Choosing experiments to accelerate collective discovery

Andrey Rzhetsky[a,b,c,1], Jacob G. Foster[d], Ian T. Foster[b,e], and James A. Evans[b,f,1]

[a]Departments of Medicine and Human Genetics, University of Chicago, Chicago, IL 60637; [b]Computation Institute, University of Chicago and Argonne National Laboratory, Chicago, IL 60637; [c]Institute of Genomic and Systems Biology, University of Chicago, Chicago, IL 60637; [d]Department of Sociology, University of California, Los Angeles, CA 90095; [e]Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60637; and [f]Department of Sociology, University of Chicago, Chicago, IL 60637

Edited by Yu Xie, University of Michigan, Ann Arbor, MI, and approved September 8, 2015 (received for review May 18, 2015)

A scientist's choice of research problem affects his or her personal career trajectory. Scientists' combined choices affect the direction and efficiency of scientific discovery as a whole. In this paper, we infer preferences that shape problem selection from patterns of published findings and then quantify their efficiency. We represent research problems as links between scientific entities in a knowledge network. We then build a generative model of discovery informed by qualitative research on scientific problem selection. We map salient features from this literature to key network properties: an entity's importance corresponds to its degree centrality, and a problem's difficulty corresponds to the network distance it spans. Drawing on millions of papers and patents published over 30 years, we use this model to infer the typical research strategy used to explore chemical relationships in biomedicine. This strategy generates conservative research choices focused on building up knowledge around important molecules. These choices become more conservative over time. The observed strategy is efficient for initial exploration of the network and supports scientific careers that require steady output, but is inefficient for science as a whole. Through supercomputer experiments on a sample of the network, we study thousands of alternatives and identify strategies much more efficient at exploring mature knowledge networks. We find that increased risk-taking and the publication of experimental failures would substantially improve the speed of discovery. We consider institutional shifts in grant making, evaluation, and publication that would help realize these efficiencies.

complex networks | computational biology | science of science | innovation | sociology of science

A scientist's choice of research problem directly affects his or her career. Indirectly, it affects the scientific community. A prescient choice can result in a high-impact study. This boosts the scientist's reputation, but it can also create research opportunities across the field. Scientific choices are hard to quantify because of the complexity and dimensionality of the underlying problem space. In formal or computational models, problem spaces are typically encoded as simple choices between a few options (1, 2) or as highly abstract "landscapes" borrowed from evolutionary biology (3–5). The resulting insight about the relationship between research choice and collective efficiency is suggestive, but necessarily qualitative and abstract.

We obtain concrete, quantitative insight by representing the growth of knowledge as an evolving network extracted from the literature (2, 6). Nodes in the network are scientific concepts and edges are the relations between them asserted in publications. For example, molecules—a core concept in chemistry, biology, and medicine—may be linked by physical interaction (7) or shared clinical relevance (8). Variations of this network metaphor for knowledge have appeared in philosophy (9), social studies of science (10–12), artificial intelligence (13), complex systems research (14), and the natural sciences (7, 15, 16). Nevertheless, networks have rarely been used to measure scientific content (2, 11, 17, 18) and never to evaluate the efficiency of scientific problem selection.

In this paper, we build a model of scientific investigation that allows us to measure collective research behavior in a large corpus of scientific texts and then compare this inferred behavior with more and less efficient alternatives. We define an explicit objective function to quantify the efficiency of a research strategy adopted by the scientific community: the total number of experiments performed to discover a given portion of an unknown knowledge graph. Comparing the modal pattern of "real-science" investigations with hypothetical alternatives, we identify strategies that appear much more efficient for scientific discovery. We also demonstrate that the publication of experimental failures would increase the speed of discovery. In this analysis, we do not focus on which strategies tend to receive high citations or scientific prizes, although we illustrate the relationship between these accolades and research strategies (2).

Our model represents science as a growing network of scientific claims that traces the accumulation of observations and experiments (see Figs. S1–S3). Earlier scientific choices influence subsequent exploration of the network (19). The addition of one redundant link is inconsequential for the topology of science. By contrast, a well-placed new link could radically rewire this network (20). Our model incorporates two key features of problem selection, importance and difficulty, which have received repeated attention in qualitative and quantitative investigations of science. We map these features onto two network properties, degree and distance, which are central to foundational models of network formation and search (21–23). First, scientists typically select "important," central, or well-studied topics on which to anchor their findings and signal their relevance to others' work (10, 24). Our model uses the degree of a concept in the network of claims (i.e., the number of distinct links in which it participates) as a measure of its importance (see Figs. S4–S6). In assuming that scientists' research choices are influenced by concept degree, we posit that scientists are influenced by the choices of others, a well-attested choice heuristic (25, 26). Second, scientists introduce novelty into their work by studying understudied topics and by combining ideas and technologies that others are unlikely to

Significance

Scientists perform a tiny subset of all possible experiments. What characterizes the experiments they choose? And what are the consequences of those choices for the pace of scientific discovery? We model scientific knowledge as a network and science as a sequence of experiments designed to gradually uncover it. By analyzing millions of biomedical articles published over 30 y, we find that biomedical scientists pursue conservative research strategies exploring the local neighborhood of central, important molecules. Although such strategies probably serve scientific careers, we show that they slow scientific advance, especially in mature fields, where more risk and less redundant experimentation would accelerate discovery of the network. We also consider institutional arrangements that could help science pursue these more efficient strategies.

