The Diversity–Innovation Paradox in Science

Bas Hofstra[a,1], Vivek V. Kulkarni[b], Sebastian Munoz-Najar Galvez[a], Bryan He[b], Dan Jurafsky[b,c], and Daniel A. McFarland[a,1]

[a]Graduate School of Education, Stanford University, Stanford, CA 94305; [b]Department of Computer Science, Stanford University, Stanford, CA 94305; and [c]Department of Linguistics, Stanford University, Stanford, CA 94305

Edited by Peter S. Bearman, Columbia University, New York, NY, and approved March 16, 2020 (received for review September 5, 2019)

Prior work finds a diversity paradox: Diversity breeds innovation, yet underrepresented groups that diversify organizations have less successful careers within them. Does the diversity paradox hold for scientists as well? We study this by utilizing a near-complete population of ∼1.2 million US doctoral recipients from 1977 to 2015 and following their careers into publishing and faculty positions. We use text analysis and machine learning to answer a series of questions: How do we detect scientific innovations? Are underrepresented groups more likely to generate scientific innovations? And are the innovations of underrepresented groups adopted and rewarded? Our analyses show that underrepresented groups produce higher rates of scientific novelty. However, their novel contributions are devalued and discounted: For example, novel contributions by gender and racial minorities are taken up by other scholars at lower rates than novel contributions by gender and racial majorities, and equally impactful contributions of gender and racial minorities are less likely to result in successful scientific careers than for majority groups. These results suggest there may be unwarranted reproduction of stratification in academic careers that discounts diversity's role in innovation and partly explains the underrepresentation of some groups in academia.

diversity | innovation | science | inequality | sociology of science

Innovation drives scientific progress. Innovation propels science into uncharted territories and expands humanity's understanding of the natural and social world. Innovation is also believed to be predictive of successful scientific careers: Innovators are science's trailblazers and discoverers, so producing innovative science may lead to successful academic careers (1). At the same time, a common hypothesis is that demographic diversity brings such innovation (2–5). Scholars from underrepresented groups have origins, concerns, and experiences that differ from groups traditionally represented, and their inclusion in academe diversifies scholarly perspectives. In fact, historically underrepresented groups often draw relations between ideas and concepts that have been traditionally missed or ignored (4–7). Given this, if demographic groups are unequally represented in academia, then one would expect underrepresented groups to generate more scientific innovation than overrepresented groups and have more successful careers (*SI Appendix*). Unfortunately, the combination of these two relationships—diversity–innovation and innovation–careers—fails to result and poses a paradox. If gender and racially underrepresented scholars are likely to innovate and innovation supposedly leads to successful academic careers, then how do we explain persistent inequalities in scientific careers between minority and majority groups (8–13)? One explanation is that the scientific innovations produced by some groups are discounted, possibly leading to differences in scientific impact and successful careers.

In this paper, we set out to identify the diversity–innovation paradox in science and explain why it arises. We provide a system-level account of science using a near-complete population of US doctorate recipients (∼1.2 million) where we identify scientific innovations (14–19) and analyze the rates at which different demographic groups relate scientific concepts in novel ways, the extent to which those novel conceptual relations get taken up by other scholars, how "distal" those linkages are (14), and the subsequent returns they have to scientific careers. Our analyses use observations spanning three decades, all scientific disciplines, and all US doctorate-awarding institutions. Through them we are able 1) to compare minority scholars' rates of scientific novelty vis-à-vis majority scholars and then ascertain whether and why their novel conceptualizations 2) are taken up by others and, in turn, 3) facilitate a successful research career.

Innovation as Novelty and Impactful Novelty in Text

Our dataset stems from ProQuest dissertations (20), which includes records of nearly all US PhD theses and their metadata from 1977 to 2015: student names, advisors, institutions, thesis titles, abstracts, disciplines, etc. These structural and semantic footprints enable us to consider students' rates of innovation at the very onset of their scholarly careers and their academic trajectory afterward, i.e., their earliest conceptual innovations and how they correspond to successful academic careers (21). We link these data with several data sources to arrive at a near-complete ecology of US PhD students and their career trajectories. Specifically, we link ProQuest dissertations to the US Census data (2000 and 2010) and Social Security Administration data (1900 to 2016) to infer demographic information on students' gender and race (i.e., name signals for white, Asian, or underrepresented minority [Hispanic, African American, or Native American]; see *Materials and Methods* and *SI Appendix*); we link ProQuest dissertations to Web of Science, a large-scale publication database with ∼38 million academic publications (1900 to 2017), to find out which students have continued research careers, and we weigh our inferential analyses by population records of the number of PhD recipients for each distinct university–year combination to render results generalizable to the population (*SI Appendix*).

Significance

By analyzing data from nearly all US PhD recipients and their dissertations across three decades, this paper finds demographically underrepresented students innovate at higher rates than majority students, but their novel contributions are discounted and less likely to earn them academic positions. The discounting of minorities' innovations may partly explain their underrepresentation in influential positions of academia.

