhn ([1959] 1977:227) acknowledged that tradition is "reinforced by subsequent life in the profession," Kuhn's scientists are above all *trained* to work within a tradition—to ignore most of the anomalies churned up by daily work. At the same time, they are acutely aware that long-term reputation depends on novel results that lead them beyond tradition (Merton 1957), hence the tension. In the Kuhnian account, tradition is externally imposed on the practicing scientist and followed obstinately until it finally breaks down. Innovation is a *felix culpa,* a fortuitous accident that befalls a scientist with just the right sensitivity to distinguish trivial from illuminating anomalies.

A more convincing version of the essential tension is presented in Bourdieu's early writing on the sociology of science. Bourdieu (1975) frames science as a competitive field in which scientists face a *strategic choice* between "succession" and "subversion." These categories map directly onto Kuhn's tradition and innovation. Rather than offering a simple story of oversocialized scientists, in which tradition is inculcated and followed, Bourdieu emphasizes that tradition is continually re-created, and deviance punished, by the strategic choices of scientists regarding what to study and what to cite. Specific practices are patterned by the scientific habitus, which disposes researchers to perceive and act in ways "objectively adapted to their outcomes without presupposing a conscious aiming at

ends" (Bourdieu 1990:53). This habitus is produced by each scientist's education and her experiences doing science and seeing science done by others. The dispositions of the habitus, in turn, are oriented toward competition for peer recognition, the primary capital of science. Tradition is maintained insofar as the habitus disposes scientists to reproduce past research in their own studies and censor novelty in the work of others. Practices too far outside tradition are neglected, while continued investment in old questions enhances the scientific capital of those with stakes in them. At the same time, scientists able to stake new positions—to muscle research questions inside the bounds of legitimacy—can subvert tradition and receive outsized recognition. Not only are these scientists early investors in a new research enterprise; they have successfully exercised *symbolic power* in defining a new question, topic, or method as legitimate (Bourdieu 1991). A scientist's recognition scales in proportion to the extent of her subversion and the scientific redefinition it entailed.

Although the use of tradition or innovation strategies in pursuit of recognition should reflect the specificity of habitus, position, capital, opportunity, and risk, Bourdieu often suggests a "direct correspondence between an agent's field position—dominant vs. dominated—and that same agent's basic intellectual stance—orthodox vs. heterodox," that is, tradition versus innovation (Camic 2011:279). Yet Kuhn ([1959] 1977) persuasively argues that tradition and innovation exist in productive tension for *all* scientists. We agree, viewing the aspiration to innovate as constitutive of the scientific habitus, even if it is rarely manifested. Blending Kuhn and Bourdieu, we develop a strategic account of the essential tension. Tradition is not pursued purely because of training; it is a reliable strategy to accumulate recognition. Innovation is not a happy accident; it is a risky gamble.[8] Like Bourdieu, we embrace the multiscale nature of innovation, ranging from modest methodological advances to "tradition-shattering novelties" (Kuhn [1959] 1977:234) We expect that tradition and innovation will coexist (in tension) within fields, within scientists—even within papers.

Here we propose three ways in which strategic dispositions of the scientific habitus are formed and adapted to the accumulation of scientific capital; we will analyze all three quantitatively. First, *scientific publications* are the visible consequences of successful research choices. Scientists sample from that distribution in their reading and learn the prevalence of particular research strategies. This sampling also yields an informal estimate of the risk that a particular strategy will fail, yielding no publishable result and no peer recognition. Second, *citations* provide evidence of others' judgments about particular research choices (Baldi 1998; Latour 1987; Merton 1942, 1988). When scientists observe how others cite different research strategies, they form an impression of the rewards each provides. Subjective assessments of the variance in citations create another impression of risk, this time regarding the outcome of a strategy once published. Some papers are massively cited, some not at all. Finally, *awards* reveal the long-term consequences of particular research choices. Awards concretize the esteem that represents the highest reward in the field of science (Merton 1973). Hence, the behavior of award-winning scientists provides an aspirational model for the younger generation. In Bourdieu's (1986) framework, modest citations and major awards represent two institutionalized pathways to scientific capital, through which scientists learn the returns to tradition and innovation.

Stepping back from this theoretical account, let us paint a practical picture linking individual action to field structure: A scientist must decide what to work on next. Her position in the field is defined by her particular history, standing, relationships, and equipment (e.g., familiar chemicals, diseases, methods, and instruments). Based on this position, a number of possible research questions present themselves. These possible questions are colored by existing categories: practical or impractical, fundable or not

fundable, patentable or not patentable (Evans 2010b; Fleming and Sorenson 2004; Washburn 2005), and traditional or innovative. Most questions, however, will not come to mind, and some, like those deemed not fundable or too innovative,[9] may be ignored. Her categorizations and subsequent research choices are informed by expectations of outcome: How many other scientists pursue similar questions? How often do they fail? How often are innovative findings rejected (how often has she herself rejected them in the article review process)? If published, what is the reaction? Do they attract citations? Are they canonized with awards?

Research on the sociology, economics, and management of innovation independently grounds several aspects of this story. A wave of recent papers suggests that commercial opportunities (Evans 2010b), pressures (Vallas and Kleinman 2007), and commercially related policies (Berman 2012) can change the composition of scientific research and the choices that guide it.[10] Other work examines the influence of high position (e.g., conferred by awards) on the reception of a scientist's work (Azoulay, Stuart, and Wang 2013). Scholars have also explored the relationship between risk and reward. In the context of patents, Fleming (2001) demonstrates how risky combinations of patentable components are associated with higher variance and lower average citations. Uzzi and colleagues (2013) recently showed how a few atypical combinations of cited journals are associated with higher scientific impact, but only when contextualized with many typical journal pairings.

Management scholars Bateman and Hess (2015: Supporting Information p. 1) found that diabetes researchers judged a hypothetical "focused and specialized project" (i.e., tradition) less risky than a "broad project that spans several topical domains" (i.e., innovation). Surveyed researchers were less likely to pursue the risky project, viewing the specialized one as "potentially very important" and a more "significant opportunity." In short, diabetes researchers perceived innovation as risky and preferred to follow the more conservative course, as our theory suggests.

Finally, research has elucidated the role of awards in stimulating scientists to take risks (Wright 1983). Azoulay, Graff Zivin, and Manso (2011) recently found that prestigious, long-term Howard Hughes Medical Institute grants, which support investigators for at least five years, are associated with the propensity to publish more highly cited articles than investigators sustained by short-term NIH grants. Our article builds on this work by (1) integrating these insights and findings within the context of a substantial empirical case; and (2) linking them to the content of research papers, the topic of research choice, and a comprehensive theoretical framework.

The evidence and theory outlined here suggest that most published findings in well-developed fields should be *expected* and *unsurprising*.[11] Such findings fit with tradition: scientists with the appropriate habitus are disposed to generate and acknowledge them as valid science. By contrast, *unexpected* findings should rarely reach publication. Tradition will be reliably but modestly rewarded with citations, whereas subversive innovation, if published, should display a higher average and variance in acclaim.[12] Finally, unexpected findings should be overrepresented in the work of high-achieving scientists. Scientific capital is disproportionately awarded for work that alters the scope of accepted knowledge.

In this article, we examine these claims quantitatively, in the context of biomedicine. We focus on knowledge about chemical relationships in biomedicine; as a shorthand, we refer to this as "biomedical chemistry." This focus on chemical relationships might seem overly narrow. To the contrary, our chemical entities range from small molecules (like most pharmaceuticals) to large macromolecules (like DNA and proteins). Knowledge about these chemical relationships stretches from organic chemistry to biochemistry to molecular biology, and proxies for everything from assays to treatments to diseases. Biomedical papers bring together methods and diseases in addition to chemicals, but the reductive paradigm of contemporary biomedicine makes

chemicals an especially effective trace of its knowledge. A method often involves a chemical intervention; a disease usually has a chemical signature.

Knowledge Networks and Network Strategies

The document is the fundamental unit of analysis in most large-scale quantitative studies of scientific behavior (Cronin and Atkins 2000; Menczer 2004). Scholars use citations between documents to identify innovations (Chen et al. 2009; Funk and Owen-Smith 2012) and outline the structure of scientific fields (Rosvall and Bergstrom 2010). In our content-centered approach, basic conceptual entities and their relationships are the fundamental units of analysis. These entities and relationships weave a network of research possibilities, corresponding to "the space of search paths" or "network of possible wanderings" involved in human problem-solving (Newell and Simon 1972:82). In this picture, scientists both navigate the network of knowledge and build it through their research efforts. In the knowledge network of biomedical chemistry, the basic entities are molecules. The "relationships" between them can be reactions, interactions, or associations. These relationships inscribe the combinatorial "space of possible position-takings" or research questions. In pursuing a project, scientists take a stand on the existence of those relationships. When a paper is published, that stance becomes public and (minimally) legitimate. If the relationships in question are already established, tradition is reproduced, little reputation is staked, and little is gained. If published relationships are new, scientists are asserting scientific power and they accumulate reputation and scientific capital to the extent that others recognize the importance of those new relationships (Bourdieu 1990).

We propose the following coarse taxonomy of research strategies, corresponding to structurally distinct contributions to the network of scientific knowledge (Newman 2003). Our taxonomy builds on extensive work in science studies and research on innovation that uses the position of article and patent elements (e.g., citations, patent classes, and coarse topics) within a broader techno-scientific network as indicators of novelty and importance (Fleming, Mingo, and Chen 2007; Hargadon and Sutton 1997; Latour 1987; Leahey, Beckman, and Stanko 2013; Leahey and Moody 2014; Uzzi et al. 2013). A researcher may propose a relationship involving completely unexplored entities, making a *jump* beyond current knowledge (Cokol et al. 2005).[13] She may also test a relationship between previously explored entities, either asserting a *new* (not previously published) relationship or *repeating* a relationship proposed before. *New* and *repeat* relations, in turn, can be either *consolidations* or *bridges*. This distinction rests on the observation that recognized relationships are not distributed evenly across the knowledge network. Socially organized scientific attention creates dense clusters of related chemicals, mirroring scientific subfields. See the Data, Methods, and Results section for an explanation of how we identify these clusters, and Table 1 for details about the most prominent ones. Joining entities within the same cluster provides further *consolidation* of the cluster, deepening knowledge in that domain and tightening established categories. Linking entities from distinct knowledge clusters creates a *bridge* between them, altering the connectivity of the knowledge network and weaving loosely connected regions more tightly together. Figure 1 illustrates the five strategies available to a scientist facing a network of known scientific relationships: *jump, new consolidation, new bridge, repeat consolidation*, or *repeat bridge*. Note that *jump, new consolidation*, and *new bridge* correspond to varying degrees of innovation (adding new relationships). *Repeat consolidation* and *repeat bridge* correspond to tradition (reproducing established relationships). This taxonomy is especially well-suited to fields like biomedicine where well-defined entities participate in a "molecular" logic of combination. Here, innovation corresponds transparently to new combinations. Other fields may

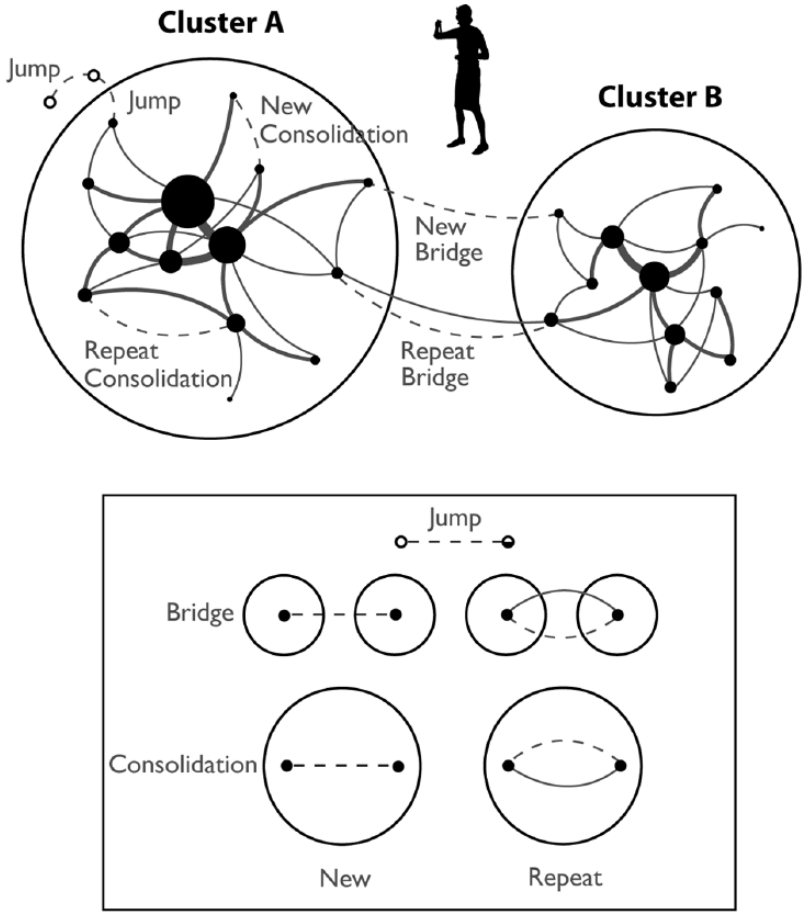

Figure 1. Scientific Strategies in a Network
Note: Nodes represent chemicals and links represent chemical relationships. The bottom panel shows our proposed taxonomy in schematic form. Labels apply to dashed links.

involve other logics, such as fractal differentiation in sociology (Abbott 2001) or subsumption of the particular to the general in mathematics (for other examples, see Lamont 2009). In such fields, tradition and innovation may require a different operationalization from that adopted here.

DATA, METHODS, AND RESULTS

Our analysis follows six distinct steps and connects to the vignette of our practicing scientist, described earlier. In brief, our broad goals are as follows: map the space of chemical knowledge (Steps 1 and 2); identify the prevalence and stability of the various strategies (Step 3); construct a simple behavioral model of scientific attention that relates strategy prevalence with opportunity (Step 4); measure risks and rewards associated with various strategies using citations (Step 5); and explore two potential motivational mechanisms for scientists' choices—maximizing citations and seeking outsized recognition (Step 6).