Author contributions: A.R., J.G.F., and J.A.E. designed research; A.R., J.G.F., and J.A.E. analyzed data; and A.R., J.G.F., I.T.F., and J.A.E. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]To whom correspondence may be addressed. Email: arzhetsky@uchicago.edu or jevans@uchicago.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1509757112/-/DCSupplemental.

connect (17, 20). Henri Poincaré (27) and many since (28) have observed that the most generative combinations are "drawn from domains that are far apart" (ref. 27, p. 24). When the concepts under study are more distant, more effort is required to imagine and coordinate their combinations (29). More risk is involved in testing distant claims, because no similar claims have been successful (30).* We operationalize the "cognitive distance" between concepts using their topological distance in the knowledge network. If two concepts are not mutually reachable through the network (i.e., in two distinct components of the network), there is no way a scientist could hypothesize a connection simply by wandering through the literature; conceptual jumps must be made. If two molecules are distant in the network but can reach one another (i.e., they are in the same component), scientists would need to read a range of research articles—likely spread across several journals and subfields—to infer a possible connection (32). Drawing together these insights, we model unlikely combinations as connections between neglected (i.e., low degree), distant, or disconnected concepts within the network of scientific claims.

The Model

In our model, scientists select a pair of entities from all possible pairs and test that pair for an empirical relationship.† Problem selection is guided by a "scientific strategy," which defines the probability of selecting a pair of entities as a function of the importance (degree) of each entity and the difficulty associated with combining them (network distance). Formalizing the studies of scientific behavior above, a strategy is determined by five parameters, which jointly define the probability of examining a relationship between entities i and j at time t:

$$p_{i,j}^{t} \propto \max\left(r_i^{t-1}, r_j^{t-1}\right)^{\alpha_\mu} \times \min\left(r_i^{t-1}, r_j^{t-1}\right)^{\alpha_\iota} \times \begin{cases} \left(\frac{d_{ij}^{t-1}}{d_{\max}}\right)^{-\beta}\left(1-\frac{d_{ij}^{t-1}}{d_{\max}}\right)^{-\gamma} & \text{if } d_{ij}^{t-1} < \infty, \\ e^{\delta} & \text{if } d_{ij}^{t} = \infty. \end{cases}$$

Two parameters define science's preference for the degree centrality r_i^{t-1} of each concept or entity at time t: α_μ controls the weight given to the degree of the more central node—$\max(r_i^{t-1}, r_j^{t-1})$—whereas α_ι controls the weight given to the degree of the less central node.‡ Two parameters (β and γ) define the preference for short and long distance between the pair, if the entities are in the same connected component. The fifth parameter (δ) governs the preference for linking entities in distinct components of a graph (6). When all parameters are zero, the strategy is random: Each pair has a uniform and independent probability of being selected for study. Note that, depending on the parameter values, this model can describe a wide range of strategies; in fact, reducing the number of parameters would involve a priori exclusion of certain functional relationships and hence eliminate some reasonable strategies from empirical consideration. A scientist may prefer a focus on important and/or obscure concepts; short, medium, or long walks between concepts; jumps between concepts in different components; etc. (see Figs. S4 and S5 for illustrations of the model's descriptive plasticity). This flexible framework allows us to dissect an empirical network tracing the history of published research choices. We can also test which network features are most important for scientific search.

We use this model to analyze the growing network of published knowledge about chemical entities in biomedicine (hereafter "biomedical chemistry") from 1976 until 2010. In this knowledge network, the core conceptual entity is a molecule. "Relationships" between molecules can take many forms (2). Molecules may react chemically, physically, or indirectly (the reaction byproduct of one may interact with another). Molecules may be put in relation because they are found in the same part of the body or involved in the same biological process. Or they may be chemically, structurally, or functionally similar. Knowledge about chemical relationships is central to biomedical disciplines at many scales, from organic chemistry, biochemistry, and molecular biology to microbiology, oncology, and pharmacology (Fig. S1). Of course, biomedicine studies entities besides chemicals. Specific diseases are often central to a given publication (oncology is a dramatic example), whereas papers can be further characterized by the methods used (19). Nevertheless, chemical entities can provide a reliable trace (see Fig. S2): A disease is often characterized by its molecular manifestations, and many methods are fundamentally molecular, from radiotracers to green fluorescent protein. To capture these chemical traces of biomedical knowledge, we leverage expert annotations of the MEDLINE database and match the relevant chemical terms into MEDLINE abstracts and the US Patent Database (34) (*Materials and Methods* and *SI Text*). We assume a relationship between chemicals that appear in the same article or patent abstract and infer the underlying research strategies by fitting our model to the resulting network of science.

Results

Before estimating our model, we explored the empirical pattern of degree and distance characterizing the chemicals combined in published articles and patents. Fig. S3 illustrates the conservative nature of most published investigations in biomedical chemistry. The vast majority of chemical relationships combine two chemicals that are moderately central by 2010; in other words, most scientists work in the "core" of various regions of biochemical knowledge (Fig. S3*A*). When chemicals are combined for the first time, they tend to be very close to one another in the network inscribed by prior published work (Figs. S3*B* and S6*D* and Table S1), reflecting a triadic closure mechanism (19). Most links, however, join chemicals that have already been linked before, i.e., chemicals at distance one (2). This conservatism contrasts with the strategies rewarded by high citations and prizes. Consistent with earlier work (2), we find that prize winners and highly cited scientists exhibit a more diverse range of strategies in their published research (see Table S2). Combinations of more and less central chemicals are associated with higher citations and scientific awards (Fig. S3*A*). Further, awards are linked with strategies more likely to bridge disconnected network components (Fig. S3*B*). We then use our generative model to infer the typical search strategy in biomedical chemistry. Encoding this strategy as a set of parameter values in our model allows us to evaluate its efficiency and identify more efficient alternatives by searching through the parameter space.