Author contributions: B. Hofstra, V.V.K., and D.A.M. designed research; B. Hofstra, V.V.K., S.M.-N.G., B. He, D.J., and D.A.M. performed research; B. Hofstra, V.V.K., S.M.-N.G., B. He, D.J., and D.A.M. contributed new reagents/analytic tools; B. Hofstra, V.V.K., S.M.-N.G., B. He, D.J., and D.A.M. analyzed data; and B. Hofstra and D.A.M. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

See online for related content such as Commentaries.

[1]To whom correspondence may be addressed. Email: bhofstra@stanford.edu or mcfarland@stanford.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1915378117/-/DCSupplemental.

First published April 14, 2020.

To measure scientific innovation, we first identify the set of scientific concepts being employed in theses. For this, we use natural language processing techniques of phrase extraction and structural topic modeling (22, 23) to identify terms representing substantive concepts in millions of documents (# concepts, mean = 56.500; median = 57; SD = 19.440; see *Materials and Methods* and *SI Appendix*, Table S1) (24). Next, we filter and identify when pairs of meaningful concepts are first related to one another in a thesis. By summing the number of novel conceptual co-occurrences within each thesis, we develop a measure of how conceptually novel a thesis and author are (# new links)—their novelty. However, not all novel conceptual linkages are taken up in ensuing works and have the same impact on scholarship. To capture impactful novelty, we measure how often a thesis's new conceptual linkages are adopted in ensuing documents of each year (uptake per new link) (Fig. 1).

Our broad perspective on innovation mirrors key theoretical perspectives on scientific innovation, where "science is the constellation of facts, theories, and methods collected in current texts" (28). Scientific development is then the process where concepts are added to the ever-growing "constellation"—i.e., our accumulating corpus of texts—in new combinations: The introduction of new links between scientific concepts (14, 15, 28–30). As such, our conception of novelty as the number of unique recombinations of scientific concepts (# new links, mean = 9.026; median = 4; SD = 13.744; 20.9% of students do not introduce links) and impactful novelty as the average future adoption of these unique recombinations (uptake per new link, mean = 0.790; median = 0.333; SD = 3.079) reflects different notions of scientific innovation. Novelty in itself does not automatically imply innovation, nor is the future adoption of novelty a prerequisite to innovation—for example, which novelty gets adopted may be in itself a function of structural processes. The advantage of our focus on conceptual recombination compared to citation metrics for innovation is that it is insensitive to 1) prioritizing some academic disciplines over others with regard to journal indexing and 2) the plethora of reasons as to why scholars cite other work (31, 32).

Results

Who introduces novelty and whose novelty is impactful? We first model individual rates of novelty (# new links) and impactful