Note that our data reveal only the outcome of *successful* choices, that is, published papers. We have no way of accessing specific choices (which may fail) or specific dispositions (a problem shared by other field-theoretic analyses). As a result, our paper strictly

documents the *pattern* of successful choices and their association with various rewards. From this we make inferences about underlying dispositions and choices.[14]

Step 1: Chemistry Data and Network Construction

Our network is built from 6,455,756 abstracts in the NIH National Library of Medicine's (NLM) MEDLINE collection, published from 1934 to 2008 and annotated by NLM with two or more chemical entities (see Part A of the online supplement [http://asr.sagepub.com/supplemental]). Chemical annotations were introduced in 1980 and applied to all articles indexed subsequently.[15] We focus analysis on 1983 to 2008 to limit the effect of the introduction of chemical annotations on our results. We construct an evolving network (Newman 2003; Rosvall and Bergstrom 2010) from these annotated articles. For each year t, we examine all abstracts published in that year for chemical annotations. Each distinct chemical is a node in the network. Every time two chemicals appear in the same abstract in year t, we add a link between the corresponding nodes in the network for year $t + 1$ (Callon, Law, and Rip 1986). Note that articles are often annotated with more than two chemicals, and therefore contribute more than one chemical relationship. For example, an abstract annotated with five chemicals would correspond to an additional 10 links in the network. All relationships appearing in a given year enter the network together, as if published simultaneously. Coarse-graining the temporal evolution of the network in year-long chunks is an approximation, but necessary given uneven article date information and the computational demands of community detection.[16] Individual links between chemicals, as documented in a particular publication, persist in the network, although we relaxed this assumption in sensitivity analyses. If the same pair of chemicals appears together in multiple publications, each appearance is added to the network as a distinct, persistent link. This construction yields a time-ordered sequence of weighted networks over chemicals. Given the network of accumulated knowledge for year $t - 1$, we classify the links associated with each publication in year t as one of the five strategies. This classification allows us to count the total number of times we observe each strategy in a given year.

Our procedure maps the rapidly growing network of chemical knowledge—published tradition—atop the space of possible chemical relationships. The number of chemicals has grown steadily since annotations were introduced in 1980 (see Figure 3). By 2008, the network has 181,078 nodes and 10,493,379 distinct links—84,709,977 links including repeats. Like many complex networks, the network of chemical knowledge is characterized by heavy-tailed distributions of degrees (number of chemicals to which a given chemical is connected) and link weights (number of co-appearances of two chemicals) (Barabási and Albert 1999; Cokol, Rodriguez-Esteban, and Rzhetsky 2007; Onnela et al. 2007); see Figure S1 in the online supplement.

This first step provides us with a detailed *map of tradition* for any given year in contemporary biomedical chemistry (1983 to 2008). To establish the space of possible research questions facing a scientist in a given year, we need to *define knowledge clusters* within that map.

Step 2: Community Detection and Knowledge Clusters

The social organization of scientific attention means that published chemical relationships bunch into clusters of chemicals that are frequently investigated together. These knowledge clusters reflect foci of scientific attention. We identify knowledge clusters using a popular community detection algorithm called the map equation (Rosvall, Axelsson, and Bergstrom 2009; Rosvall and Bergstrom 2008). It is one of the few algorithms capable of analyzing a network of our size, and it performs well in benchmark tests (Lancichinetti and Fortunato 2009). Knowledge clusters reflect

the structure of chemical knowledge as expressed in journal articles: chemicals that appear together frequently in articles will be clustered together. We confirmed the robustness of cluster-dependent results against alternative subfield definitions induced by journal classifications and MEDLINE's MeSH ontology (see Part B and Table S1 in the online supplement).

Community detection algorithm. Community detection algorithms reveal intermediate structures in large, complex networks (Bruggeman, Traag, and Uitermark 2012; Shwed and Bearman 2010). What structure is relevant to our scientists? Imagine these scientists as they explore the network of chemical knowledge, wandering mentally from one chemical to another across previously published connections. An individual scientist's impression of the structure of chemical knowledge (tradition) in any given year can be coarsely approximated by the trajectory of a random walk on the knowledge network (see Austerweil, Abbott, and Griffiths 2012). *Knowledge clusters* are regions where the wandering process tends to get stuck before moving on.

The map equation provides a community detection scheme that precisely mirrors this picture (Rosvall et al. 2009; Rosvall and Bergstrom 2008). Formally, the algorithm minimizes the description length of a random walk on the network given a two-level labeling (higher-level labels for communities and lower-level labels for nodes within communities). Heuristically, the algorithm finds partitions such that the random walker spends a long time *within* a given community, on average, before transitioning to a new community. It thereby picks out subsets of nodes with dense intra-connection. This heuristic corresponds to the dynamic perception of subfields by scientists: as they explore the chemical knowledge network, they group together chemicals that frequently appear in the same papers via associative learning (Hebb 1949).

Knowledge clusters. To extract the knowledge clusters used to categorize links added at time t, we consider the *entire network* uncovered in all years prior to t. This choice for modeling temporality captures the gradual accumulation of chemical knowledge and is naturally conservative. While it intrinsically biases cluster detection toward stability, we view this as both desirable and realistic. Under this approach, knowledge clusters, which proxy for scientists' shared perception about the natural divisions of chemical knowledge, change substantively only in response to sustained changes in patterns of attention and published relationships. In the intervening period, the articles that drive change are likely to be perceived as altering the organization of chemical knowledge—and they are classified in exactly this way by our approach.[17]

Our community detection procedure discovers scientifically plausible knowledge clusters in this network. Given the network up to time t, we define knowledge clusters for time t by selecting the best partition (i.e., the one that minimizes the description length) out of 50 randomly seeded iterations of the map equation community detection algorithm. Multiple iterations are performed to avoid a local minimum in the solution space (Rosvall et al. 2009). We selected 50 as a practical number of iterations; more iterations would have been prohibitive computationally. The coherence of the clusters over time (Figure 2) and the face validity of clusters (Table 1) give us confidence that we are not in a local minimum of the partition-space.

On the whole, knowledge clusters are quite stable. To demonstrate this, we use a visualization technique developed by Rosvall and Bergstrom, the "alluvial diagram" (Bruggeman et al. 2012; Rosvall and Bergstrom 2010). This technique tracks nodes from their assigned knowledge clusters at time t to their locations at time $t+\Delta t$ and represents the movement via a ribbon. Figure 2 shows a sequence of visualizations covering the 15 largest clusters for the period 1983 to 2003. These represent the vast majority of flow in the network. Different shades of gray are assigned based on cluster membership in 2003 and follow chemicals back to earlier

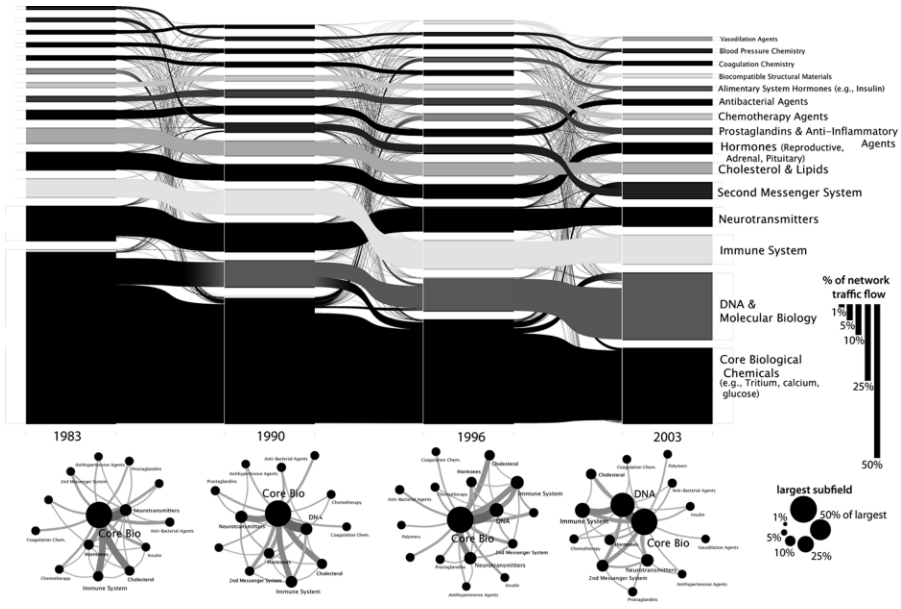

Figure 2. Largest MEDLINE Knowledge Clusters, 1983 to 2003
Source: Using the map equation algorithm and associated alluvial and network diagram software (Rosvall and Bergstrom 2008, 2010).
Note: Here we show an alluvial diagram for the chemical network at four time points: 1983, 1990, 1996, and 2003. Only the top 15 knowledge clusters are shown. Different shades of gray are assigned based on cluster membership in 2003 and follow chemicals back to the earlier networks. Clusters are quite stable. The major events visible are the emergence of a knowledge cluster associated with DNA and RNA (DNA and Molecular Biology) from Core Biological Chemistry and the growing importance of Second Messenger System. These shifts occur against a background of continuity. For example, the clusters Neurotransmitters and Hormones (Reproductive, Adrenal, Pituitary) persist with little change across 20 years. The network diagrams below each year emphasize the structural stability of the network; the links shown correspond to the highest flow between subfields.

networks. Most of the alluvial diagram consists of thick ribbons, illustrating the stability of these knowledge clusters. We further quantify stability as follows. For clusters that persist from time t to time $t+\Delta t$, we find the cluster in $t+\Delta t$ with the largest overlap and calculate the percent of chemicals in the cluster at time t that are clustered together in time $t+\Delta t$. In Table 1, we list these percentages for the clusters shown in Figure 2.[18] We also compute the average percentage over all clusters, weighted by the number of chemicals in the cluster. For clusters in 1983 and 1990, the average overlap is 73 percent; for 1990 and 1996, it is 76 percent; and for 1996 and 2003, it is 78 percent. These results demonstrate the stability of these knowledge clusters over long periods.

Knowledge clusters change over time as shifts in scientific attention and publication add new chemicals, move chemicals to another cluster, merge clusters together, or split them apart. Because we allow chemical knowledge to accumulate, large changes in the cluster structure reflect *sustained* changes in publication patterns. The major events visible in Figure 2 are meaningful: for example, the emergence of a knowledge cluster associated with DNA and RNA (DNA and Molecular Biology) from the large cluster labeled Core Biological Chemistry. Such shifts, however, occur against a background of continuity. For example, the clusters Neurotransmitters and Hormones (Reproductive, Adrenal, Pituitary), persist with little change across the decades, while Second Messenger System

grows in importance, corresponding to a well-established increase in research on intercellular communication.[19] The network diagrams beneath each year emphasize the structural stability of the network, with links corresponding to the highest flows between subfields.

Table 1 lists the top 10 chemicals (by PageRank) for each knowledge cluster in the 2003 network (shown in Figure 2). The coherence of these lists demonstrates the scientific plausibility of the knowledge clusters.[20] Similarly, the relationships and strategies identified by our network analysis are scientifically meaningful. For example, a 2007 article (Zemlyak et al. 2007) includes a *new consolidation* within a neurobiology-related knowledge cluster. This paper shows that neuronal tubulin-preferring agent NAPVSIPQ, a neuroprotective oligopeptide, defends neurons against neurotoxic kainic acid. Not only is the link between NAPVSIPQ and kainic acid a new and meaningful one, it also makes sense for both chemicals to be located in the same neurochemical cluster, and for their connection to be typed as a *new consolidation* of that cluster.[21]

Now that we have extracted the knowledge clusters for each year, we can fully define the *space of possible research questions* and sort them into our taxonomy of research strategies. In the third step, we explore the prevalence of each strategy by establishing how many published papers pursue questions in each category.

Step 3: Prevalence and Stability of Strategies

Strategy definitions. Our taxonomy of strategies, defined briefly earlier, is specified as follows:[22] *jumps* involve at least one chemical that joined the network in the current year *t*; *new consolidations* connect known chemicals from the same knowledge cluster with no link between them in year $t - 1$; *new bridges* connect known chemicals from *different* knowledge clusters with no link between them in year $t - 1$; *repeat consolidations* reconnect chemicals from the same cluster that already have a link between them in year $t - 1$; and *repeat bridges* reconnect chemicals from different clusters that already have a link between them in year $t - 1$. Using these strategy definitions, we assign a strategy type to every link that joins the network in year *t* and count the instances of each strategy.

Strategy frequency. Pooling counts across all years, we find that the frequency of each strategy in the published literature from 1983 to 2008 is inversely related to its plausible risk of failure, as expected for a mature science. *Repeat* strategies, which correspond to *tradition,* were six times more frequent than *new* or *jump* strategies, which correspond to *innovation* (85.8 versus 14.2 percent). *New bridges* and *new consolidations* were more common than *jumps* by roughly the same proportion, tracking more and less extreme forms of innovation (12.4 versus 1.8 percent). Strategies that build incrementally on past knowledge appear more frequently not only because they are pursued more frequently, but also because non-incremental strategies are prone to fail. Whether a research project does not work or the resulting publication is censored by peer review, the results remain unpublished and invisible to science.

Yearly strategy distributions. We refer to the relative frequency of strategies within a year as the strategy distribution (see Figure 3A, solid lines). To assign confidence intervals to the relative frequencies, we model them as draws from a multinomial distribution with five types. We use the confidence interval proposed in Quesenberry and Hurst (1964) with the tighter bound introduced by Goodman (1965). The resulting distribution of scientific strategies remained remarkably stable over the period studied. The black arrow indicates the introduction of chemical annotations in 1980. From 1983 to 2008, the fraction of annual links corresponding to innovation strategies changed little, with *jumps* most rare, followed by *consolidations,* then *bridges*.