When we estimate model parameters from the data (Table 1 and *Materials and Methods*), we find that typical (modal) strategies are more likely to combine a relatively "famous" chemical (one with a high degree) with a more obscure one, given the opportunity; $\alpha_\mu > 0$ and $\alpha_\iota < 0$. This is consistent with the emergent empirical degree–degree distribution in Fig. S3*A*. When molecules are in the same component, biomedical scientists typically prefer to combine those close in the network ($\beta < 0$, $\gamma < 0$, $\beta \gg \gamma$). Only rarely do they study chemicals in different connected components (Figs. S3*B* and S6).§ In sum, the typical strategy is oriented toward exploitation, extracting further value

*The notion that linking distant literatures is hard but potentially fruitful underwrites Swanson's work on literature-based discovery (31).

†Scientists often study several entities in combination. This complicates the modeling, so we approximate the discovery process with dyadic strategies.

‡Some values of α_μ and α_ι describe a mechanism analogous to preferential attachment (21, 33), in which researchers choose concepts in proportion to the product of their degrees. Our model encodes many types of preferential attachment, e.g., versions that are superlinear in the degrees. We find that such preferential attachment strategies can be much more efficient for discovery.

§Observed behavior is generated by the interaction between preferences and the evolving set of opportunities. This makes interpretation subtle. For example, when considering chemicals in different connected components, a specific opportunity to combine them would be preferred (i.e., has a higher probability than an opportunity to connect similar chemicals at finite distance). Over time, however, more nodes enter the giant component. Hence, fewer opportunities exist to connect nodes in different components, leading to their small absolute number (Figs. S3*B* and S6).

Table 1. Maximum-likelihood estimates of strategies used in articles and patents, 1980–2010

Model parameter	MEDLINE articles*	US patents*
Preference for degree of the more central chemical, α_μ	1.375 (1.374, 1.377)	1.508 (1.505, 1.512)
Preference for degree of the less central chemical, α_ι	−0.280 (−0.281, −0.280)	−0.172 (−0.175, −0.169)
Preference for network distance between chemicals, β	−5.312 (−5.328, −5.297)	−5.503 (−5.536, −5.469)
Preference for network distance between chemicals, γ	−45.369 (−45.470, −45.268)	−49.579 (−49.799, −49.345)
Preference for bridging disconnected network components, δ	−15.483 (−15.529, −15.428)	−16.303 (−16.399, −16.200)

*Modal estimates; 99% credible intervals in parentheses.

from well-explored regions of the knowledge network (30). Strategies estimated from MEDLINE articles and US patents are very similar, with patents reflecting a slightly more conservative strategy (Fig. S6). When we estimate model parameters for 5-y windows (Fig. 1), we find that scientists have come to focus on more central chemicals (higher α_μ and α_ι). They have become more conservative. They put increased weight on opportunities that explore slightly larger distances, e.g., preferring experiments that result in triadic closure, but these preferences interact with the space of opportunities (including the joint degree distribution) to produce a decreasing fraction of links at distances greater than 1 (Fig. S6*D*). The relative preference for bridging disconnected network components has increased over time, leading to a slight increase in the fraction of links that bridge disconnected components (Fig. S6*A*). In other words, biomedical chemistry has largely become more conservative and more reliant on the exploitation of established knowledge, although it has become slightly more adventurous in bridging disconnected components.

We quantify the efficiency of a scientific strategy in terms of the total number of experiments performed, relative to the number of discoveries made, i.e., the number of new connections identified. We define $X\%$ loss as the total number of experiments performed (edges tested) before discovering $X\%$ of all edges in the target network. Relative loss is the number of experiments performed divided by the number of novel edges discovered plus one. Relative loss measures the number of experiments performed to discover one new chemical relationship (network edge). Strategies with larger relative loss continue to investigate previously explored relationships or test relationships that do not exist. This definition of efficiency implies that the objective function of science prizes the discovery of novel relationships above all else. We consider alternative objective functions in *Discussion*.

To estimate the efficiency of the inferred strategy and compare it with alternatives, we drew a subsample from the empirical network with a similar degree distribution (Figs. S7–S10 and *SI Text*) (14, 21). Then we used a supercomputer to simulate the exploration of this sample network with thousands of different strategies. Based on these simulations, we calculated the average cost (relative loss) associated with each strategy and minimize this cost with simulated annealing. By simulating the discovery process hundreds of times for each parameter setting, we identify strategies that, on average, discover a given proportion of the network with greatest efficiency. This exploration was computationally intensive and could not be conducted on the full network; in fact, our project was a use case for the development of novel parallel programming approaches (35, 36). Recall that each strategy (i.e., set of parameters) prioritizes different kinds of experiments. A given experiment "succeeds" if it proposes a relationship that is realized in the empirical (sample) network. Successful experiments are added to the network ("published"). An experiment "fails" if it proposes a relationship that is never realized in the empirical network. In the simplest scenario considered here, failures are not published. In our model, both "success" and "failure" are error free; i.e., there are no false positives or false negatives; this could be relaxed at the cost of considerable complexity and additional modeling assumptions.