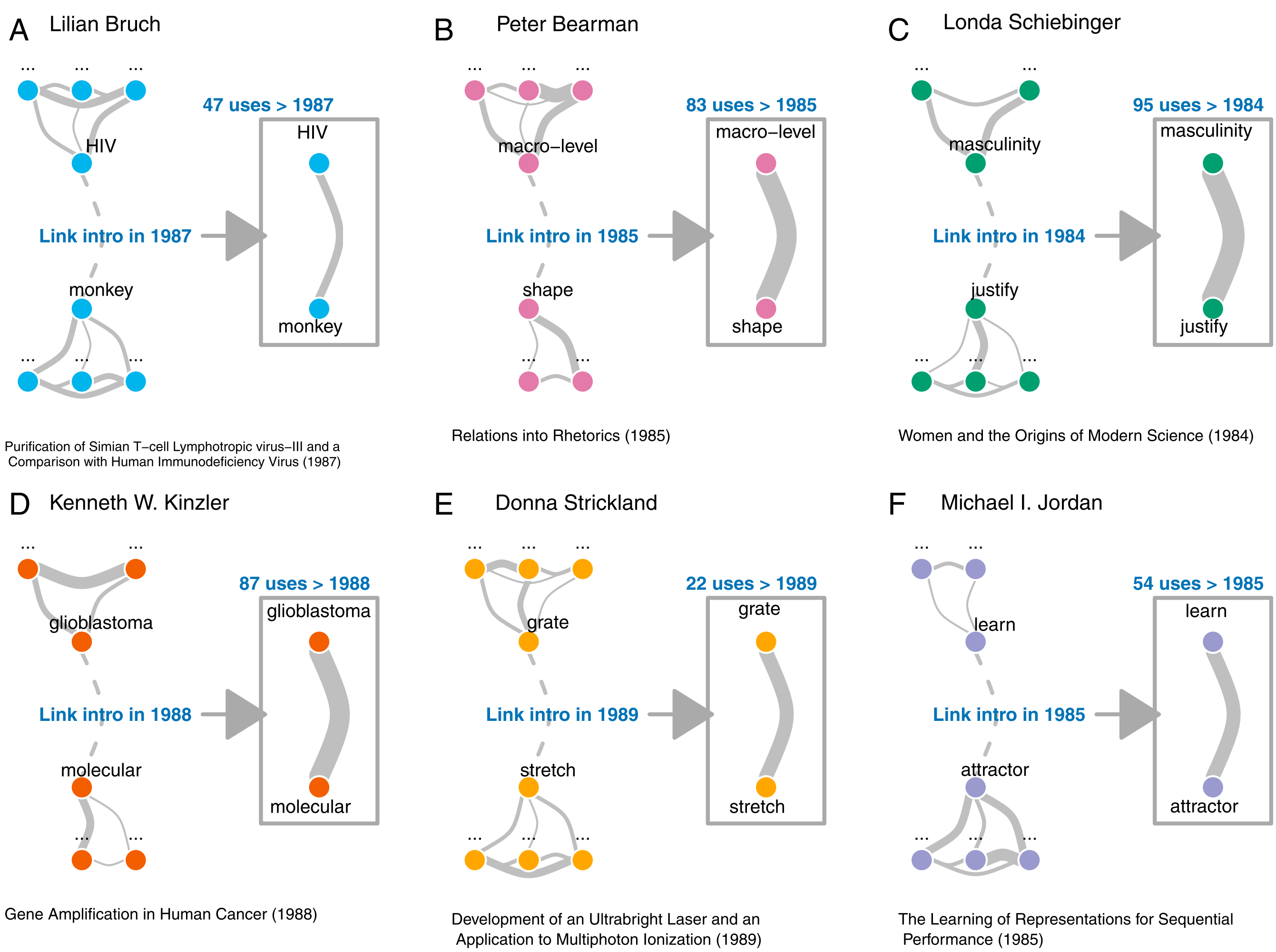

Fig. 1. The introduction of innovations and their subsequent uptake. (*A*–*F*) Examples drawn from the data illustrate our measures of novelty and impactful novelty. Nodes represent concepts, and link thickness indicates the frequency of their co-usage. Students can introduce new links (dotted lines) as their work enters the corpus. These examples concern novel links taken up at significantly higher rates than usual (e.g., 95 uses of Schiebinger's link after 1984). The mean (median) uptake of new links is 0.790 (0.333), and ∼50% of new links never gets taken up. (*A*) Lilian Bruch was among the pioneering HIV researchers (25), and her thesis introduced the link between "HIV" and "monkeys," indicating innovation in scientific writing as HIV's origins are often attributed to nonhuman primates. (*C*) Londa Schiebinger was the first to link "masculinity" with "justify," reflecting her pioneering work on gender bias in academia (26). (*E*) Donna Strickland won the 2018 Nobel Prize in Physics for her PhD work on chirped pulse amplification, utilizing grating-based stretchers and compressors (27).

novelty (uptake per new link) by several notions of demographic diversity, the gender and racial representation in a student's discipline, and by gender/race indicators reflecting historically underrepresented groups (Fig. 2). We keep institution, academic discipline, and graduation year constant (33, 34) (see *Materials and Methods* and *SI Appendix*, Figs. S1 and S4 and Table S2). We find that the more students are underrepresented genders ($P < 0.001$) or races ($P < 0.05$) in their discipline, the more they are likely to introduce novel conceptual linkages (# new links). Yet the more students are surrounded by peers of a similar gender in their discipline, the more their novel conceptual linkages are taken up by others ($P < 0.01$): That is, the less a student's gender is represented, the less their novel contributions are adopted by others (uptake per new link). Findings for binary gender and race indicators follow similar patterns. Women and nonwhite scholars introduce more novelty (both $P < 0.001$) but have less impactful novelty (both $P < 0.05$) when compared to men and white students. Additionally, intersectional analyses of gender–race combinations suggest that nonwhite women, white women, and nonwhite men all have higher rates of novelty compared to white men (all $P < 0.001$) but that white men have higher levels of impactful novelty compared to the other groups (all $P < 0.01$). Combined, these findings suggest that demographic diversity breeds novelty and, especially, historically underrepresented groups in science introduce novel recombinations, but their rate of adoption by others is lower, suggesting their novel contributions are discounted.