The observed prevalence of each research strategy reveals aspects distinctive to the

Table 1. Largest Chemical Clusters Induced by Map Equation from the MEDLINE Chemical Term Network

Title	Percentage of Chemicals in the Cluster at the Beginning of the Period Also Present at the End		
	1983 to 1990	1990 to 1996	1996 to 2003
Core Biological Chemicals	60%	69%	81%
	Glucose, Amino Acids, Peptides, Adenosine Triphosphate, Potassium, Sodium, Tritium, Isoenzymes, Magnesium, Carbon Isotopes		
DNA and Molecular Biology	n/a	57%	85%
	Messenger RNA, DNA, DNA-Binding Proteins, Recombinant Proteins, Proteins, Transcription Factors, Carrier Proteins, Membrane Proteins, Bacterial Proteins, Peptide Fragments		
Immune System	87%	77%	75%
	Monoclonal Antibodies, Membrane Glycoproteins, Immunoglobulin G, Tumor Necrosis Factor-alpha, Glycoproteins, Antibodies, CD Antigens, Epitopes, Cytokines, Lipopolysaccharides		
Neurotransmitters	67%	87%	86%
	Norepinephrine, Serotonin, Epinephrine, Acetylcholine, Isoproterenol, Pyridines, Propanolol, Histamine, Catecholamines, Piperazines		
Second Messenger System	66%	49%	65%
	Calcium, Enzyme Inhibitors, Cyclic AMP, Protein Kinase C, Tetradecanoylphorbal Acetate, GTP-Binding Proteins, Indoles, Calcium Channel Blockers, Adenylate Cyclase, Ion Channels		
Cholesterol and Lipids	82%	80%	87%
	Cholesterol, Lipids, Phospholipids, Fatty Acids, Triglycerides, Lipoproteins, Phosphatidylcholines, Liposomes, Nonsterified Fatty Acids, Dietary Fats		
Hormones (Reproductive, Adrenal, Pituitary)	93%	94%	87%
	Estradiol, Hydrocortisone, Progesterone, Testosterone, Luteinizing Hormone, Dexamethasone, Adrenocorticotropic Hormone, Prolactin, Follicle Stimulating Hormone, Estrogens		
Prostaglandins and Anti-inflammatory Agents	83%	82%	83%
	Indomethacin, Non-Steroidal Anti-Inflammatory Agents, Dinoprostone, Prostaglandins, Sulfonamides, Arachidonic Acid, Aspirin, Arachidonic Acids, Prostaglandin-Endoperoxide Synthases		
Chemotherapy Agents	84%	92%	92%
	Antineoplastic Agents, Cyclophosphamide, Doxorubicin, Cisplatin, Methotrexate, Prednisone, Vincristine, Fluorouracil, Antineoplastic Antibiotics, Phytogenic Antineoplastic Agents		
Anti-bacterial Agents	84%	90%	89%
	Anti-Bacterial Agents, Drug Combinations, Penicillins, Anti-Infective Agents, Chloramphenicol, Quinolines, Tetracycline, Streptomycin, Cephalosporins, Gentamicins		
Alimentary System Hormones (Insulin, etc.)	n/a	n/a	59%
	Insulin, Blood Glucose, Growth Hormone, Glucagon, Insulin-Like Growth Factor I, Hypoglycemic Agents, Somatostatin, Insulin Receptor, Glycosylated Hemoglobin A, Somatomedins		
Biocompatible Structural Materials	86%	94%	94%
	Polymers, Biocompatible Materials, Artificial Membranes, Silicon Dioxide, Composite Resins, Methacrylates, Acrylic Resins, Titanium, Hydroxyapatites, Silver		
Coagulation Chemistry	81%	86%	88%
	Heparin, Thrombin, Fibrinogen, Anticoagulants, Platelet Aggregation Inhibitors, Fibrinolytic Agents, Tissue Plasminogen Activator, Fibrin, Urokinase-Type Plasminogen Activator, Blood Coagulation Factors		
Blood Pressure Chemistry	76%	72%	69%
	Imidazoles, Angiotensin II, Antihypertensive Agents, Renin, Aldosterone, Diuretics, Angiotensin-Converting Enzyme Inhibitors, Bradykinin, Angiotensin Receptors, Benzimidazoles		
Vasodilation Agents	n/a	n/a	79%
	Nitric Oxide, Nitric Oxide Synthase, Vasodilator Agents, Nitrates, Nitric Oxide Synthase Type II, NG-Nitroarginine Methyl Ester, Nitroprusside, Nitrites, Guanylate Cyclase, Penicillamine		

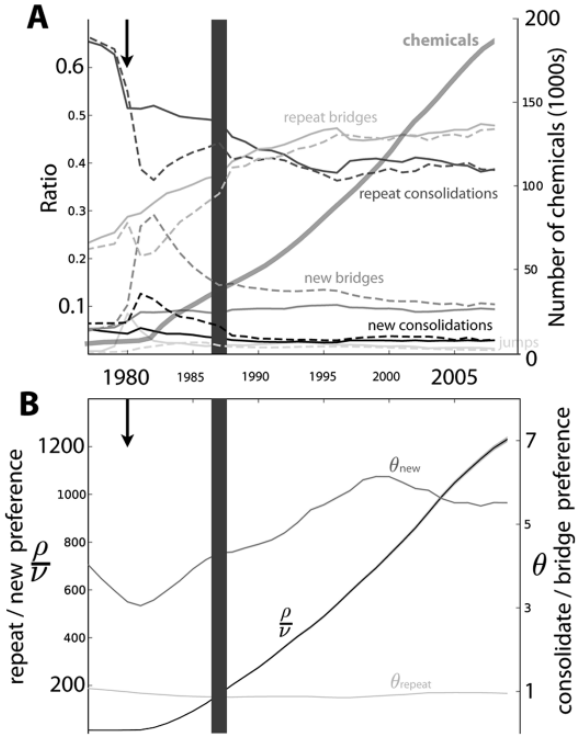

Figure 3. Stable Strategies and Dynamical Attention
Note: (A) The empirical frequency of each strategy (solid line) with 95 percent confidence intervals smaller than the solid lines. Dotted lines show the predictions of our generative model. In 1980 (black arrow) chemical annotation is introduced in MEDLINE (see Part A of the online supplement). This distorts parameter estimates until 1987 (parameters for year *t* inferred from the six previous years). (B, left axis) The ratio ρ/ν indicates an increasing concentration on established knowledge relative to new possibilities. (B, right axis) The preference for new consolidations over new bridges θ_{new} grows and stabilizes. This preference does not carry over to repeat work θ_{repeat}. 95 percent Bayesian credible intervals, shown in lighter shades of gray, are very small.

development of biomedical chemistry. The discovery of new, important chemicals (jumping) is particularly uncommon and risky, but it represents the only mechanism by which the chemical universe can grow. The rarity of new consolidations may result from dense "presumed" knowledge within a subfield, especially *negative* knowledge. For example, many possible combinations are viewed as fruitless because they were previously attempted without success (Swanson 1990). The number of legitimate, unpublished consolidating links is relatively small, but these will tend to be surprising and consequential for the field. The higher frequency of new bridges, by contrast, is driven by combinatorial possibilities. As clusters multiply and shrink in size relative to the whole network, more opportunities arise to make links between versus within knowledge clusters.

Linear regressions of frequency on year show no significant temporal trend for new bridges, and significant but slight trends for jumps and new consolidations; their shifts each alter the mix of strategies less than 1 percent over the study period (see Part C and Table S2 in the online supplement). Prais-Winsten estimation, which takes autocorrelation into account, eliminated the detection of significant trends in jumps, new consolidations, and new bridges ($p > .05$), confirming the stability of the distribution of these strategies. We find mild, statistically significant trends in the repeat strategies, with repeat bridges passing repeat consolidations in 1990 and remaining stable thereafter.

This stability in the strategy distribution is surprising given the enormous changes in biomedical chemistry over the past few decades. The number of distinct chemicals that appear in MEDLINE annotations increased by more than an order of magnitude between 1980 and 2008 (see Figure 3). Because of this, there are many more *new* links to explore than *known* links (a factor of 22 more in 1983, growing to a factor of 188 by 2008). Because knowledge clusters are small relative to the entire system (the largest shrinks from 27.5 percent of the network in 1983 to 12.8 percent in 2008), the number of possible bridging links grows much faster than the number of possible consolidating links. The stability of strategic choices suggests that scientists are not responding to the changing space of research possibility. This excludes a simple Mertonian account of research choice, in which originality is the prime requisite of reward, and choice is determined largely by the opportunities that nature provides. Even if an overwhelming fraction of potential new links can be discarded as physically impossible or otherwise unworthy of exploration, biomedical scientists are not taking full advantage of opportunities to explore new relationships between chemicals.

To explore this further, in the next section we develop a simple generative behavioral model that relates the *opportunities* to pursue each strategy to the *observed* distribution of strategies. This model provides a coarse estimate of scientists' *preference for tradition over innovation,* a core disposition of the biomedical habitus.

Step 4: Connecting Stability to Attention; A Generative Model of the Biomedical Habitus

The stability of the strategy distribution, despite dramatic expansions in research opportunity, implies that scientific attention is narrowing in focus. Indeed, both Kuhnian and Bourdieusian theories of the essential tension suggest potential narrowing mechanisms: for Kuhn, the paradigms provided by research training; for Bourdieu, stabilization of a field-specific (in this case, biomedical) habitus. Under both mechanisms, some opportunities will be discounted relative to others. A simple behavioral model of strategy selection allows us to measure this narrowing of scientific attention, connecting individual choices to aggregate stability. Our model takes the distribution of strategic opportunities into account and thus allows us to infer strategic preferences embodied in the biomedical habitus.

We assume that scientists choose probabilistically from the five possible strategies: *jump, new consolidation, new bridge, repeat consolidation,* and *repeat bridge*. For the purposes of this model, we treat *each* instance of strategy choice as an independent draw.[23] If $P(jump)$ is the probability of choosing a jump strategy, then

$$P(jump) + P(new, consolidation) + P(new, bridge) + P(repeat, consolidation) + P(repeat, bridge) = 1$$

as the five strategies partition the sample space. It is plausible (and consistent with our motivating theory) that the distinction between jump, new, and repeat is most salient to choosing scientists. This distinction maps directly onto extreme innovation, innovation, or tradition, the categories most relevant to risk (and, as we show in Steps 5 and 6, to reward). We therefore model choice as a two-step process, in which researchers first choose between *jump, new,* or *repeat* and then choose to *consolidate* or *bridge*[24] between subfields. Under these assumptions, the probabilities factorize:

$$P(j) + P(n)\cdot P(c\,|\,n) + P(n)\cdot P(b\,|\,n) + P(r)\cdot P(c\,|\,r) + P(r)\cdot P(b\,|\,r) = P(j) + P(n)\cdot(P(c\,|\,n) + P(b\,|\,n)) + P(r)\cdot(P(c\,|\,r) + P(b\,|\,r)) = 1$$

Once a scientist chooses new or repeat, she must either consolidate or bridge conditional on that choice:

$$P(c\,|\,n) + P(b\,|\,n) = P(c\,|\,r) + P(b\,|\,r) = 1$$

Substituting this into the previous equation, we have the following:

$$P(j) + P(n) + P(r) = 1$$

In other words, extreme innovation, innovation, and tradition provide a coarse partition of the sample space. The partition into jump, new consolidation, new bridge, repeat consolidation, and repeat bridge can be viewed as a fine partition of the sample space.

Researchers, here modeled by a representative agent,[25] independently and randomly choose a strategy at time *t* based on the number of possible links in the network corresponding to each strategy at that time. This number is then weighted according to a *bias parameter* that coarsely summarizes the factors influencing this decision. Let J_t be the number of potential jumps, C_t the number of potential new consolidations, B_t the number of potential new bridges, C'_t the number of repeat consolidations, and B'_t the number of repeat bridges. $R_t = C'_t + B'_t$ is thus the number of links that could be repeated and $N_t = C_t + B_t$ is the number of potential new links. Repeat variables track the *total* number of repeat links.[26] This accounting scheme is consistent with our probabilistic framework: the likelihood that a researcher encounters the opportunity for a repeated link is proportional to its number of repetitions. The probability of each strategy is as follows:

$$P(j) = \frac{J_t}{\nu N_t + \rho R_t + J_t}$$

$$P(n,c) = P(n) \cdot P(c \mid n) = \frac{\nu N_t}{\nu N_t + \rho R_t + J_t} \cdot \frac{\theta_n C_t}{\theta_n C_t + B_t}$$

$$P(n,b) = P(n) \cdot P(b \mid n) = \frac{\nu N_t}{\nu N_t + \rho R_t + J_t} \cdot \frac{B_t}{\theta_n C_t + B_t}$$

$$P(r,c) = P(r) \cdot P(c \mid r) = \frac{\rho R_t}{\nu N_t + \rho R_t + J_t} \cdot \frac{\theta_r C'_t}{\theta_r C'_t + B'_t}$$

$$P(r,b) = P(r) \cdot P(b \mid r) = \frac{\rho R_t}{\nu N_t + \rho R_t + J_t} \cdot \frac{B'_t}{\theta_r C'_t + B'_t}$$

where ν is the bias for new relationships, ρ the bias for repeats, θ_n for new consolidation, and θ_r for repeat consolidation. We fix the bias for jumps at 1 and calculate these other parameters relative to it.[27]

Conceptually, we imagine that the agent repeats this procedure until it has made the same number of choices as we observe in a given year. Using the aggregate history of strategic choices discussed earlier, we infer bias parameters that maximize the likelihood our agent will generate the observed history. The likelihood of all publication data observed between years t_0 and t_f under this model is as follows:

$$L(data[t_0, t_f] \mid \nu, \rho, \theta_n, \theta_r) =$$

$$\prod_{t=t_0}^{t=t_f} L(data[t]) = \prod_{t=t_0}^{t=t_f} P(j)^{\eta_J(t)}\, P(n,c)^{\eta_C(t)}$$

$$P(n,b)^{\eta_B(t)}\, P(r,c)^{\eta_{C'}(t)}\, P(r,b)^{\eta_{B'}(t)}$$

The likelihood function depends on the bias parameters through the strategy probabilities (see previous equation). Parameter values are obtained by maximum likelihood estimation. To estimate the parameters for a given year, we use the observed number of links of each type (e.g., $\eta_J(t)$ = the number of *jumps* in year *t*) from the six previous years as the data in the likelihood function; see Part D of the online supplement for details.[28] Estimated bias parameters provide an average description of the preferences held by the scientific community for each type of strategic opportunity, relative to jumping.[29] We assign 95 percent Bayesian credible intervals to the parameter point estimates by Monte Carlo sampling. We performed 200,000 iterations for each parameter estimate and constructed intervals in Figure 3B as conservatively as possible (see Part D of the online supplement).