Fig. 1 and Table S1 show the optimal parameters we discovered for uncovering $X\%$ of the network ($X\%$ relative loss optimized; *SI Text*). We find that connecting multiple important, high-degree chemicals (positive α_μ and α_ι) is critical for early exploration—discovering the first 10–20%—of the network. Instead of exploiting the local neighborhood of a high-profile chemical (like MEDLINE), this strategy prioritizes connections between important chemicals at either very short or infinite distance, a combinatorial exploitation strategy (driven by preferential attachment) similar to interdisciplinary approaches that mine connections between fields. By contrast, efficient discovery of 100% of the target network requires a disposition to link distant chemicals (in the same network component and disconnected components), whereas degree becomes less important. This strategy engages considerable risk, as it attempts to establish links spanning substantial cognitive distance and relies less on imitating prior scientists' chemical choices. Fig. 1 illustrates how the history of inferred strategies for chemical discovery diverges from our estimated optimal discovery strategies. Historical strategies have become more conservative each year, as scientists focus on more central (i.e., higher-degree) chemicals. By contrast, optimal discovery strategies trend in the opposite direction.

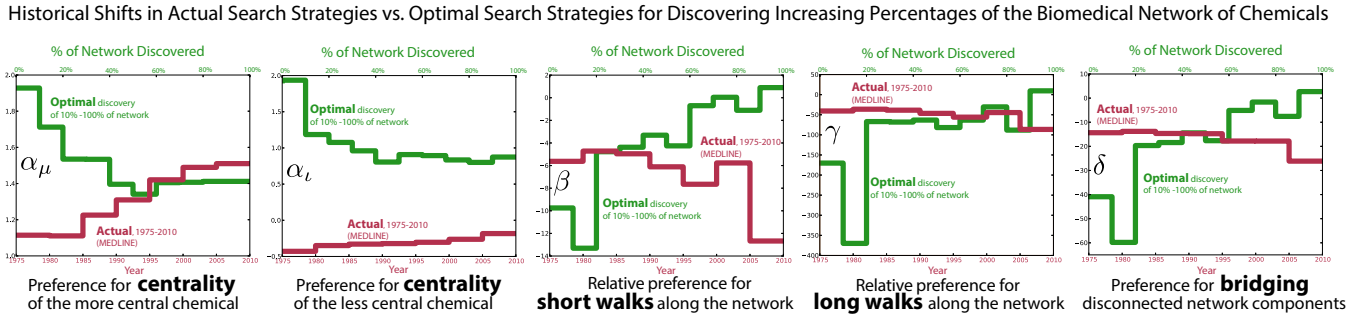

Fig. 1. Red lines show model parameters estimated from the network of published chemical relationships over historical time, 1975–2010, every 5 y. The preference for more central chemicals (α_μ, α_ι) increases consistently over time. The parameters controlling preference for walk length (β, γ) and for jumping to disconnected network components (δ) also decrease consistently between 1975 and 2010, although the interpretation is somewhat subtle (main text). The green lines illustrate the optimal 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, and 100% strategies against the historical trend and highlight the contrast between the trajectories.

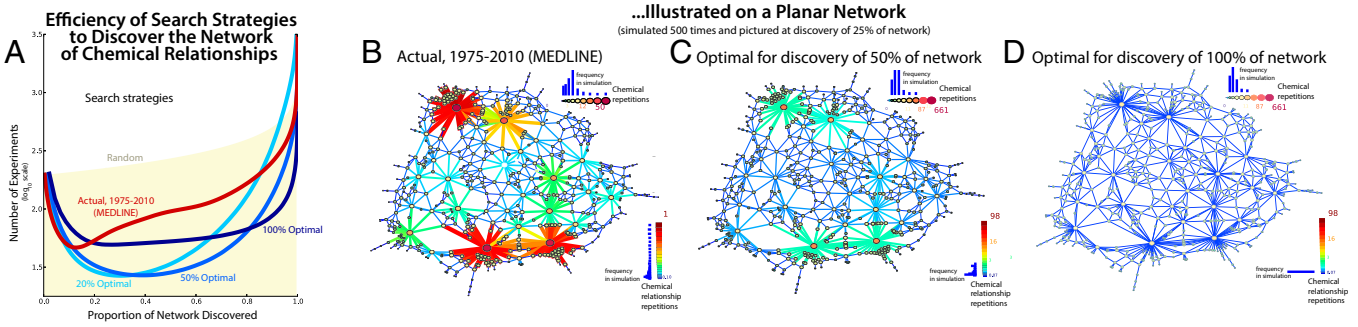

Fig. 2. (*A*) Comparison of the efficiency of discovery for different search strategies. Efficiency is quantified as the estimated number of experiments required to discover from 1% to 100% of a representative sample of the 2010 MEDLINE network. Compared strategies include random choice, the inferred MEDLINE strategy, and optimal strategies for discovering 20%, 50%, and 100% of the network. Results show that contemporary scientific activity (MEDLINE) may have been nearly optimal for discovering 10% of the chemical network, but becomes increasingly inefficient for discovering more than 30%. Parameters for "optimal" strategies are drawn from multistage collections of simulated annealing and subsequent MCMC search procedures. (*B–D*) Actual and optimal search processes illustrated on a planar network of chemical relationships. Each panel represents the average from 500 independent runs of the strategy, at the point where 25% of the possible chemical relationships have been discovered. The node and edge legends for each network strategy (*Upper Right* and *Lower Right* of each panel) are normalized to highlight differences between the strategies and are paired with histograms to illustrate the frequencies with which chemicals and chemical relationships of various degree centralities are selected for experimentation. Panels compare the strategies used by biomedical scientists publishing MEDLINE-indexed articles with alternative strategies that most efficiently discover the first 50% or 100% of the network.