So why is the novelty introduced by (historically) underrepresented groups less impactful? We test the common hypothesis that innovations that draw together concepts from very different fields or using distal metaphorical links receive less reward. If (historically) underrepresented groups combine distal concepts, this may partly explain their less impactful novelty. We first identify how semantically distal or proximal newly linked concepts are from one another in the space of accumulated concepts using word embedding techniques (35) (see Fig. 3, detailed in *Materials and Methods*). Word embedding techniques enable us to estimate the semantic location of concepts in a vast network of interrelated concepts and compare how distally (or proximally) positioned newly linked concepts are to one another in that space using cosine distance. For the set of newly linked concepts in each thesis, we average their semantic distance and model whether some groups introduce more distal forms of novelty in their theses than other groups. We find that students whose gender is underrepresented in a discipline introduce slightly more concept linkages that are semantically distant (see Fig. 3*C*; $P < 0.001$) and women introduce more distal novelty in comparison to men ($P < 0.001$). In turn, distal novelty relates inversely to impactful novelty; more distal new links between concepts receive far less uptake (see Fig. 3*D*; $P < 0.001$). Hence, underrepresented groups introduce novelty, and the discounting

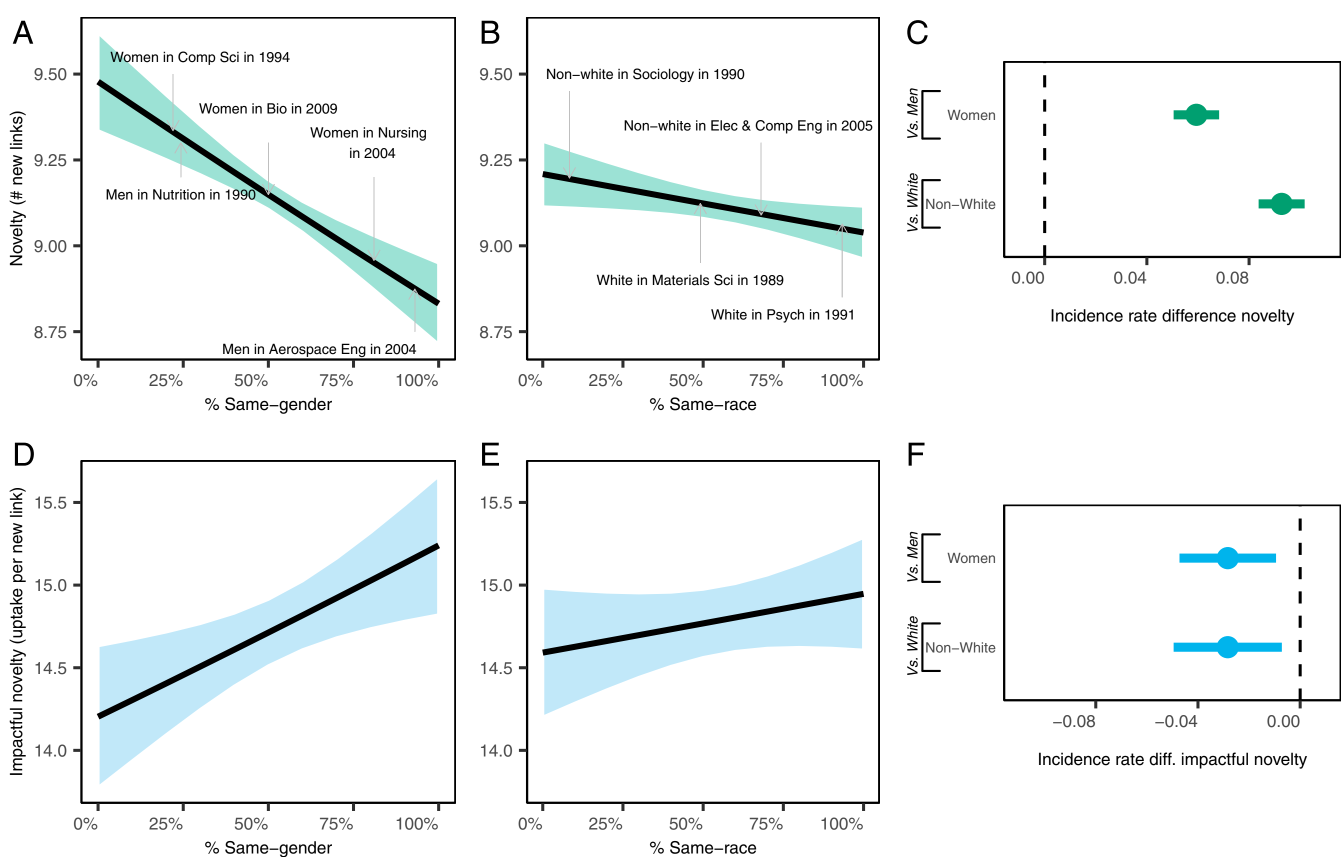

Fig. 2. Gender and race representation relate to novelty and impactful novelty. (*A*) Introduction of novelty (# new links) by the percentage of peers with a similar gender in a discipline (n = 808,375). Specifically, the results suggest that the more students' own gender is underrepresented, the more novelty they introduce. (*B*) Similarly, the more students' own race is underrepresented, the more novelty they introduce. (*C*) Binary gender and race indicators suggest that historically underrepresented groups in science (women, nonwhite scholars) introduce more novelty (i.e., their incidence rate is higher). (*D*) In contrast, impactful novelty decreases as students have fewer peers of a similar gender and suggests underrepresented genders have their novel contributions discounted (n = 345,257). (*E*) There is no clear relation between racial representation in a discipline and impactful novelty. (*F*) Yet the novel contributions of women and nonwhite scholars are taken up less by others than those of men and white students (their incidence rate is lower).