Figure 3A (dashed line) shows that this behavioral model predicts observed behavior reasonably well, with a high correlation between known and predicted values (Pearson's R = .983). The thick vertical line in Figures 3A and 3B indicates the point (1987) at which parameter estimates are no longer strongly affected by the introduction of annotations in 1980. Significant trends in parameters

(Figure 3B) suggest that scientists filter out more new opportunities and become more locally focused as knowledge grows. The left axis of Figure 3B indicates a sharply increasing preference for repeating known links over exploring new ones. The right axis of Figure 3B shows that the preference for new consolidation over bridging has also grown, and may be leveling off, indicating a local focus in the exploration of new relationships. In other words, biomedical chemistry remains in a stable, normal science regime precisely because opportunities for novelty are persistently and increasingly ignored.[30] The biomedical habitus, crudely approximated by these parameters, sustains tradition at the expense of innovation.

In our Bourdieusian account of scientific choice, the relevant dispositions must be shaped by more than training or observed strategy prevalence. They must reflect and reinforce distributions of risk and reward. In the next section, we *relate strategies to the risks and rewards* revealed by the first and second moments of the associated citation distribution. The mean number of citations defines the expected reward for a strategy if published. Standard deviation in citations traces the uncertainty in this reward, conditional on publication. We perform this analysis by applying regression models to pooled and disaggregated, article-level data.

Step 5: Measuring the Relationship between Strategy, Risk, and Reward

We now connect the disposition to pursue tradition (*repeat* links) over innovation (*new* and *jump* links) to the reward mechanisms that drive scientists in their pursuit of recognition. We conjecture that innovation strategies, which are rare and risky, should be more highly rewarded, on average, but face greater uncertainty in reward. In this section, we focus on citations as a form of reward and recognition.

Publication-level strategies. Careful analysis of the relationship between strategy, risk, and reward demands a taxonomy that classifies each *publication* uniquely. Unique classification assigns a given paper's citations to only one strategy. Our earlier taxonomy works at the level of individual links. We thus construct a taxonomy at the publication level, based on the link-level strategies deployed. A *jump* publication has at least one *jump* link, the least frequent type. A *new consolidation* publication has at least one *new consolidation* link, the next least frequent, but no *jump* links. A *new bridge* publication has at least one *new bridge,* but no *jumps* or *new consolidations,* and so on through *repeat consolidations* and *repeat bridges.* Each publication is thus defined as the outcome of its most distinguishing (and surprising) strategy. Call this first taxonomy "fine-grained." As in the generative model, we also create a "coarse-grained" taxonomy, separating articles into *jump, new,* and *repeat* publications. A *jump* publication is defined as above; a *new* publication has at least one *new* link but no *jumps*; and a *repeat* publication has no *jumps* or *new* links.[31]

Surprisal. In information theory, self-information or *surprisal* measures the information associated with observing outcome i of a discrete random variable (Cover and Thomas 1991). Surprisal is defined as $I(p_i) = \log(1/p_i) = -\log(p_i)$, such that highly improbable outcomes are surprising (high surprisal) and more commonplace outcomes are unsurprising (low surprisal). In our case, the outcomes are observations of a given research strategy. Observations of rare strategies are more surprising and therefore more novel. Strategy surprisal can be defined for the publication-level taxonomy using the relative frequency of the distinctive or eponymous strategy in the prior year as an estimate for its probability in the current year—$P_t(strategy)$. For example, the surprisal of a *new bridge* publication in year t would be

$$-\log\left[\frac{\eta_B(t-1)}{\eta_J(t-1)+\eta_C(t-1)+\eta_B(t-1)+\eta_{C'}(t-1)+\eta_{B'}(t-1)}\right].$$

Citation counts. Citations are assigned to abstracts by linking MEDLINE abstracts to the ThomsonReuters *Web of Science* citation

database. Each abstract is assigned the total citations it received in the three years following initial publication (the median half-life for scientific citations); this effectively implements a medium-term time horizon on scientists' assessment of reward.[32] If an abstract is not in the *Web of Science,* it is omitted from this analysis. Approximately two-thirds of the MEDLINE abstracts link to a *Web of Science* record with citation information. We restrict analysis to 1983 to 2002 due to decreased citation coverage post 2005.[33]

Aggregate analysis: strategy-year models. In these models, we pool all abstracts corresponding to each publication-level strategy in a given year and compute the mean citations and the standard deviation in citations for that pool. We then perform robust regressions (HC3 heteroscedasticity correction) in Stata. For the data partitioned into a fine taxonomy of five strategies (*jump, new consolidation, new bridge, repeat consolidation, repeat bridge*), we predict the mean and standard deviation of citations using two different models: strategy surprisal and year in one model, and strategy type and year in another.[34] For the coarse taxonomy with three strategies (*jump, new, repeat*) we predict mean citations and standard deviation in citations with strategy type and year.

Aggregate results. Results from these models confirm our conjecture that strategy surprisal positively varies with mean citations, conditional on that strategy being published. Rare *innovation* strategies have high surprisal. They likely have a harder time being published but garner more attention when they are. Common *tradition* strategies have low surprisal. Robust regression shows that surprisal alone explains 49 percent of the variation in mean citations; a model including year explains 79 percent. This pattern is also observed when surprisal is replaced with strategy type, treated as an indicator variable. In the five-strategy (fine-grained) model, strategy type alone explains 50 percent of the variation, and the regression coefficients become larger as the strategy becomes more innovative, exactly as predicted. When year is added, we explain 81 percent. Results are very similar for the three-strategy (coarse-grained) model. Strategy type and year together explain 74 percent of the variation.

We also confirm our conjecture that surprisal is positively correlated with the uncertainty of reward, which we operationalized as the standard deviation in citations. The same results obtain when strategy type is used, with the rare, risky strategies again having larger regression coefficients (see Table 2).[35] Figure 4A provides a simple visualization of the relationship between tradition, innovation, and mean citations.

Article-level analysis. Because each article can be associated with its distinguishing strategy, we can also test the association between strategies and citations on disaggregated article-level data, accounting directly for overdispersion in citations (article level of analysis in Table 2). Citation values are discrete and typically have a broad distribution, with some articles receiving very high citation counts. Indeed, many more highly cited articles occur than would be the case for Poisson-distributed data. This overdispersed behavior defies assumptions of the standard linear regression model, which is why we used a pooling strategy for the OLS analysis. For some model specifications, we can estimate the relationship between citations and research strategies using a negative binomial model, a generalization of the Poisson model that accounts for overdispersion in the data.[36] Like Poisson models (Hausman, Hall, and Griliches 1984), negative binomial models assume that the logarithm of the expected value of the dependent variable can be modeled by a linear combination of known predictors. In this sense, it is similar to estimating a simple linear regression with the logarithm of citations as the dependent variable.

To test the effect on citations of strategy type, we assume the following:

$$Citations_a \sim NegativeBinomial(\mu_a, \alpha)$$

Table 2. Models Regressing Citations on Strategy Surprisal or Type

Level of Analysis:	Strategy-Year				Article	
Model/Estimation:	Robust HC3				NBREG	
Dependent Variable:	Mean Citations		SD Citations		Citations	
	β	(SE)	β	(SE)	IRR	(SE)
Strategy Surprisal Model						
Strategy Surprisal	1.446	(.099)	2.619	(.379)	1.143	(.0009)†
Year	.252	(.022)	.365	(.072)	1.022	(.0002)
Constant	4.179	(.323)	11.682	(.921)	5.921	(.013)
R^2 (observations)	.785	(100)	.471	(100)	.003‡	(2975434)
Five-Strategy Model						
Strategies (comparison: repeat bridge)						
Repeat consolidation	1.212	(.187)	2.223*	(.922)	1.165	(.003)
New bridge	2.817	(.218)	2.800**	(1.049)	1.381	(.004)
New consolidation	4.492	(.235)	7.268	(1.113)	1.630	(.006)
Jump	5.029	(.556)	10.002	(1.917)	1.720	(.009)
Year	.261	(.022)	.380	(.073)	1.023	(.0002)
Constant	4.706	(.258)	13.087	(.970)	5.753	(.018)
R^2 (observations)	.812	(100)	.528	(100)	.003‡	(2975434)
Three-Strategy Model						
Strategies (comparison: repeat)						
New	2.437	(.180)	2.704	(.612)	1.300	(.002)
Jump	4.022	(.550)	8.062	(1.728)	1.510	(.007)
Year	.277	(.030)	.423	(.106)	1.020	(.0002)
Constant	5.554	(.370)	14.616	(1.015)	6.550	(.012)
R^2 (observations)	.744	(60)	.475	(60)	.003‡	(2975434)

$$\text{Surprisal of } strategy \text{ in } t = -\log\left[P_t(strategy)\right] = -\log\left[\frac{\eta_{strategy}(t-1)}{\sum_{strategies}\eta_{strategy}(t-1)}\right]$$

Note:
‡Pseudo-R^2 for negative binomial regression models.
†Standard errors for the non-exponentiated estimates (e.g., the natural logarithm of the presented IRR).
*$p < .05$; **$p < .01$; all other results $p < .001$. Model significance: F-test; coefficient significance: two-tailed t-tests. Coefficients are unstandardized.

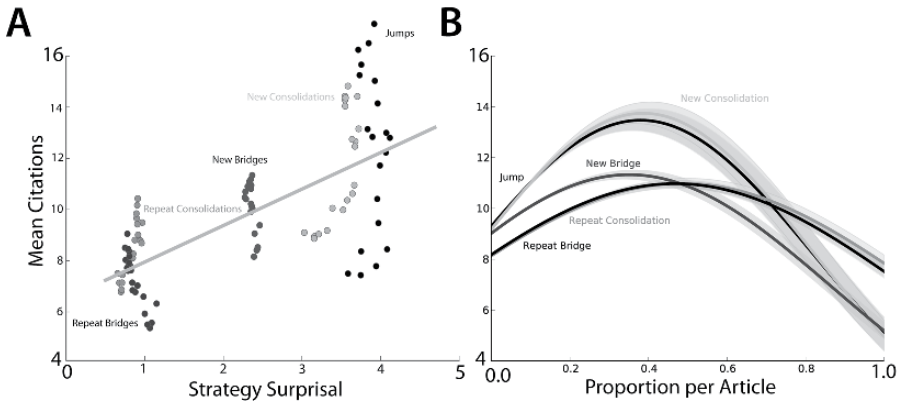

Figure 4. Strategies and Citation Impact
Note: (A) Rare innovation strategies are correlated with higher mean citations (49 percent of variation explained). The surprisal of a strategy in year *t* (horizontal axis) plotted against mean citations received by papers published in year *t* distinguished by that strategy. Much of the remaining variation is explained by adding year of observation to the model, as citations tend to increase with time (79 percent of variation explained by the combined model; the surprisal coefficient remains significant). (B) The mean citations for a paper depend quadratically on the fraction of links in the paper that correspond to a given strategy. Here we show the results of a negative binomial regression model; the peak occurs at around 40 percent for jump, new consolidation, and new bridge. Note that the response to strategy fraction is strongest for rare innovation strategies.

where $Citations_a$ is the number of citations received by article a in the year a was published and the subsequent three years. We consider two classes of model. For the five-strategy (fine-grained) taxonomy, we construct models using either surprisal or strategy type as a predictor, along with year since 1983. In the first case, the model is

$$\log(\mu_a) = \beta_0 + surprisal_a\beta_1 + year_a\beta_2 + \varepsilon_a$$

where strategy surprisal is assigned as above. In the second case, the model is

$$\log(\mu_a) = \beta_0 + jump_a\beta_1 + new,con_a\beta_2 + new,bri_a\beta_3 + repeat,con_a\beta_4 + year_a\beta_5 + \varepsilon_a$$

where $jump_a$, new,con_a, and such are indicator variables; $year_a$ is a continuous predictor; and ε_a, the error, is assumed to be a gamma distributed random variable. This formulation makes *repeat bridge* a reference category. For the three-strategy (coarse-grained) taxonomy, we use strategy type and year as predictors:

$$\log(\mu_a) = \beta_0 + jump_a\beta_1 + new_a\beta_2 + year_a\beta_3 + \varepsilon_a$$

where $jump_a$ and new_a are indicator variables and $year_a$ and ε_a are treated as before. In this model, *repeat* is used as the reference category.

Article-level results. Our results for this analysis are summarized in Table 2. All coefficients are highly significant. Once again, these models reveal higher average returns to rare, risky, innovation strategies, as predicted by our hypothesis connecting rarity, risk, and reward. In the three-strategy model, for example, a paper deploying a *new* strategy has 30 percent more citations, on average, than a paper studying only *repeat* links. A paper deploying a *jump* strategy has 51 percent more citations, on average, than a more conservative *repeat* paper.[37]

Optimal strategy analysis. We also use negative binomial regressions to test a related hypothesis, connecting expected citations to the fraction of links in a given article that correspond to a particular link-level strategy. The interplay of strategy, risk, and reward

suggests that articles should be penalized for myopically focusing on one strategy and excluding all others. For example, it is difficult to locate a paper with no *repeat* links in the existing knowledge network. Such a paper is a radical break with tradition and is likely to be misunderstood or ignored. Likewise, a paper with no *new* or *jump* links is at lower risk of being misunderstood, but also at lower risk of being a subversive breakthrough with corresponding high citations. This analysis of the optimal mix of strategies within an article connects directly to recent findings in Uzzi and colleagues (2013).