Early on they leverage existing knowledge by linking high-profile chemicals and then become riskier as they attend less to prior chemical choices and attempt more distant combinations.

Fig. 2*B* visualizes the strategy estimated from MEDLINE data (Table 1) on a planar network, whereas Fig. 2 *C* and *D* illustrates the 50% and 100% optimal strategies for exploring that network. Results are shown after 25% of the network has been discovered. The planar nature of the network makes the aggregate effect of these strategies apparent (see bit.ly/16QRviz for an animation of the MEDLINE strategy and *SI Text* for other animation URLs). In particular, Fig. 2*B* highlights the tendency of the MEDLINE strategy to explore the neighborhood of prominent chemicals. Fig. 2*A* compares the efficiency of several search strategies, quantified as the estimated number of experiments each strategy requires to discover from 1% to 100% of the network. Efficiency results use the sample network drawn from MEDLINE (Fig. S7 and *SI Text*). The MEDLINE strategy—which works well for generating coherent, thoroughly explored islands of knowledge in a young knowledge network—rapidly becomes inefficient, as effort is wasted by excessive focus on a few key entities and repetition of known connections. The MEDLINE strategy reaches maximum efficiency at 13% of the network discovered (Fig. 2*A*); it becomes increasingly inefficient at discovering larger fractions of the network. Although more efficient than the random strategy, which tests all edges with equal probability, the MEDLINE strategy is over three times more costly than the most efficient alternative when discovering the bulk of the network. By contrast, the optimal strategy for uncovering 50% of the network is nearly 10 times more efficient than the random strategy for discovering a range of network fractions (Table S1, 50% of network discovered).

Judging from the distribution of distances spanned in the empirical network (Figs. S3*B* and S6*D*), researchers rarely wander far across the knowledge network or bridge disconnected chemicals. Such behavior is critical for advance in mature areas of science (Table S1), and award-winning scientists appear to do so more frequently ("Prizes" in Fig. S3*B*). Scientists may hesitate to undertake a long walk or jump because of the low chance of success, even though a successful outcome could reveal the structure of the larger network and stimulate further work. In this way, individual incentives for productivity, reinforced by institutions like tenure, may be at odds with science's collective interest in maximizing discovery.

Patterns in the parameters associated with efficient discovery reveal strategies that are consistently important. Preference for a high degree of the more central chemical is consistently associated with efficient discovery, which underscores the importance of preferential attachment as a mechanism for guiding research (21, 33). Preference for the degree of the less central chemical declines more rapidly (Fig. 1). Distinct preferences for the network distance between chemicals are rarely (clearly) implicated in optimal search (i.e., efficient discovery is robust to considerable variation in these parameters; see the ranges containing 95% of sampled parameters in Table S1).

Beyond alternative strategies, we also consider an alternative institution: "coordinated" discovery, in which scientists publish all findings, positive and negative, and do not repeat experiments. We calculate the efficiency of this regime analytically for a random strategy and numerically estimate the efficiency of coordinated MEDLINE and optimal discovery strategies on the sample network (*SI Text*). Coordinating research decreases the costs of discovery—the number of failed or duplicate experiments—regardless of strategy (Fig. S11).

Discussion

Our paper provides a quantitative method for inferring research strategies from data, examining their consequences, and discovering more efficient alternatives. Nevertheless, our model has several limitations. First, by modeling discovery as the exploration of a hidden network, we imply that new discoveries in biomedical chemistry always link chemicals never linked before. This is not true; a tie between two chemicals may be novel because previously linked chemicals are now linked in a new way (e.g., in the context of a new disease). Second, we evaluate efficiency only in relation to a single objective function—maximizing discovery of novel links. Many other objective functions exist, including minimizing error or increasing the robustness of discovered knowledge (37, 38). Development of useful medical and industrial technologies relies on the productive repetition of molecular relationships, which our current objective function does not reward. These alternative goals could be incorporated into the evaluation of our model—for example, by allowing repeat explorations of a known relationship to contribute to the objective function, but with diminishing marginal returns. Note that individual scientists may "locally" hold objectives that are very different from the global objective function of science. They may optimize their total number of publications or the predicted number of future citations to their work (2, 20). That said, we believe that the broader scientific community does have an implicit objective to traverse the space of possible research problems in search of novel and useful knowledge, and so we use that as a baseline here (39, 40). Third, we found that the most efficient discovery strategies are dominated by preferential attachment to the most central chemical in a problem over preferences

for the degree of the less central chemical or the distance between chemicals. In the empirical network, most published links connect chemicals that have already been connected, and most novel links connect chemicals at distance two. Chemicals in disconnected components are next most frequent. All other distances are extremely rare. We could thus construct more parsimonious models that focus on preferential attachment and a few distinct categories of distance to describe the most efficient discovery strategies. Finally, our empirical network of published chemical relationships represents an imperfect sample of research effort, as effort is screened by experimental failure and the greater challenge of publishing an unconventional paper vs. an incremental one. The sample is drawn from virtually every publishing biomedical scientist, but publications overwhelmingly document successful experiments. Our data almost certainly underrepresent the risky but unsuccessful choices made by individual scientists. It nevertheless represents an informative trace of the scientific process—the very trace that scientists themselves use as they read the literature and design new experiments to build upon it. A more complete record of failures would both deepen our understanding of research behavior and improve its efficiency.[¶] Future models could introduce these layers as additional features, penalties, and rewards associated with the "game" of science.