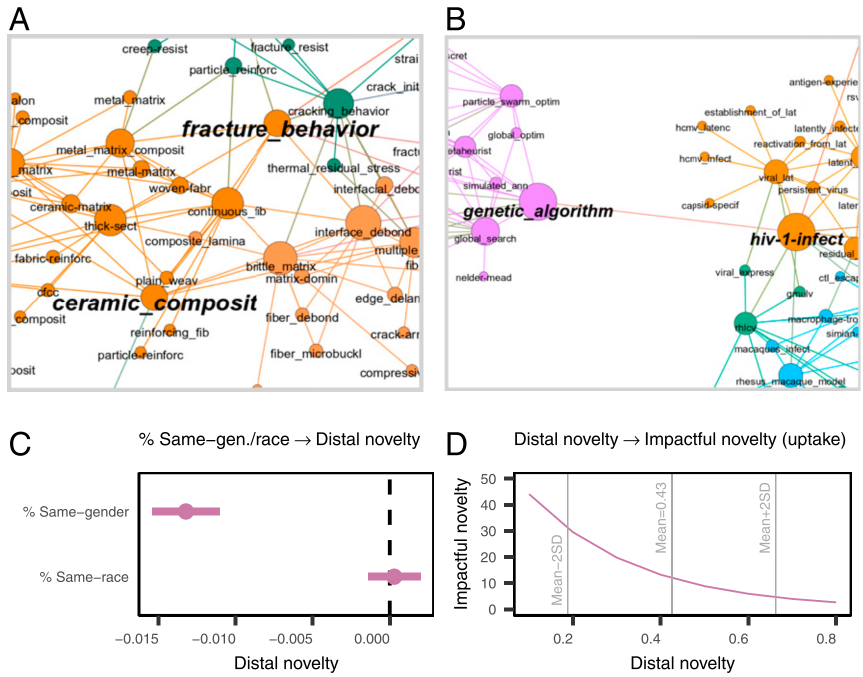

Fig. 3. Underrepresented genders introduce distal novelty, and distal novelty has less impact. (*A* and *B*) Apparent network communities (colors) of concepts and their linkages. (*A*) The link between "fracture_behavior" and "ceramic_composition" arises within a semantic cluster. Both concepts are proximal in the embedding space of scientific concepts, and as such, their distal novelty score is low. (*B*) In contrast, the conceptual link between "genetic_algorithm" and "hiv-1" spans distinct clusters in the semantic network. As such, the concepts are distal in the embedding space of scientific concepts, and their distal novelty score is high. (*C*) Students of an overrepresented gender introduce more proximal novelty, and students from an underrepresented gender introduce more distal novelty in their theses. (*D*) In turn, the average distance of new links introduced in a thesis is negatively related to their future uptake.

of their novel contributions may be partly explained by how distal the conceptual linkages are that they introduce.

Finally, we examine how levels of novelty and impactful novelty relate to extended faculty and research careers. We model careers as (a) obtaining a research faculty position and (b) as continuing research endeavors (Fig. 4 and *SI Appendix*, Table S2). The former reflects PhDs who go on to become primary faculty advisors of PhDs at US research universities, while the latter reflects the broader pool of PhDs who continue to conduct research even if they do not have research advisor roles (e.g., in industry, nontenure line role, etc.). For the latter, we identify which students become publishing authors in the Web of Science (36) 5 y after obtaining their PhD. The conceptual novelty and impactful novelty of a student's thesis is positively related to their likelihood of becoming both a research faculty member or continued researcher (all $P < 0.001$). This suggests that students are more likely to become professors and researchers if they introduce novelty or have impactful novelty.

However, consistent with prior work (8–13), we find that gender and racial inequality in scientific careers persists even if we keep novelty and impactful novelty constant (as well as year, institution, and discipline). Numerically underrepresented genders in a discipline have lower odds of becoming research faculty (~5% lower odds) and sustaining research careers (6% lower odds) compared to gender majorities (all $P < 0.001$). Similarly, numerically underrepresented races in a discipline have lower odds of becoming research faculty (25% lower odds) and continuing research endeavors (10% lower odds) compared to majorities (all $P < 0.001$). Most surprisingly, the positive correlation of novelty and impactful novelty on career recognition varies by gender and racial groups and suggests underrepresented groups' innovations are discounted. The long-term career returns for novelty and impactful novelty are often lower for underrepresented rather than overrepresented groups. At a low level of (impactful) novelty gender minorities and majorities have approximately similar probabilities of faculty careers. But with increasing (impactful) novelty the probabilities diverge at the expense of gender minorities' chances (both slope differences $P < 0.01$). For instance, a 2SD increase from the median of (impactful) novelty increases the relative difference in probability of becoming a faculty researcher for gender minorities and majorities from about 3.5% (4.3%) to 9.5% (15%). These results hold over and above of the distance between newly linked concepts. This innovation discount also holds for traditionally underrepresented groups (i.e., women versus men, nonwhite versus white scholars).