We use negative binomial regressions to test the connection between expected citations and the fraction of links in a given paper that correspond to a particular link-level strategy. We assume that citations are distributed according to a negative binomial distribution and fit the following model:

$$\log(\mu_a) = \beta_0 + f(jumps_a)\beta_1 + f(jumps_a)^2\beta_2 + \varepsilon_a$$

where $f(jumps_a)$ refers to the proportion of chemical relationships in article *a* that represent jumps; $f(jumps_a)^2$ is the same quantity squared; and ε_a, the error, is again assumed to be a gamma distributed random variable. We estimated this model for every other type of relationship (new consolidations, new bridges, repeat consolidations, and repeat bridges) in Stata.

Optimal strategy results. Negative binomial regressions of expected article citations on strategy fraction and fraction-squared bear out our intuition that there should be a robust unimodular relationship between strategy proportion and expected citations (see Part F and Table S4 in the online supplement); Figure 4B depicts that relationship. We find that research attention in the form of citations responds most strongly to variation in strategy proportion for innovation strategies, with *pure innovation* (all links are *jump, new consolidation,* or *new bridge*) highly penalized but modest fractions (around 40 percent) highly rewarded. The citation-optimizing configurations can be explained in the typical case of five chemicals per publication. By adding a single new chemical to four previously connected chemicals from the same subfield, a paper can approach the optimal level: 40 percent *jump* links and 60 percent *repeat consolidation* links. Similarly, by newly linking a single known chemical in one knowledge cluster to four connected chemicals in another, a paper can achieve the near optimum: 40 percent *new bridge* links and 60 percent *repeat consolidation* links.

At first blush, these proportions appear to disagree with Uzzi and colleagues' (2013) recent finding that peak impact is associated with articles in the 85th to 95th percentile of median conventionality in their references, implying an overwhelming number of repeat reference combinations. The difference between 60 percent conventional chemical combinations and 90 percent conventional journal combinations is easy to resolve, however, if authors cite familiar literature concerning repeat chemical associations more often than the unfamiliar literature used to justify novel associations.[38]

We have demonstrated that innovation is more richly rewarded with citations than is tradition. Citations represent one of the fundamental currencies of scientific recognition, so we might reasonably ask *whether observed levels of innovation could be motivated purely by citation accumulation,* or if additional motivations are needed.[39] We take this up in the last empirical section.

Step 6: Analyzing Possible Motivations

Maximizing citations. We verified our conjecture that innovation strategies should be more highly rewarded, on average, than tradition strategies. We also found that they face greater variability in reward following publication (although not a greater risk of being ignored; see note 37). We now consider whether these risk-reward relationships are sufficient to explain the observed distribution of strategies, with innovation (*jump, new*) rare and tradition (*repeat*) common. The simplest strategic model of scientific motivation

assumes that scientists are trying to maximize citations on a paper-by-paper basis to accumulate scientific capital (Bourdieu 1975). This corresponds to the picture of professional pressure described earlier. To maintain their current position in the scientific field, scientists must reliably publish articles and garner citations for those articles.

This analysis requires that we estimate the chance of utter failure, in which nothing is found or the resulting paper fails to pass peer review. Do innovation strategies attract enough reward to balance this risk? In other words, is a scientist who chooses to innovate pursuing a reasonable strategy, if her goal is to maximize expected citations for that project? We cannot directly observe the probability that a particular strategy will succeed or fail from our data: our sample is drawn only from projects that *did* produce a publication. We can, however, provide bounds on the failure probability, under the assumption that scientists select strategies only to maximize citations (Bourdieu 1975).[40]

To analyze whether citation maximization is a credible motivation for the observed distribution of scientists' research choices, we again study paper-level strategies, now taking into account the *a priori* probability that a particular strategy yields a publication. The expected citation count for a particular strategy E(*citations* | *S*) is the product of the probability of success given that strategy, Pr(*publication* | *S*), with the mean citations for published articles employing that strategy—the expected reward conditional on publication, R(*citations* | *publication, S*). Note that Pr(*publication* | *S*) incorporates both *actual* failure, in which a project goes nowhere, and *publication* failure, where the resulting article is censored by peer review.[41] Our use of mean citations (Wuchty et al. 2007) reflects the simplest possible utility function for scientists with no risk-aversion or risk-seeking.

For a risk-neutral, citation-maximizing scientist to be indifferent between strategies—for the choice of *new* or *jump* instead of *repeat* to be rational—then E(*citations* | *repeat*) = E(*citations* | *new*) = E(*citations* | *jump*). This equality allows us to place bounds on the probability of success Pr(*publication* | *S*) for each strategy. We compute mean citations from the data, averaging over the period 1983 to 2002:[42] 8.38 for *repeat*, 11.00 for *new*, and 12.90 for *jump*. For the equality to hold, Pr(*publication* | *new*) = .76 × Pr(*publication* | *repeat*) and Pr(*publication* | *jump*) = .65 × Pr(*publication* | *repeat*). This means that if scientists were risk-neutral and sought to maximize citations in each paper, a pathbreaking *new* or *jump* project is only 24 or 35 percent less likely to succeed than a conservative *repeat* project. We argue that this is extremely unlikely, and innovation projects are *much* more likely to fail.

This result suggests that the probability of reaching publication for innovation strategies is sufficiently small that *repeat* is likely to be the dominant strategy for a scientist seeking to maximize expected citations on each project. Citation-maximization thus supplies plausible motivation for one half of the essential tension, that is, the substantial fraction of scientists who choose tradition over innovation. Systems of professional evaluation like hiring and tenure, which often rest on reliable productivity and citations, strongly push scientists to pursue conservative *repeat* strategies. But we are left with a puzzle. While scientists eschew *most* opportunities for high-risk, high-impact work, they still engage more risk than would be rational if they only sought to maximize citations—if they were only concerned with job security and a stable career. This leads us to evaluate another potential motivation: rewards from exceptional scientific achievement, which allow scientists to move *up* in the field.

Exceptional achievement and prize-winners. Beyond job security, scientists are also motivated by the desire for *significant* impact and recognition. They wish to leave their mark on history (Merton 1973) and move up in the field (Bourdieu 1975). This level of achievement is captured by the most highly cited papers (Cokol, Rodriguez-Esteban, and Rzhetsky 2007; Wuchty et al. 2007)

and even more so by awards and prizes. Such lofty recognition is thought to require riskier, original contributions (Kuhn [1959] 1977; Merton 1949). Hence, we predict that top-cited articles and prizewinners will deploy rare, risky innovation strategies more frequently than will the typical article or scientist.

To analyze the distribution of strategies among award-winning scientists, we compiled a list of 137 prizes and awards in biomedicine and chemistry, drawing on the category pages for biology awards, medicine awards, and chemistry awards on Wikipedia.[43] We confirmed the resulting list with several biology, medical, and chemistry researchers.

All prestigious and many less prestigious prizes are listed, providing a broad sample of achievement—from field-specific (e.g., the Anselme Payen award, related to cellulose chemistry) to world-recognized (e.g., the Nobel Prize in Physiology or Medicine). We eliminated all prizes awarded for non-research reasons (e.g., teaching or history) as well as those awarded to students. We then searched the lists of prizewinners against PubMed and hand-selected three articles by each prizewinner, where possible. We expanded the initial set of articles by linking the name of the prizewinner in the MEDLINE entry for each of the seed articles with clusters of disambiguated author names assigned by the Author-ity tool (Smalheiser and Torvik 2009). We then retrieved and associated with the prizewinner all publications authored by Author-ity–linked name variants written up to 30 years before the date of the award.[44]

To examine truly outstanding achievement, we compiled a list of "elite, general" prizes (see Figure 5B). We included the two Nobel Prizes; prizes explicitly mentioned as being precursors to a Nobel; top general awards given by professional societies in the United States and the United Kingdom (the American Chemical Society, the National Academy of Science, and the Royal Society of Chemistry); and the Wolf Prizes, which have similar stature and prestige. The resulting list of 12 prizes is as follows: Nobel Prize in Physiology or Medicine, Nobel Prize in Chemistry, Louisa Gross Horwitz Prize, Lasker-DeBakey Clinical Medical Research Award, Albert Lasker Award for Basic Medical Research, Gairdner Foundation International Award, NAS Award in Chemical Sciences, Priestley Medal, Corday-Morgan Medal, Grand Prix Charles Leopold Mayer, Wolf Prize in Medicine, and Wolf Prize in Chemistry.

Figure 5 verifies with an aggregate analysis that top-cited articles deploy significantly more innovation strategies than all articles as a fraction of links contributed, with the strongest enrichment in jumping and new consolidation. The strategy distribution of typical articles is shown in Figure 5A. The relative enrichment in the top 10, 5, and 1 percent most highly cited articles is shown in Figure 5B, with multinomial confidence intervals at the 95 percent level shown in boxes.

Articles written by authors who won one of 137 different prizes in biomedicine and chemistry show a similar pattern of enrichment (see Figure 5B). This population includes 7,594 awardees and 241,176 articles. Articles written by winners of the *most* elite awards, including Nobel prizes, not only introduce new chemicals more frequently than those written by typical scientists, but they also more frequently introduce new relationships *within* knowledge clusters. A novel, integrating link within a chemical cluster attracts the attention of the existing research community whose articles inscribe the connections that define the cluster. Identifying new connections within dense, established clusters of chemicals may also be more difficult, on average, as these connections are more likely to have been tried unsuccessfully in the past by members of the community. In this way, elite award winners may define or transform our understanding of these clusters. The same pattern obtains when we analyze awards by field, grouping them into biomedicine (Figure 5C) and chemistry (Figure 5D).[45] Note that these populations also produce

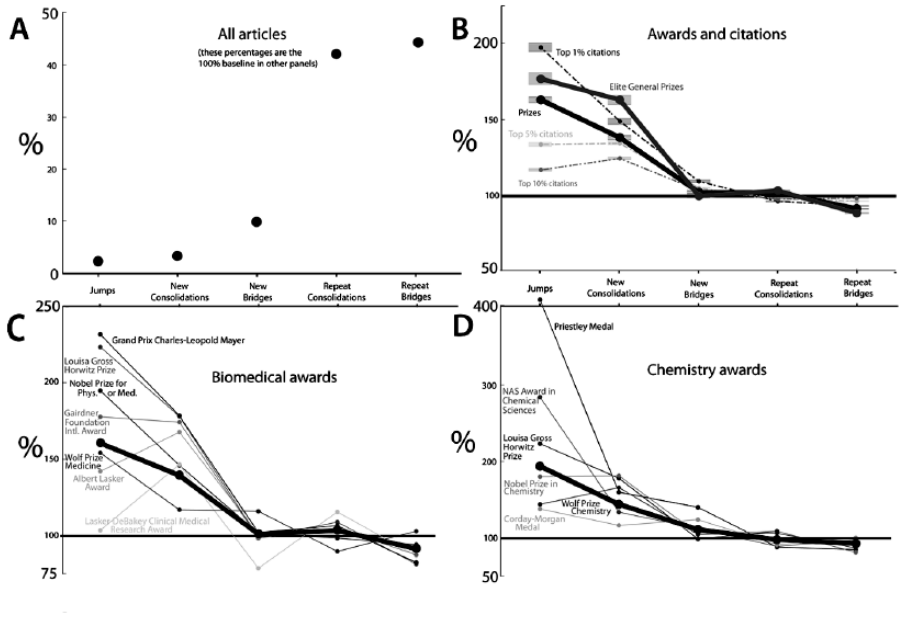

Figure 5. Strategies Shift for High Impact Scientists
Note: (A) Pooling all papers published from 1983 to 2002, we find a distribution of strategies similar to that observed in any year (Figure 3A). We judge pools of high impact articles and authors against this baseline. (B) Highly cited articles or those published by prizewinners in the 30 years before receipt of a prize show an enhancement of jumps and new consolidations. The vertical axis measures the observed strategy frequency in each pool as a percentage of the baseline value. 95 percent confidence intervals are indicated by the boxes outside each data point. (C) and (D) As in (B) but for biological and medical prizes, and chemical prizes, respectively. The heavy line shows the aggregate behavior of all prizes in the category; confidence intervals are smaller than the symbols. We also show the most prestigious, general prizes in these fields; confidence intervals are not shown.

(many) tradition links. Highly cited papers and prizewinners simply display a larger *proportion* of innovation strategies in their output than does the typical scientist.

These results provide a plausible motivational mechanism for the other half of the essential tension, linking risky innovation to extraordinary scientific achievement. As Merton (1968) argues, highly successful scientists are important models of emulation, and insofar as they use risky innovation strategies more frequently, this may inspire other scientists to engage in the occasional risky gamble. This mechanism may be quite general; a similar process has recently been described in filmmaking, where the pursuit of an award also carries considerable risk (Rossman and Schilke 2014).

DISCUSSION

Strategic Origins of the Essential Tension

Taken together, our results provide strong quantitative evidence that scientists' choice of research problem is indeed shaped by their strategic negotiation of an essential tension between productive tradition and risky innovation (Kuhn [1959] 1977). Kuhn's basic picture, when enriched with Bourdieu's conception of research choice as strategic action and his notion of scientific habitus, provides a compelling framework in which to understand our findings. Using a network representation of chemical knowledge, we demonstrated that biomedical chemistry is in a normal science regime, with conservative

strategies (tradition) common and riskier strategies (innovation) rare. This conservatism is further supported by our finding that knowledge has crystallized into stable knowledge clusters. The distribution of research claims is also remarkably stable. This would be surprising if scientists attended only to research opportunities, but becomes entirely expected if their choices are shaped by a complex of socially inflected dispositions and incentives. Using a simple behavioral model, we relate these stable choices to exploding opportunities for biomedical novelty, and we measure the growing biases that focus scientists on the known. We then provide a quantitative analysis of some of the incentives that undergird those biases. In this, we confirm that innovation strategies are not only rare but also more highly rewarded and, in all likelihood, more risky. Finally, we sketch out and quantitatively confirm two motivational explanations for tradition and innovation. We find that a disposition toward tradition is plausibly adapted to maximizing productivity and reliably accumulating scientific capital, while a disposition toward occasional innovation is motivated by a gamble for posterity and a desire to achieve higher position in the field. We believe our results, taken as stylized facts about stability, dispositions, risk-reward, and motivational mechanisms, are robust to the necessary modeling assumptions we made and the limitations of the data outlined in the next section.