Despite these limitations, our model reveals patterns of discovery in biomedical chemistry and shows that more efficient discovery strategies would incorporate more "interdisciplinarity" and more risk, with the latter particularly important as a field matures. Efficient discovery of radically new knowledge in a mature field, including many areas of biomedicine (42), requires abandoning the current focus on important, nearby chemicals. Adopting a more efficient approach would lead to greater risk, but our findings suggest that scientists pursue progressively less risk, focusing more and more on the immediate neighborhood of high-degree chemicals, with the slight increase in bridging links as a silver lining. Successful research that goes against the crowd is more likely to garner high citations and prizes (Fig. S3), but these incentives may not be sufficient or flexible enough to motivate sustained advance in mature fields. A shift to riskier research would lead to more failures, which typically remain unpublished under current publication norms. We find that publication of failures substantially increases the speed of discovery. Thus, science policy could improve the efficiency of discovery by subsidizing more risky strategies, incentivizing strategy diversity, and encouraging the publication of failed experiments.

Policymakers could design institutions that cultivate intelligent risk-taking by shifting evaluation from the individual to the group, as was done at Bell Labs (43). They could also fund promising individuals rather than projects, like the Howard Hughes Medical Institute (44). Both approaches incentivize the spreading of risk across a portfolio of experiments that reflect multiple research strategies, instead of evaluating each experiment separately and selecting safer opportunities. Science and technology policy might also promote risky experiments with large potential benefits by lowering barriers to entry and championing radical ideas, emulating the Gates Foundation's Grand Challenges program. Finally, new incentives to publish failures, like those mandating web publication of clinical trials at https://www.clinicaltrials.gov, should be considered if risk-taking increases. With carefully designed incentives and institutions, scientists will choose the next experiment to benefit themselves, science, and society.

Materials and Methods

Data. We examine scientific discovery by analyzing the growth of the knowledge network in biomedical chemistry since 1976. We constructed this network by matching a large lexicon of 52,654 distinct chemical terms extracted from MEDLINE metadata into MEDLINE from 1976 to 2010 (34) and then inferring a chemical relationship when the terms appeared in the same abstract. We used the same procedure to extract chemical relationships within US patents (*SI Text*). This process resulted in 30,060 unique chemicals with at least one link to others and 12,342,474 links between chemicals, corresponding to 1,338,753 unique chemical relationships. This network represents accumulated chemical knowledge within 2,363,858 articles and 295,812 patents. The combined network has a broad, approximately log-normal degree distribution (45, 46) (Fig. S8).

Estimating and Simulating Strategies. In estimating parameters from the MEDLINE and US Patent networks, we considered time-dependent snapshots of the "visible" connectivity of each chemical within the growing chemical network to compute time-dependent choice probabilities. We can then compute the full likelihood of selecting a sequence of edge sets for experimentation, given model parameters. Parameter estimates are obtained by maximizing this likelihood function with respect to parameter values (*SI Text*). We used simulated annealing to find the maximum-likelihood estimates and Markov chain Monte Carlo (MCMC) to explore the parameter space around these estimates and assign Bayesian credible intervals (Table 1 and *SI Text*). We used the same approach to explore the parameter space of the model on the sample network: simulated annealing to identify an initial estimate, followed by MCMC to explore the objective function in the neighborhood of that estimate. Because of the heuristic nature of simulated annealing and MCMC, there are no formal guarantees on the global optimality of discovered strategies. Our purpose is less to establish claims of global optimality and more to demonstrate that much more efficient strategies exist and can be discovered.

Appendix

See *SI Text* for more details about the data, further definition and characterization of the model, analysis of the empirical network, and strategy comparisons.

ACKNOWLEDGMENTS. We thank Mike Wilde and Tim Armstrong for Swift implementations of our code on the Beagle supercomputer; Mike Papka and Ti Leggett for help parallelizing our code; research assistants Mahmoud Bahrani, Simo Huang, David Kates, Val Michelman, and Nathan Worchester for help compiling prize-winner data, as well as testing our annotations; Stefano Allesina, Carl Bergstrom, and anonymous reviewers for comments on the manuscript; the Computation Institute at University of Chicago for access to and help with the Peta-Scale Active Data Store and Beagle; Thomson Reuters for citation information; Jeff Alstott for help with his Python package powerlaw; and Enthought, Inc., for help with Python programming. This work was supported by National Science Foundation Grant SBE 0915730, National Institutes of Health Grants 1P50MH094267 and U01HL108634-01, Defense Advanced Research Projects Agency Big Mechanism contract W911NF1410333, AFOSR Grant FA9550-15-1-0162, and a John Templeton Foundation grant to the Metaknowledge Network.