Discussion

In this paper, we identified the diversity–innovation paradox in science. Consistent with intuitions that diversity breeds innovation, we find higher rates of novelty across several notions of demographic diversity (2–7). However, novel conceptual linkages are not uniformly adopted by others. Their adoption depends on which group introduces the novelty. For example, underrepresented genders have their novel conceptual linkages discounted and receive less uptake than the novel linkages presented by the dominant gender. Traditionally underrepresented groups in particular—women and nonwhite scholars—find their novel contributions receive less uptake. For gender minorities, this is partly explained by how "distal" the novel conceptual linkages are that they introduce. Entering science from a new vantage may generate distal novel connections that are difficult to integrate into localized conversations within prevailing fields. Moreover, this discounting extends to minority scientific careers. While novelty and impactful novelty both correspond with successful scientific careers, they offer lesser returns to the careers of gender and racial minorities than their majority counterparts (8–13). Specifically, at

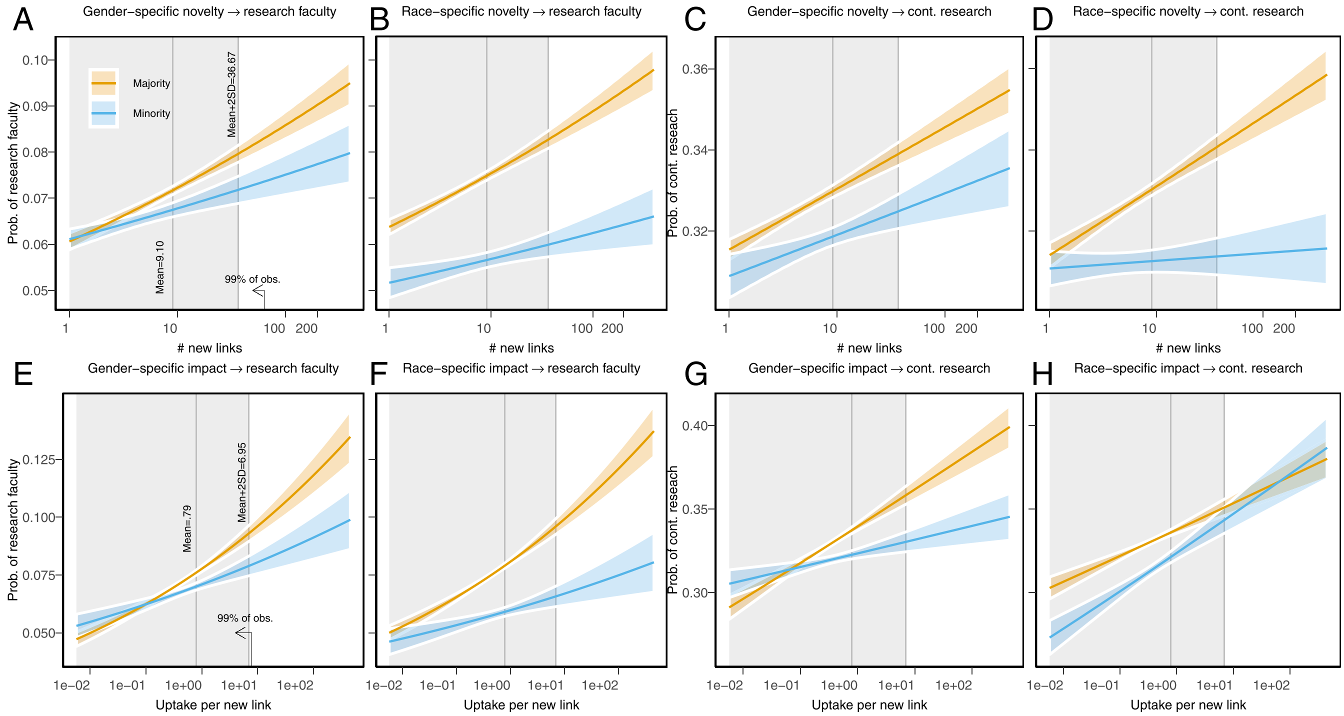

Fig. 4. The novelty and impactful novelty minorities introduce have discounted returns for their careers. (*A*–*H*) Each of the observed patterns holds with and without controlling for distal novelty. (*A*–*D*) Correlation of gender- and race-specific novelty with becoming research faculty or continued researcher (n = 805,236). As novelty increases, the probabilities of becoming faculty (for gender and race) and continuing research (for race) have diminished returns for minorities. For instance, a 2SD increase from the median level of novelty (# new links) increases the relative difference in probability to become research faculty between gender minorities and majorities from 3.5 to 9.5%. (*E*–*H*) Correlation of gender- and race-specific impactful novelty with becoming research faculty and a continued researcher (when novelty is nonzero, n = 628,738). With increasing impactful novelty, the probabilities of becoming faculty (for gender and race) and continuing research (for gender) start to diverge at the expense of the career chances of minorities. For instance, a 2SD increase from the median of impactful novelty (uptake per new link) increases the relative difference in probability of becoming research faculty between gender minorities and majorities from 4.3 to 15%.