Limitations of the Current Study and Robustness of Results to Selection

Broad limitations. Our analysis builds on much recent research in the social science of innovation but draws together many effects in a single case; it thus serves as the first large-scale confirmation of a key concept in the qualitative study of science and innovation. Despite our novel methods and our analysis of a new, large dataset, this study has several limitations. Protocols for chemical annotation at the National Library of Medicine are not uniform over time. Our method necessarily misses relationships whose chemicals are not in the NLM annotation. By considering all co-mentioned chemicals, we also allow some "false relationships," failing to distinguish between substantial and incidental associations. Moreover, our taxonomy of research strategies does not reflect the full variety of factors scientists consider. More precise extraction of chemical relationships (Evans and Foster 2011; Evans and Rzhetsky 2011), a broader full-text corpus including research proposals, and richer models can transcend these limits.

Selection effect. Strategies that failed to yield publishable findings are excluded from our analysis because we focus on publications. Low-risk incremental strategies may be more prevalent in our corpus not just because scientists choose them more frequently, but also because they fail less frequently, leading to publication more often. Risky strategies, as suggested by the name, fail more frequently, and thus remain unpublished, a phenomenon often known as the "file drawer" problem (Rosenthal 1979). This suggests that taking the distribution of strategies *in the literature* as representative of the underlying distribution of strategies may misestimate the actual prevalence of a given strategy in scientists' aggregate research effort.

In fact, our results are largely unaffected by this filtering process. Our finding of *stability* stands, unless there is temporal variation in the probability that various strategies succeed or fail. Our proposed connection between *risk and reward* also stands: if research failures were routinely published, it would enhance the correlation between strategy surprisal and variance in citations by disproportionately adding unpublished and uncited risky attempts to our sample. The claim that *high-achieving scientists* visibly produce more high-risk research is also unaffected by this selection. We note that the simplest explanation for a higher prevalence of risky strategies among high-achieving scientists is that they engage in risky research efforts more often, consistent with our theory and

analysis. Another plausible explanation, which we pose as a research question, is that these scientists *succeed* more frequently in their risky efforts, either by skill in problem selection, pure luck, or greater success in the exercise of symbolic power (i.e., getting innovations accepted).[46] These alternatives do not affect the conclusion that high achievement is visibly associated with riskier strategies and consequently likely to motivate risk-taking in other scientists. Finally, *policy prescriptions* that promote more risk-taking will lead to a higher prevalence of risky strategies. If incentives for risk-taking are extreme, diminishing returns may obtain as scientists are pushed into outlandish projects, but it is implausible that we are near this point in the current policy regime. A myopic focus on novelty contributes to increased publication of novel yet spurious results (Ioannidis 2005a, 2005b), while a focus on tradition increases the likelihood of needless repetition and spurious incremental findings. An increase in the supply of risky projects will lead to their higher prevalence in publications.

The only claim directly affected by the selection effect is the relative rank of strategy prevalence in science. It is impossible to validate this using our data. Recall that repeated chemical relationships are half an order of magnitude more frequent in published findings than new relationships, and new relationships half an order of magnitude more frequent than jumps. The size of these differences suggests that the underlying effort distribution in science is probably reflected in the rank order of published strategy prevalence. Its plausibility is further bolstered by a reasonable risk assessment of the various strategies and widely documented, subjectively valid concerns about pressures for productivity, which prioritize conservative, reliable *repeat* strategies.

CONCLUSIONS

Understanding the research process remains a central challenge for science studies, and improved understanding will be key to improved science policy (Evans and Foster 2011). Science is a complex system (Foote 2007), and new methods like ours can identify some of the processes that govern its evolution. Our research suggests one reason why unexpected findings that change the landscape of science are infrequent. Pursuing innovation is a gamble, without enough payoff, on average, to justify the risk. Nevertheless, science benefits when individuals overcome the dispositions that orient them toward established islands of knowledge (Cokol et al. 2005) in the expanding ocean of possible topics. Early breakthroughs in literature-based discovery illustrated the power of linking islands of knowledge together (Swanson 1990). Our results explain why such discovery methods remain fruitful.

To be sure, not all scientists should pursue risky strategies. Normal science that characterizes a known relationship more deeply has its own value, as recognized by Kuhn ([1959] 1977). Nevertheless, stimulating innovation is an important goal of science policy, and we suggest two policy levers to promote risk-taking. The first involves decoupling job security from productivity, which can encourage originality, as was the case at Bell Labs (Gertner 2012). So can funding scientists rather than projects, as at the Howard Hughes Medical Institute and a new family of grants at the National Institutes of Health.[47] These policies follow from the insight that career pressure represents a powerful incentive for conservative behavior.[48] The other lever sits outside the tension between productivity and posterity: agencies can lower the barriers to risky projects by funding them more aggressively, like the Gates Foundation. These interventions may be able to counter the stable conservatism we uncover here.

Our paper also has implications for the sociology of science and for sociological methodology more broadly. First, our findings suggest new research questions in the sociology of science. For example, how do strategies change over a scientific career? How does local context shape strategic choice? Are high-achieving scientists initiators of more risky projects, more far-sighted,

more capable of exercising symbolic power, or simply luckier? And how are the dispositions of risk-taking scientists (Bateman and Hess 2015) constituted? Our study also suggests that new methods for representing scientific knowledge and modeling scientific incentives and behavior can bring big data to bear in evaluating the scope of rich qualitative insights from science studies, the sociology of scientific knowledge, and the anthropology and history of science (Evans and Foster 2011). Our method should extend to other fields, as long as we can identify plausible building blocks of content. For example, consider PACS codes in physics, MSC codes in mathematics, USPTO classification codes in patents, and classification codes or author-assigned keywords in sociology. We believe that new methods should be developed for mining building blocks with finer granularity—comparable to chemicals—and for defining strategies beyond the combinatorial. The broad contours of our findings should carry over to other relatively stable disciplines characterized by substantial consensus on core elements and on which combinations of these elements are valid and interesting. It would be especially intriguing to extend our analysis along the lines suggested by Panofsky (2011), combining the two forms of scientific capital examined here (citations and awards) with more and less external and exchangeable capitals (e.g., reputation in other fields, funding, and patents).

As we noted earlier, the essential tension has been studied many times under many names. Our empirical findings and theoretical framework make important advances beyond simply bringing together multiple phenomena identified in isolated studies from the sociology, economics, and management of innovation. First, contrary to Kuhn and in keeping with Bourdieu, we find a coexistence of conservative tradition and radical innovation strategies within a field—and within a paper. Second, we find increased emphasis on tradition relative to novel opportunities as the space of possibilities expands. Third, we find that novel connections *within* fields are more highly rewarded—innovation within tradition—which is unanticipated under many theories of interdisciplinarity (e.g., those based on structural holes). Fourth, we provide and substantiate a fundamentally strategic account of the essential tension, sharply at odds with Kuhn's internal cognitive account and building usefully on Bourdieu's. Under our account, a disposition toward tradition is reflected and reproduced by the relationship between expected reward and expected risk, which is further modulated by the screening processes of scientific peer review. We find it surprising that innovation is relatively unrewarded in the day-to-day currency of scientific credit (citations), given how rare and risky innovation is and how highly novelty is prized. This may, in part, reflect the insensitivity of our approach to innovation beyond the combination of chemicals (e.g., the linkage of a chemical with a novel method or disease). It may also reflect the absolute difficulty of moving new relationships, once accepted as provisionally legitimate, into the category of legitimate and interesting. Cultivating such interest may require substantial mobilization (Frickel and Gross 2005). Finally, the apparent division of motivational labor between citations (as an anticipated reward for traditional work) and awards (as a hard-to-anticipate reward for innovation) is novel and has not been documented before.

Methodologically, our analysis makes three contributions. First, it provides a template for applying the machinery of network analysis to scientific content, and suggests that such methods can be fruitfully applied to other forms of content. Second, it demonstrates that, if one is willing to tolerate simplification, rich sociological constructs like the habitus can be represented in formal models and measured (as stylized facts) on large scales, extending the quantitative aspects of Bourdieu's program beyond correspondence analysis (Bourdieu 1984) and creating surprising connections with analytic sociology (Hedström 2005). Finally, our paper argues for the utility of simple probabilistic models of behavior coupled with maximum

likelihood inference. Such generative models are a happy compromise between the behavioral plausibility of agent-based models and the closeness to data of classical regression analysis, and there is much room for extending our simple generative model to incorporate rich representations of individual dispositions and field positions. To put it reflexively: our paper is grounded in several traditions from the sociology of science, network analysis, and statistical modeling. In making novel connections within and between them, it contributes substantive or methodological innovations to each.

Acknowledgments

We are grateful for helpful comments from Carl Bergstrom, David Blei, Charles Camic, Erica Cartmill, and Steve Epstein. We also received valuable feedback from audiences at Duke University, Princeton University, the University of California-Los Angeles, the University of Chicago, the University of Michigan, Harvard Business School, and the American Sociological Association. We thank Martin Rosvall and Daniel Edler for generous assistance with use of their MapEquation software package; Jeff Alstott for help with his Python package powerlaw; research assistants Mahmoud Bahrani, Simo Huang, David Kates, and Val Michelman for help compiling data on prizewinners in biomedicine and chemistry; the National Library of Medicine and Jane Rosov for assistance with MEDLINE chemical annotations; and ThomsonReuters for making their citation information available. This work was supported by NSF grant SBE 0915730 and a grant (ID: 39147) to the Metaknowledge Network by the John Templeton Foundation.

Notes

1. All of these factors are *internal* to the scientific field (Camic 2011, 2013). As pressure to patent and commercialize scientific research increases, the scientific field necessarily becomes less autonomous and more subject to *external* forces impinging from without (Berman 2008, 2012; Camic 2011, 2013; Davis, Larsen, and Lotz 2011; Gibbons et al. 1994; Moses et al. 2005; Nowotny, Scott, and Gibbons 1994; Powell and Owen-Smith 1998; Thursby and Thursby 2011; Woolf 2008; Wright 1983).
2. As Camic (2011, 2013) notes, while Bourdieu often adopts this multiplex strategy in his empirical work, he downplays external factors in programmatic statements of his sociology of science, emphasizing internal competitive dynamics within the field.
3. Kuhn (1959) initially wrote and presented "The Essential Tension: Tradition and Innovation in Scientific Research" at the third University of Utah Research Conference on the Identification of Scientific Talent.
4. It is worth pointing out that risk-taking strategies are easier to sustain when scientists have already accumulated enough recognition to live off the returns for a while. This was in fact the case for Wiles and even more for Sanger, who had already won his first Nobel.
5. The distinction between "tradition" and "innovation" in organizational action is also made in classical organization theory (Lawrence and Lorsch 1969); these authors observed that the two were not mutually exclusive.
6. While we use the language of research "choice" throughout, we do not imply that the selection of one problem over another is necessarily the outcome of a deliberate or deliberative process. We mean that out of many possibilities, one is pursued and possibly published. This selection is influenced by a wealth of factors.
7. ". . . revolutionary shifts of a scientific tradition are relatively rare, and extended periods of convergent research are the necessary preliminary to them" (Kuhn [1959] 1977:227).
8. Tradition and innovation have different stakes for those at different positions in the scientific enterprise (e.g., different stocks of scientific capital and different demands on their capital accumulation). Although a fascinating topic, we do not explore it in this article, reserving it for future work.
9. In a scientific field that is *truly* autonomous, that is, governed only by its internal competitive dynamics, the dichotomy between tradition and innovation would be most consequential. This is an ideal type, for no such field can exist.
10. Some of this work shows, however, that patents are often a natural byproduct of scientific work, rather than a driver (Azoulay, Ding, and Stuart 2009), and patent policies, like the Bayh-Dole Act of 1980, may not stimulate a shift to patenting and commercialization as much as reflect preexisting currents (Mowery et al. 2001).
11. Papers must clear some minimal threshold of novelty to be publishable, but—particularly in less prestigious journals—that threshold can be very low. Here we focus on strong forms of novelty, at the level of fundamental objects of study (i.e., chemicals). As we will discuss, novelty takes different forms in different disciplines.
12. Fleming (2001) finds that when inventors combine unexpected components in new patents, they have a lower average but higher variance in citations. Patents have the explicit burden of novelty, but this assessment is made by the examiner and does not require peer recognition. We argue that the burden is greater for papers that reveal not only technical but also substantive novelty (which cannot be hedged with limited patent claims). As a result, we

expect that published combinations of novel scientific content will receive *higher* average citations, as they have already received the recognition of peer reviewers as legitimate.