[¶]We assume that published research reflects the underlying distribution of research effort in a relatively undistorted way. Recent survey data on scientific choice are consistent with this assumption (41). Although we interpret Fig. 1 and Fig. S6*D* to imply that scientists pursue less risky projects over time, it is possible that scientists pursue such projects with the same intensity, but that fewer succeed and are published in later periods. We cannot tackle this issue directly, but consider how effort is screened by experimental failure, publication bias, etc., to produce the distribution of published choices. Our interpretation assumes that although a priori "risky" strategies (like combining two distant, low-degree chemicals) may fail more often than conservative alternatives, the risk is not so high that the published record no longer reflects the underlying distribution of effort. It also requires that risky strategies do not become much riskier over time. If the selection process has these plausible properties—i.e., it is well behaved and near stationary—then changes in the observed distribution and inferred parameters will track changes in the unobserved distribution of research effort and scientific choice.

1. Kleinberg J, Oren S (2011) *Mechanisms for (Mis)Allocating Scientific Credit, STOC '11* (Association for Computing Machinery, New York), pp 529–538.
2. Foster JG, Rzhetsky A, Evans JA (2015) Tradition and innovation in scientists' research strategies. *Am Sociol Rev* 80(5):875–908.
3. Weisberg M, Muldoon R (2009) Epistemic landscapes and the division of cognitive labor. *Philos Sci* 76(2):225–252.
4. Mason W, Watts DJ (2012) Collaborative learning in networks. *Proc Natl Acad Sci USA* 109(3):764–769.
5. Tria F, Loreto V, Servedio VDP, Strogatz SH (2014) The dynamics of correlated novelties. *Sci Rep* 4:5890.
6. Cokol M, Iossifov I, Weinreb C, Rzhetsky A (2005) Emergent behavior of growing knowledge about molecular interactions. *Nat Biotechnol* 23(10):1243–1247.
7. Grzybowski BA, Bishop KJM, Kowalczyk B, Wilmer CE (2009) The 'wired' universe of organic chemistry. *Nat Chem* 1(1):31–36.
8. Rzhetsky A, Wajngurt D, Park N, Zheng T (2007) Probing genetic overlap among complex human phenotypes. *Proc Natl Acad Sci USA* 104(28):11694–11699.