low impactful novelty we find that minorities and majorities are often rewarded similarly, but even highly impactful novelty is increasingly discounted in careers for minorities compared to majorities. And this discounting holds over and above how distal minorities' novel contributions are.

In sum, this article provides a system-level account of innovation and how it differentially affects the scientific careers of demographic groups. This account is given for all academic fields from 1982 to 2010 by following over a million US students' careers and their earliest intellectual footprints. We reveal a stratified system where underrepresented groups have to innovate at higher levels to have similar levels of career likelihoods. These results suggest that the scientific careers of underrepresented groups end prematurely despite their crucial role in generating novel conceptual discoveries and innovation. Which trailblazers has science missed out on as a consequence? This question stresses the continued importance of critically evaluating and addressing biases in faculty hiring, research evaluation, and publication practices.

Materials and Methods

Data. This study focuses on a dataset of ProQuest dissertations filed by US doctorate-awarding universities from 1977 to 2015 (20). The dataset contains 1,208,246 dissertations and accompanying dissertation metadata such as the name of the doctoral candidate, year awarded, university, thesis abstract, primary advisor (37.6% of distinct advisors mentor one student), etc. These data cover ∼86% of all awarded doctorates in the US over three decades across all disciplines. We describe below how we follow PhD recipients going on into subsequent academic and research careers.

Concept Extraction from Scientific Text. How do we extract concepts from text? Not all terms are scientifically meaningful; combining function words like "thus," "therefore," and "then" is substantively different from combining terms from the vocabulary of a specific research topic, like "HIV" and "monkey." We argue that innovation entails combining relevant terms from topical lexicons. Hence, we set out to define the latent themes in our corpus of dissertations and the most meaningful concepts in every theme. We employ structural topic models (STMs) (22), commonly used to detect latent thematic dimensions in large corpora of texts (*SI Appendix*).

We fit topic models at K = [50–1000] (K is commonly used to specify the number of topics). Fit metrics (*SI Appendix*, Fig. S1) plateau at K = 400, 500, and 600, and we use those three in this paper. To extract concepts, we identify the terms of relative importance to each latent theme in the dissertation corpus. Using the STM output, we obtain terms that are most frequent and most exclusive within a topic. This helps identify concepts that are both common and distinctive to balance generality and exclusivity. To get at this, we extract the top terms based on their FRequency-EXclusivity (FREX) score (24). FREX scores compound the weighted frequency and exclusivity of a term in a topic. Here we explore three weighing schemes: equally balancing frequency and exclusivity (50/50), attaching more weight to frequency and less to exclusivity (75/25), and attaching more weight to exclusivity and less to frequency (25/75). As such, we analyze nine hyperparameter scenarios (three K and three FREX scenarios) for which sensitivity analyses provide robust results (*SI Appendix*, Table S2). For the results depicted in the main text, we report the scenario where frequency and exclusivity are equally balanced at K = 500.

We use all doctoral abstracts (1977 to 2015) as input documents for a semantic signal for the students' scholarship at the onset of their careers. However, in our inferential analyses, we utilize theses from 1982 to 2010 1) to allow for the scientific concept space to accumulate 5 y before we measure which students start to introduce links and 2) to allow for the most recently graduated students (up until 2010) to have opportunities (5 y) for their novelty to be taken up. Additionally, *SI Appendix*, Fig. S2 suggests that the "stable" years for link introductions and uptake per new link start at ∼1982. The year fixed effects in our inferential analyses (detailed below) further account for left and right censoring: That is, year fixed effects enable comparisons of students within rather than across years. *SI Appendix*, Fig. S3 depicts four exemplary topics and their concepts resulting from the structural topic models.

Outcome Variable: Novelty and Impactful Novelty. Using the extracted scientific concepts, we aggregate co-occurring concepts in abstracts for each year, identifying which students first introduce each novel link. We remove spurious links (due to chance, combinations of extremely rare terms, etc.) by computing a significance score for each link: the log-odds ratio of the probability of link occurrence (computed over all extracted concepts and all documents in the corpus) to the probability of each component concept term occurring independently over the corpus (37, detailed in the *SI Appendix*). In sum, we identify "meaningful" links by filtering the documents for the top FREX terms via structural topic models and then filtering for spurious links through a link significance score. If a link is introduced by two students in the same year, they both get counted. (The percentage of links concurrently introduced is only 1.6%, and the majority of concurrent link introductions arise from students getting their doctorate in the same year [99.7%].) This metric—the number of new link introductions—we call the novelty of a student's thesis (# new links, mean = 9.026; median = 4; SD = 13.744; 20.9% of students do not introduce new links).