13. She may explicitly set out to design a new compound (e.g., creating a new pharmaceutical); find evidence of a new entity and set out to characterize it (e.g., describing a new receptor or antibiotic); or simply introduce a substance characterized *outside* biomedical chemistry into the network of biomedical knowledge.
14. The core data connecting chemicals to papers, chemicals to clusters, strategies to papers, and papers to awards is available at http://asr.sagepub.com/supplemental.
15. All journals added after 1980 were indexed, including their pre-1980 content. This accounts for the pre-1980 annotations in the dataset.
16. It also captures the realistic delay between publication of an article and awareness of the relationships discussed therein.
17. Note that our results and conclusions are not sensitive to this treatment of time. We analyzed a large subset of data with a shifting window biased toward more recent links, assigning all links a weight that decays exponentially as a function of time, with a three-year half-life. The results were consistent with the temporal evolution of strategies shown in Figure 3A (solid lines).
18. For clusters having no clear antecedent (i.e., those that break off from earlier clusters or form from the merger of multiple earlier clusters), we list the overlap as n/a in Table 1.
19. See Results by Year for Second Messenger System in PubMed (http://www.ncbi.nlm.nih.gov/pubmed?term=second%20messenger%20system).
20. PageRank, an established centrality measure, may also be a good index of a chemical's "prominence in memory" for scientists (see Griffiths, Steyvers, and Firl 2007).
21. This example (NAPVSIPQ and kainic acid) is a particularly tight relationship. Relationships can also be compositional, type of, or associational. For example, the same abstract is annotated with oligopeptides; NAPVSIPQ is a type of oligopeptide, and kainic acid is associated with oligopeptides because they occur in the same neurobiological context. The frequency of incidental, weak relationships may differ by research strategy. Consolidations between nearby chemicals in the same knowledge cluster may be more likely to represent a substantive chemical relationship than bridges between distant chemicals in separate clusters, which may be more likely to be incidental associations.
22. Three definitional points. First, some papers contain a single chemical, which we ignore; we focus on strategies that weave relationships. The first link that adds a chemical to the network is treated as a jump. Second, if multiple abstracts in the same year contain the same "new" link, all instances count as new. This elides some sequencing behavior, but the lag between submission and publication suggests that such grouping is warranted. Third, recall that the knowledge network is *cumulative,* so chemicals with no link between them in $t - 1$ have never been linked in the published literature before time t.
23. In reality, these choices are correlated at the article level, but this vastly complicates the analysis. Patterns of past choice *in aggregate* are incorporated in the generative model, but patterns of past choice by individual scientists are not. This could be introduced, along with heterogeneity in scientists' preferences; both would alter the assumption of identically distributed strategies. We are developing these additions in future work, but they are beyond the scope of this project.
24. Several alternative models can be formulated, including a model that associates a parameter to each strategy or one that makes the choice of jump, consolidation, or bridge primary and new or repeat secondary. All models fit the data reasonably well, but the model presented here gives a good fit while also presenting a behaviorally plausible sequence of decisions, and it is consistent with our overarching theory and subsequent analyses of risk and reward.
25. This assumes a uniform habitus across scientists acting from an average position. This is, of course, a departure from both reality and the spirit of field theoretic analysis. As noted earlier, these assumptions could be relaxed at the cost of considerable complexity. It should also be noted that most choices are in fact made by teams of scientists, rather than individual scientists. We reserve such complications for later work.
26. If a link has been repeated in 3,000 article abstracts, it contributes 3,000 to the count.
27. We assume the total number of chemicals available to be discovered is 250,000. Sensitivity tests indicate that changing this number shifts parameter values but not their pattern. We also assume that the full space of combinatorial possibility is accessible. While the number of valid combinations lies below this maximal number, our assumption is defensible. First, our results will be qualitatively unchanged unless the current network is essentially correct and exhaustive, which is implausible. Second, in a *biomedical* context, relationships between chemicals are plausible even when physical interaction is impossible. For example, chemicals may interact indirectly through metabolites or physiological consequences.
28. To test the robustness of the generative model to the size of this smoothing window, we estimated parameters with two-, four-, eight-, and ten-year windows as well. The qualitative behavior of the parameters is identical in all cases; shorter windows exhibit more volatility on short time-scales. A six-year window strikes a balance between tracking change and capturing the durability of dispositions.

29. Although we presented the model from the perspective of a choosing scientist, it is most accurate to view it as distilling the dispositions of the community—because we examine *published* strategies, the parameters fold together the preferences of choosing scientists and evaluating reviewers.
30. We use "normal science" quite narrowly, to index the stability of chemical knowledge in biomedicine. A paradigm shift in biomedical chemistry should occasion the introduction of a large number of newly relevant chemicals, *or* a significant uptick in the discovery of new relationships, *or* a substantial re-organization of the knowledge clusters. Although systems biology or bioinformatics may represent paradigm shifts in biomedicine, these shifts only weakly overlap with our study period, and in both cases primarily affect the scope and method of investigation. They thus have limited implications for the structure of chemical knowledge.
31. These coarser classifications make no reference to the knowledge clusters; results for the coarse taxonomy thus demonstrate that the relationship between strategy and reward is not an artifact of community detection.
32. As a robustness test, we repeated all analyses linking strategy type to mean citation and standard deviation in citations, but assigned each abstract the citations received five, seven, and ten years following initial publication. Results for mean citations (both aggregate and disaggregate) are robust across all windows. Results for standard deviation are somewhat less robust, as the longer run allows highly successful papers in each category to wash out differences apparent in the short-run.
33. This decreased coverage reflects the date when we obtained a complete version of *Web of Science* in May 2006.
34. Citations display a significant positive trend with year, defined as number of years since 1983. These results were subjected to many additional robustness tests (see Part E of the online supplement), including using the surprisals predicted by the generative model as a predictor; truncating the study period to 1983–1997 to remove the effect of changing annotation practices; excluding jumps from the analysis entirely; using median rather than mean citations as the outcome; and making subfield assignments on the basis of journal co-appearance or externally curated ontologies (see Table S3 in the online supplement).
35. In regression models of standard deviation under the coarse taxonomy, coefficients for *new* remain significant at five years, while those for *jump* remain significant (five, seven) or marginally significant (ten) at all citation windows. Under the fine taxonomy, coefficients for *new consolidation* remain significant at five and seven years, while *jump* remains significant for all citation windows. Radical innovation (*jump*) is robustly uncertain, even over long time horizons.
36. Negative binomial models contain an extra parameter to capture overdispersion (i.e., a thick right-hand tail); this parameter (α) equals 0 in the case of a Poisson. In all our cases, the overdispersion parameter is significantly different from zero according to a likelihood ratio test comparing the negative binomial model to a Poisson model, $p < .001$, so our use of the negative binomial is justified throughout.
37. Papers using innovative strategies are also *more* likely to be cited at all, if they are published. Binomial logit models (distinguishing "zero" from "one or more" citations) as well as multinomial logit models (with yearly citation quartiles as the outcome variable) confirm that shifting from tradition to innovation increases the odds of being cited and of moving from lower to higher quartiles. The true risk associated with these strategies is in failure to publish.
38. Note that optimal levels of *jumping* or *new bridging* only involve one chemical new to the conventional literature, so it is reasonable to expect this would be referenced much less.
39. A strictly field-theoretic formulation would ask whether the dispositions generating observed levels of innovation are objectively adapted to citation accumulation and forgo explicit reference to motivation or choice; we retain this spirit, but opt for more transparent language.
40. Again, we can put this in a less voluntaristic way: assuming that scientists' dispositions over research strategies are adapted to maximize citations.
41. Note that this risk analysis puts bounds on how strong the selection effect would have to be to preserve a simple story in which accumulating citations is a sufficient motivation for the distribution of published strategies.
42. We obtain comparable results if we compute success probabilities on a yearly basis.
43. http://en.wikipedia.org/wiki/Category:Biology_awards;
http://en.wikipedia.org/wiki/Category:Medicine_awards;
http://en.wikipedia.org/wiki/Category:Chemistry_awards.
44. This process could introduce false positives for common names if we misassigned seed articles, associating a prizewinner with an article written by another scientist. False positives, however, cause us to underestimate the size of the prize effect. Prizewinners with non-English characters in their names are also underrepresented, as Author-ity is less likely to cluster these into an identified author. Such false negatives also lead us to underestimate the enrichment of innovation strategies in the strategy distribution of prizewinners. Our results should be viewed as a conservative estimate of the difference between prizewinners and the pool of all scientists.
45. The list of papers associated with awards is available at http://asr.sagepub.com/supplemental.

46. We emphasize that the strategies of award winners are examined *before* their consecration by awards. That said, the more prestigious a scientist becomes, the less risky it is for her to take risks. Her symbolic power, greater stock of resources, and the operation of the Matthew effect make it easier for her to publish innovation strategies and capture a larger share of the resulting recognition (Merton 1968).
47. These interventions raise further questions about how to identify the promising scientists who receive such support; this is a fascinating research topic in its own right.
48. Although we do not analyze the impact of funding here, it is likely to be a conservative force. Funding tracks and hence reinforces current distributions of research attention (see Yao et al. forthcoming).

References

Abbott, Andrew D. 1999. *Department & Discipline: Chicago Sociology at One Hundred*. Chicago: University of Chicago Press.

Abbott, Andrew D. 2001. *Chaos of Disciplines*. Chicago: University of Chicago Press.

Albert, Mathieu, and Daniel Lee Kleinman. 2011. "Bringing Pierre Bourdieu to Science and Technology Studies." *Minerva* 49(3):263–73.

Austerweil, Joseph L., Joshua T. Abbott, and Thomas L. Griffiths. 2012. "Human Memory Search as a Random Walk in a Semantic Network." Pp. 3041–49 in *Advances in Neural Information Processing Systems 25*, edited by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger. Curran Associates, Inc.

Azoulay, Pierre, Waverly Ding, and Toby Stuart. 2009. "The Impact of Academic Patenting on the Rate, Quality and Direction of (Public) Research Output." *Journal of Industrial Economics* 57(4):637–76.

Azoulay, Pierre, Joshua S. Graff Zivin, and Gustavo Manso. 2011. "Incentives and Creativity: Evidence from the Academic Life Sciences." *RAND Journal of Economics* 42(3):527–54.

Azoulay, Pierre, Toby Stuart, and Yanbo Wang. 2013. "Matthew: Effect or Fable?" *Management Science* 60(1):92–109.

Baldi, Stéphane. 1998. "Normative versus Social Constructivist Processes in the Allocation of Citations: A Network-Analytic Model." *American Sociological Review* 63(6):829–46.

Barabási, Albert-Lászlo, and Réka Albert. 1999. "Emergence of Scaling in Random Networks." *Science* 286(5439):509–512.

Bateman, Thomas S., and Andrew M. Hess. 2015. "Different Personal Propensities among Scientists Relate to Deeper vs. Broader Knowledge Contributions." *Proceedings of the National Academy of Sciences* 112(12):3653–58.

Berman, Elizabeth Popp. 2008. "Why Did Universities Start Patenting? Institution-Building and the Road to the Bayh-Dole Act." *Social Studies of Science* 38(6):835–71.

Berman, Elizabeth Popp. 2012. "Explaining the Move toward the Market in US Academic Science: How Institutional Logics Can Change without Institutional Entrepreneurs." *Theory and Society* 41(3):261–99.

Bourdieu, Pierre. 1975. "The Specificity of the Scientific Field and the Social Conditions for the Progress of Reason." *Social Science Information* 14(6):19–47.

Bourdieu, Pierre. 1984. *Distinction: A Social Critique of the Judgement of Taste*. New York: Routledge.

Bourdieu, Pierre. 1986. "The Forms of Capital." Pp. 241–58 in *Handbook of Theory and Research for the Sociology of Education*, edited by J. Richardson. New York: Greenwood.

Bourdieu, Pierre. 1990. *The Logic of Practice*. Stanford, CA: Stanford University Press.

Bourdieu, Pierre. 1991. "The Peculiar History of Scientific Reason." *Sociological Forum* 6(1):3–26.

Bourdieu, Pierre. 2004. *Science of Science and Reflexivity*. Chicago: University of Chicago Press.

Bruggeman, Jeroen, V. A. Traag, and Justus Uitermark. 2012. "Detecting Communities through Network Data." *American Sociological Review* 77(6):1050–63.

Busch, Lawrence, William B. Lacy, and Carolyn Sachs. 1983. "Perceived Criteria for Research Problem Choice in the Agricultural Sciences: A Research Note." *Social Forces* 62(1):190–200.

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

Camic, Charles. 2011. "Bourdieu's Cleft Sociology of Science." *Minerva* 49(3):275–93.

Camic, Charles. 2013. "Bourdieu's Two Sociologies of Knowledge." Pp. 183–214 in *Bourdieu and Historical Analysis*, edited by P. S. Gorski. Durham, NC: Duke University Press.

Chen, Chaomei, Yue Chen, Mark Horowitz, Haiyan Hou, Zeyuan Liu, and Donald Pellegrino. 2009. "Towards an Explanatory and Computational Theory of Scientific Discovery." *Journal of Informetrics* 3(3):191–209.

Cokol, Murat, Ivan Iossifov, Chani Weinreb, and Andrey Rzhetsky. 2005. "Emergent Behavior of Growing Knowledge about Molecular Interactions." *Nature Biotechnology* 23(10):1243–47.

Cokol, Murat, Raul Rodriguez-Esteban, and Andrey Rzhetsky. 2007. "A Recipe for High Impact." *Genome Biology* 8(5):406.

Cover, Thomas M., and Joy A. Thomas. 1991. *Elements of Information Theory*. New York: Wiley.

Cowen, Tyler. 2011. *The Great Stagnation: How America Ate All The Low-Hanging Fruit of Modern History, Got Sick, and Will (Eventually) Feel Better*. New York: Dutton Adult.

Cronin, Blaise, and Helen Barsky Atkins, eds. 2000. *The Web of Knowledge: A Festschrift in Honor of Eugene Garfield*. Medford, NJ: Information Today.

Davis, Lee, Maria Theresa Larsen, and Peter Lotz. 2011. "Scientists' Perspectives Concerning the Effects of University Patenting on the Conduct of Academic

Research in the Life Sciences." *Journal of Technology Transfer* 36(1):14–37.

Diamond, Arthur M., Jr. 1994. "The Determinants of a Scientist's Choice of Research Projects." Pp. 167–210 in *Scientific Failure*, edited by T. Horowitz and A. I. Janis. Lanham, MD: Rowman & Littlefield.

Ding, Waverly W., Fiona Murray, and Toby E. Stuart. 2006. "Gender Differences in Patenting in the Academic Life Sciences." *Science* 313(5787):665–67.

Evans, James A. 2010a. "Industry Collaboration, Scientific Sharing and the Dissemination of Knowledge." *Social Studies of Science* 40(5):757–91.

Evans, James A. 2010b. "Industry Induces Academic Science to Know Less about More." *American Journal of Sociology* 116(2):389 –452.

Evans, James A., and Jacob G. Foster. 2011. "Metaknowledge." *Science* 331(6018):721–25.