9. Quine WV (1951) Main Trends in Recent Philosophy: Two Dogmas of Empiricism. *The Philosophical Review* 60(1):20–43.
10. Latour B (1987) *Science in Action: How to Follow Scientists and Engineers Through Society* (Harvard Univ Press, Cambridge, MA).
11. Callon M, Law J, Rip A (1986) *Mapping the Dynamics of Science and Technology: Sociology of Science in the Real World* (Macmillan, New York).
12. Evans JA (2010) Industry collaboration, scientific sharing and the dissemination of knowledge. *Soc Stud Sci* 40(5):757–791.
13. Newell A, Simon HA (1972) *Human Problem Solving* (Prentice Hall, Englewood Cliffs, NJ).
14. Newman M (2003) The structure and function of complex networks. *SIAM Rev* 45(2): 167–256.
15. Hewett M, et al. (2002) PharmGKB: The pharmacogenetics knowledge base. *Nucleic Acids Res* 30(1):163–165.
16. Gothard CM, et al. (2012) Rewiring chemistry: Algorithmic discovery and experimental validation of one-pot reactions in the network of organic chemistry. *Angew Chem Int Ed Engl* 51(32):7922–7927.
17. Uzzi B, Mukherjee S, Stringer M, Jones B (2013) Atypical combinations and scientific impact. *Science* 342(6157):468–472.
18. Beam E, Appelbaum LG, Jack J, Moody J, Huettel SA (2014) Mapping the semantic structure of cognitive neuroscience. *J Cogn Neurosci* 26(9):1949–1965.
19. Shi F, Foster JG, Evans JA (2015) Weaving the fabric of science: Dynamic network models of science's unfolding structure. *Soc Networks* 43:73–85.
20. Bourdieu P (1975) The specificity of the scientific field and the social conditions for the progress of reason. *Soc Sci Inf (Paris)* 14(6):19–47.
21. Barabasi AL, Albert R (1999) Emergence of scaling in random networks. *Science* 286(5439):509–512.
22. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. *Nature* 393(6684):440–442.
23. Kleinberg JM (2000) Navigation in a small world. *Nature* 406(6798):845.
24. Latour B, Woolgar S (1986) *Laboratory Life: The Construction of Scientific Facts* (Princeton Univ Press, Princeton).
25. Easley D, Kleinberg J (2010) *Networks, Crowds, and Markets: Reasoning About a Highly Connected World* (Cambridge Univ Press, Cambridge, UK).
26. Salganik MJ, Dodds PS, Watts DJ (2006) Experimental study of inequality and unpredictability in an artificial cultural market. *Science* 311(5762):854–856.
27. Poincaré H (1952) *The Creative Process: A Symposium*, ed Ghiselin B (Univ of California Press, Berkeley, CA).
28. Carnabuci G, Bruggeman J (2009) Knowledge specialization, knowledge brokerage and the uneven growth of technology domains. *Soc Forces* 88(2):607–641.
29. Leahey E, Beckman C, Stanko T (2013) Prominent but Less Productive: The Impact of Interdisciplinarity on Scientists' Research. arXiv:1510.06802.
30. March JG (1991) Exploration and exploitation in organizational learning. *Organ Sci* 2(1):71–87.
31. Swanson DR (1990) Medical literature as a potential source of new knowledge. *Bull Med Libr Assoc* 78(1):29–37.
32. Burt RS (2004) Structural holes and good ideas. *Am J Sociol* 110(2):349–399.
33. Barabasi AL, et al. (2002) Evolution of the social network of scientific collaborations. *Physica A Stat Mech Appl* 311(3-4):590–614.
34. Krauthammer M, Rzhetsky A, Morozov P, Friedman C (2000) Using BLAST for identifying gene and protein names in journal articles. *Gene* 259(1-2):245–252.
35. Wozniak J, et al. (2013) Swift/T: Large-scale application composition via distributed-memory dataflow processing. 95–102. *13th International Symposium on Cluster, Cloud and Grid Computing* (Institute for Electrical and Electronics Engineers, New York), pp 95–102.
36. Wozniak JM, et al. (2012) Turbine: A distributed-memory dataflow engine for extreme-scale many-task applications. *Proceedings of the 1st ACM SIGMOD Workshop on Scalable Workflow Execution Engines and Technologies*. (Association for Computing Machinery, New York), pp 5:1–5:12.
37. Ioannidis JP (2005) Why most published research findings are false. *PLoS Med* 2(8): e124.
38. Ioannidis JPA (2014) How to make more published research true. *PLoS Med* 11(10): e1001747.
39. Merton RK (1957) Priorities in scientific discovery: A chapter in the sociology of science. *Am Sociol Rev* 22:635–659.
40. Gieryn TF (1978) Problem retention and problem change in science. *Sociol Inq* 48(3–4): 96–115.
41. Bateman TS, Hess AM (2015) Different personal propensities among scientists relate to deeper vs. broader knowledge contributions. *Proc Natl Acad Sci USA* 112(12): 3653–3658.
42. Jones BF (2009) The burden of knowledge and the "death of the renaissance man": Is innovation getting harder? *Rev Econ Stud* 76(1):283–317.
43. Gernter J (2012) *The Idea Factory: Bell Labs and the Great Age of American Innovation* (Penguin, New York).
44. Azoulay P, Stuart T, Wang Y (2014) Matthew: Effect or Fable? *Management Science* 60(1):92–109.
45. Clauset A, Shalizi C, Newman M (2009) Power-law distributions in empirical data. *SIAM Rev* 51(4):661–703.
46. Alstott J, Bullmore E, Plenz D (2014) Powerlaw: A Python package for analysis of heavy-tailed distributions. *PLoS One* 9(1):e85777.
47. Medawar PB (1967) *The Art of the Soluble* (Methuen, London).
48. Peirce CS (1878) The probability of induction. *Popular Science Monthly* 12:705–718.
49. Busch L, Lacy W, Sachs C (1983) Perceived criteria for research problem choice in the agricultural sciences-A research note. *Soc Forces* 62(1):190–200.
50. Zuckerman H (1978) Theory choice and problem choice in science. *Sociol Inq* 48(3–4): 65–95.
51. Knorr KD (1981) *The Social Process of Scientific Investigation*, Sociology of the Sciences A Yearbook, eds Knorr KD, Krohn R, Whitley R (Springer, Dordrecht, The Netherlands) Vol 4, pp 25–52.
52. Kitcher P (1990) The division of cognitive labor. *J Philos* 87(1):5–22.
53. Payette N (2012) *Models of Science Dynamics, Understanding Complex Systems*, eds Scharnhorst A, Brner K, Besselaar Pvd (Springer, Berlin), pp 127–157.
54. Ding WW, Murray F, Stuart TE (2006) Gender differences in patenting in the academic life sciences. *Science* 313(5787):665–667.
55. Guimerà R, Uzzi B, Spiro J, Amaral LAN (2005) Team assembly mechanisms determine collaboration network structure and team performance. *Science* 308(5722):697–702.
56. Shwed U, Bearman PS (2010) The temporal structure of scientific consensus formation. *Am Soc Rev* 75(6):817–840.
57. Wuchty S, Jones BF, Uzzi B (2007) The increasing dominance of teams in production of knowledge. *Science* 316(5827):1036–1039.
58. Petersen AM, et al. (2014) Reputation and impact in academic careers. *Proc Natl Acad Sci USA* 111(43):15316–15321.
59. Kuhn T (1962) *The Structure of Scientific Revolutions* (Univ of Chicago Press, Chicago).
60. Johnson NL, Kotz S, Balakrishnan N (1995) *Continuous Univariate Distributions* (Wiley, New York).
61. Snijders TA (2002) Markov chain Monte Carlo estimation of exponential random graph models. *J Soc Struct 3* 3(2):1–40.
62. Kirkpatrick S, Gelatt CD, Jr, Vecchi MP (1983) Optimization by simulated annealing. *Science* 220(4598):671–680.
63. Torvik VI, Weeber M, Swanson DR, Smalheiser NR (2005) A probabilistic similarity metric for medline records: A model for author name disambiguation. *J Am Soc Inf Sci Technol* 56(2):140–158.
64. Goodman LA (1965) On simultaneous confidence intervals for multinomial proportions. *Technometrics* 7(2):247–254.
65. Leskovec J, Faloutsos C (2006) Sampling from large graphs. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, New York), pp 631–636.
66. Hidiroglou MA (1978) An approximation of the inverse moments of the positive hypergeometric distribution. *Communications in Statistics-Theory and Methods* 7(15): 1475–1487.