Evans, James A., and Andrey Rzhetsky. 2011. "Advancing Science through Mining Libraries, Ontologies, and Communities." *Journal of Biological Chemistry* 286(27):23659–66.

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

Fleming, Lee, Santiago Mingo, and David Chen. 2007. "Collaborative Brokerage, Generative Creativity, and Creative Success." *Administrative Science Quarterly* 52(3):443–75.

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

Foote, Richard. 2007. "Mathematics and Complex Systems." *Science* 318(5849):410–12.

Frickel, Scott, and Neil Gross. 2005. "A General Theory of Scientific/Intellectual Movements." *American Sociological Review* 70(2):204–232.

Funk, Russell J., and Jason Owen-Smith. 2012. "A Dynamic Network Approach to Breakthrough Innovation." *arXiv:1212.3559 [physics]*, December (http://arxiv.org/abs/1212.3559).

Gertner, Jon. 2012. *The Idea Factory: Bell Labs and the Great Age of American Innovation*. New York: Penguin.

Gibbons, Michael, Camille Limoges, Helga Nowotny, Simon Schwartzman, Peter Scott, and Martin Trow. 1994. *The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies*. Thousand Oaks, CA: Sage.

Gieryn, Thomas F. 1978. "Problem Retention and Problem Change in Science." *Sociological Inquiry* 48(3–4):96–115.

Goodman, Leo A. 1965. "On Simultaneous Confidence Intervals for Multinomial Proportions." *Technometrics* 7(2):247–54.

Griffiths, Thomas L., Mark Steyvers, and Alana Firl. 2007. "Google and the Mind: Predicting Fluency with PageRank." *Psychological Science* 18(12):1069–76.

Guetzkow, Joshua, Michèle Lamont, and Grégoire Mallard. 2004. "What Is Originality in the Humanities and the Social Sciences?" *American Sociological Review* 69(2):190–212.

Guimerà, Roger, Brian Uzzi, Jarrett Spiro, and Luís A. Nunes Amaral. 2005. "Team Assembly Mechanisms Determine Collaboration Network Structure and Team Performance." *Science* 308(5722):697–702.

Hargadon, Andrew, and Robert I. Sutton. 1997. "Technology Brokering and Innovation in a Product Development Firm." *Administrative Science Quarterly* 42(4):716–49.

Hausman, Jerry, Bronwyn H. Hall, and Zvi Griliches. 1984. "Econometric-Models for Count Data with an Application to the Patents–R&D Relationship." *Econometrica* 52(4):909–938.

Hebb, Donald O. 1949. *The Organization of Behavior*. New York: Wiley & Sons.

Hedström, Peter. 2005. *Dissecting the Social: On the Principles of Analytical Sociology*. Cambridge, UK: Cambridge University Press.

Ioannidis, John P. A. 2005a. "Why Most Published Research Findings Are False." *PLoS Med* 2(8):e124.

Ioannidis, John P. A. 2005b. "Contradicted and Initially Stronger Effects in Highly Cited Research." *Journal of the American Medical Association* 294(2):218–28.

Kevles, Daniel J. 1978. *The Physicists: The History of a Scientific Community in Modern America*, 1st ed. New York: Knopf.

Kim, Kyung-Man. 2009. "What Would a Bourdieuan Sociology of Scientific Truth Look Like?" *Social Science Information* 48(1):57–79.

Kleinberg, Jon, and Sigal Oren. 2011. "Mechanisms for (Mis)Allocating Scientific Credit." Pp. 529–38 in *Proceedings of the Forty-Third Annual ACM Symposium on Theory of Computing*. New York: ACM.

Knorr-Cetina, Karin. 1999. *Epistemic Cultures: How the Sciences Make Knowledge*. Cambridge, MA: Harvard University Press.

Kuhn, Thomas S. [1959] 1977. "The Essential Tension: Tradition and Innovation in Scientific Research." Pp. 162–74 in *The Third (1959) University of Utah Research Conference on the Identification of Scientific Talent*, edited by C.W. Taylor. Salt Lake City: University of Utah Press. Reprint, pp. 225–39 in *The Essential Tension: Selected Studies in Scientific Tradition and Change*. Chicago: University of Chicago Press.

Kuhn, Thomas S. 1962. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Lamont, Michèle. 2009. *How Professors Think: Inside the Curious World of Academic Judgment*. Cambridge, MA: Harvard University Press.

Lancichinetti, Andrea, and Santo Fortunato. 2009. "Community Detection Algorithms: A Comparative Analysis." *Physical Review E* 80(5):056117.

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

Lawrence, Paul R., and Jay William Lorsch. 1969. *Organization and Environment: Managing Differentiation and Integration*. Homewood, IL: R. D. Irwin.

Leahey, Erin, Christine Beckman, and Taryn Stanko. 2013. "Prominent but Less Productive: The Impact of Interdisciplinarity on Scientists' Research." Working paper.

Leahey, Erin, and James Moody. 2014. "Sociological Innovation through Subfield Integration." *Social Currents* 1(3):228–56.

Malmgren, R. Dean, Julio M. Ottino, and Luís A. Nunes Amaral. 2010. "The Role of Mentorship in Protege Performance." *Nature* 465(7298):622–26.

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

Menczer, Filippo. 2004. "Evolution of Document Networks." *Proceedings of the National Academy of Sciences USA* 101(Suppl. 1, April):5261–65.

Merton, Robert K. 1938. "Science, Technology and Society in Seventeenth Century England." *Osiris: Studies on the History and Philosophy of Science, and on the History of Learning and Culture* 4(1938):362–632.

Merton, Robert K. 1942. "Science and Technology in a Democratic Order." *Journal of Legal and Political Sociology* 1(1):115–26.

Merton, Robert K. 1949. *Social Theory and Social Structure: Toward the Codification of Theory and Research*. Glencoe, IL: Free Press.

Merton, Robert K. 1957. "Priorities in Scientific Discovery: A Chapter in the Sociology of Science." *American Sociological Review* 22(6):635–59.

Merton, Robert K. 1968. "The Matthew Effect in Science." *Science* 159(3810):56–63.

Merton, Robert K. 1973. *The Sociology of Science: Theoretical and Empirical Investigations*. Chicago: University of Chicago Press.

Merton, Robert K. 1988. "The Matthew Effect in Science, II: Cumulative Advantage and the Symbolism of Intellectual Property." *Isis* 79(299):606–623.

Merton, Robert K., and Elinor Barber. 2011. *The Travels and Adventures of Serendipity: A Study in Sociological Semantics and the Sociology of Science*. Princeton, NJ: Princeton University Press.

Moses, Hamilton, III, E. Ray Dorsey, David H. M. Matheson, and Samuel O. Thier. 2005. "Financial Anatomy of Biomedical Research." *Journal of the American Medical Association* 294(11):1333–42.

Mowery, David C., Richard R. Nelson, Bhaven N. Sampat, and Arvids A. Ziedonis. 2001. "The Growth of Patenting and Licensing by U.S. Universities: An Assessment of the Effects of the Bayh-Dole Act of 1980." *Research Policy* 30(1):99–119.

Murray, Fiona, and Scott Stern. 2007. "Do Formal Intellectual Property Rights Hinder the Free Flow of Scientific Knowledge? An Empirical Test of the Anti-Commons Hypothesis." *Journal of Economic Behavior & Organization* 63(4):648–87.

Newell, Allen, and Herbert Alexander Simon. 1972. *Human Problem Solving*. Upper Saddle River, NJ: Prentice Hall.

Newman, M. E. J. 2003. "The Structure and Function of Complex Networks." *SIAM Review* 45(2):167–256.

Nowotny, Helga, Peter Scott, and Michael Gibbons. 1994. *The New Production of Knowledge: The Production of Science and Research in Contemporary Societies*. London, UK: Sage Publications.

Onnela, J.-P., J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási. 2007. "Structure and Tie Strengths in Mobile Communication Networks." *Proceedings of the National Academy of Sciences* 104(18):7332–36.

Panofsky, Aaron L. 2011. "Field Analysis and Interdisciplinary Science: Scientific Capital Exchange in Behavior Genetics." *Minerva* 49(3):295–316.

Polanyi, Michael. 1969. *Knowing and Being: Essays by Michael Polanyi*, edited by M. Grene. Chicago: University of Chicago Press.

Powell, Walter W., and Jason Owen-Smith. 1998. "Universities and the Market for Intellectual Property in the Life Sciences." *Journal of Policy Analysis and Management* 17(2):253–77.

Quesenberry, C. P., and D. C. Hurst. 1964. "Large Sample Simultaneous Confidence Intervals for Multinomial Proportions." *Technometrics* 6(2):191–95.

Rosenthal, Robert. 1979. "The File Drawer Problem and Tolerance for Null Results." *Psychological Bulletin* 86(3):638–41.

Rossman, Gabriel, and Oliver Schilke. 2014. "Close, But No Cigar: The Bimodal Rewards to Prize-Seeking." *American Sociological Review* 79(1):86–108.

Rosvall, Martin, D. Axelsson, and Carl T. Bergstrom. 2009. "The Map Equation." *European Physical Journal Special Topics* 178(1):13–23.

Rosvall, Martin, and Carl T. Bergstrom. 2008. "Maps of Random Walks on Complex Networks Reveal Community Structure." *Proceedings of the National Academy of Sciences, USA* 105(4):1118–23.

Rosvall, Martin, and Carl T. Bergstrom. 2010. "Mapping Change in Large Networks." *PLoS ONE* 5(1):e8695.

Shapin, Steven. 1995. "Here and Everywhere: Sociology of Scientific Knowledge." *Annual Review of Sociology* 21(January):289–321.

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(October):73–85.

Shwed, Uri, and Peter S. Bearman. 2010. "The Temporal Structure of Scientific Consensus Formation." *American Sociological Review* 75(6):817–40.

Smalheiser, Neil R., and Vetle I. Torvik. 2009. "Author Name Disambiguation." *Annual Review of Information Science and Technology* 43(1):1–43.

Stuart, Toby E., and Waverly W. Ding. 2006. "When Do Scientists Become Entrepreneurs? The Social Structural Antecedents of Commercial Activity in the Academic Life Sciences." *American Journal of Sociology* 112(1):97–144.

Swanson, Don R. 1990. "Medical Literature as a Potential Source of New Knowledge." *Bulletin of the Medical Library Association* 78(1):29–37.

Thursby, Jerry G., and Marie C. Thursby. 2011. "Has the Bayh-Dole Act Compromised Basic Research?" *Research Policy* 40(8):1077–83.

Uzzi, Brian, Satyam Mukherjee, Michael Stringer, and Ben Jones. 2013. "Atypical Combinations and Scientific Impact." *Science* 342(6157):468–72.
Vallas, Steven Peter, and Daniel Lee Kleinman. 2008. "Contradiction, Convergence and the Knowledge Economy: The Confluence of Academic and Commercial Biotechnology." *Socio-Economic Review* 6(2):283–311.
Washburn, Jennifer. 2005. *University, Inc.: The Corporate Corruption of American Higher Education*. New York: Basic Books.
Weisberg, Michael, and Ryan Muldoon. 2009. "Epistemic Landscapes and the Division of Cognitive Labor." *Philosophy of Science* 76(2):225–52.
Whitley, Richard. 2000. *The Intellectual and Social Organization of the Sciences*. Oxford, UK: Oxford University Press.
Woolf, Steven H. 2008. "The Meaning of Translational Research and Why It Matters." *Journal of the American Medical Association* 299(2):211–13.
Wright, Brian D. 1983. "The Economics of Invention Incentives: Patents, Prizes, and Research Contracts." *American Economic Review* 73(4):691–707.
Wuchty, Stefan, Benjamin F. Jones, and Brian Uzzi. 2007. "The Increasing Dominance of Teams in Production of Knowledge." *Science* 316(5827):1036–39.
Yao, Lixia, Ying Li, Soumitra Ghosh, James A. Evans, and Andrey Rzhetsky. Forthcoming. "Health ROI as a Measure of Misalignment of Biomedical Needs and Resources." *Nature Biotechnology*.
Zemlyak, Ilona, Nathan Manley, Robert Sapolsky, and Illana Gozes. 2007. "NAP Protects Hippocampal Neurons against Multiple Toxins." *Peptides* 28(10):2004–8.
Zuckerman, Harriet. 1978. "Theory Choice and Problem Choice in Science." *Sociological Inquiry* 48(3–4):65–95.

Jacob G. Foster is Assistant Professor of Sociology at the University of California-Los Angeles. His research interests include science studies, computational social science, social theory, complex networks, and the evolutionary dynamics of culture. He has published in journals like *Science* ("Metaknowledge"), *PNAS* ("Edge Direction and the Structure of Networks"), and *Sociological Science* ("Finding Cultural Holes: How Structure and Culture Diverge in Networks of Scholarly Communication").

Andrey Rzhetsky is Professor of Medicine and Human Genetics at the University of Chicago, where he is Director of the Conte Center for Computational Neuropsychiatric Genomics. His research focuses on the bioinformatics of complex human phenotypes (including disease); biomedical text mining; and the science of science. He has published in journals like *Cell* ("A Nondegenerate Code of Deleterious Variants in Mendelian Loci Contributes to Complex Disease Risk"), *PLoS Computational Biology* ("Quantifying the Impact and Extent of Undocumented Biomedical Synonymy"), and *PNAS* ("Microparadigms: Chains of Collective Reasoning in Publications about Molecular Interactions").

James A. Evans is Associate Professor of Sociology and Senior Fellow of the Computation Institute at the University of Chicago, where he is Director of Knowledge Lab (http://knowledgelab.org) and directs the program in Computational Social Science. He works on the sociology of science and knowledge, computational social science, innovation, complex networks, and content analysis. His work has appeared in journals like *Science* ("Electronic Publication and the Narrowing of Science and Scholarship"), the *American Journal of Sociology* ("Industry Induces Academic Science to Know Less about More") and *Social Studies of Science* ("Industry Collaboration, Scientific Sharing and the Dissemination of Knowledge").