Second, we measure impactful novelty, the uptake of a thesis's new links in ensuing theses. We count the total number of times theses in following years use the links first introduced by a prior thesis, normalized by the number of new links. We use the resulting metric, uptake per new link (mean = 0.790; median = 0.333; SD = 3.079), to quantify the average scientific impact of an individual student's thesis. See *SI Appendix*, Fig. S2 for the distributions and correlations of these outcome variables across the different K and FREX scenarios. Both metrics positively correlate with publication productivity and citation among those students that publish (*SI Appendix*, Table S3).

Outcome Variable: Distal Novelty. Some links are "distal" in that they link concepts that are located in distinct clusters of co-occurring concepts. Other links are "proximal" because they link concepts in the same semantic cluster or proximate location. For instance, genetic_algorithm–hiv-1 is distal because it links concepts from distinct research areas: "genetic algorithms" (evolutionary computing) with "hiv-1" (medicine). In contrast, fracture_behavior–ceramic_composition is proximal because the concepts are from the same field.

To operationalize this notion of semantic distance, we embed each concept in a semantic network of cumulated co-occurring concepts and then estimate its location in a vector space, representing each concept c by a fixed dimension vector (or "embedding") $v(c)$. We use the skip-gram model (35), a standard approach that models co-occurrences between concepts by their usage in text (window size is five) and learns a vector for each concept such that concepts with similar co-occurrence patterns have similar embeddings. The result is a space in which concepts with similar embeddings have similar meaning and concepts with dissimilar embeddings have different meanings.

We learn embeddings (of the FREX concepts in the dissertation abstracts) of 100 dimensions, but the metric is robust to 100, 200, or 300 dimensions as well as to stochasticity. We capture the dominant meaning of a concept globally over time. (Although concepts may evolve over time, we use the globally dominant meaning of the concept because we also model uptakes of links globally, and modeling concept embeddings over time is computationally intensive and suffers from data sparsity. Sensitivity analyses for one year [2000] provided very high correlations [$r = 0.931$] between global and time-dependent distal novelty scores.)

Having learned concept embeddings, we calculate how distant newly linked FREX concepts' embeddings are to one another using cosine distance (35) (*SI Appendix*, Table S4). We then average those scores for all novel links introduced in each thesis (distal novelty, mean = 0.426; median = 0.419; SD = 0.118). We validate these automatic measures of concept distance with expert human coders, finding moderate intercoder agreement between distal/proximal assignments to a random set of 100 links and three coders (average Cohen's kappa = 0.46), and together, coder assignments predict ∼95% of the true distal links (i.e., distance score > 0.8). This validation further suggests that distal links are often between concepts from different fields or creative metaphors, and only a fraction of links between distal concepts are hard to interpret substantively (15 to 20%).

Outcome Variable: Careers. To measure innovation reception, we study how innovations relate to two science career outcomes. The first is a conservative proxy of whether graduate students become research faculty after their graduation (research faculty, mean = 0.066). This outcome is measured as graduating PhDs who go on to become a primary advisor of other PhD students in the dissertation corpus. Ultimately, this captures who transitions from student to mentor at a PhD-granting US university and who was able to secure a faculty job with a lineage of students. For those that graduated up until 2010 (i.e., the last graduating cohort we follow), we do consider whether they transitioned to faculty between 2010 and 2015. The second outcome is a more liberal proxy of career success that reflects whether graduating PhDs continue their career in research or not. To capture this, we match students to article authors in the Web of Science (WoS) (*SI Appendix*). The WoS database consists of ∼38 million publication records and their associated metainformation from 1900 to 2017 (disambiguated authors, title, abstracts, etc.). The linkage across datasets allows us to follow students' ensuing careers and research output. Using the ProQuest–WoS link, we measure whether students publish academically at least once in the 5-year period after obtaining their PhD or if they become research faculty, which we interpret as scholars who continue research endeavors (continued research: mean = 0.319). This metric captures a broader range of those who continue to pursue research: scholars who continue to pursue science at institutions that might not grant PhDs (e.g., liberal arts colleges, think-tanks, industry jobs, etc.) or move internationally. Individuals from underrepresented groups might disproportionally move toward such institutions rather than US PhD-granting universities. Hence, examining both metrics indicates whether our results are robust to different academic strata.

Main Covariates. The ProQuest dissertation data do not contain direct reports of student gender and race characteristics, but we identify the degree to which their name corresponds to the race or gender reported by persons with particular first (gender) and last (race) names. We compiled datasets from the US censuses (38) to predict race and from the US Social Security Administration (39) to predict gender. We matched these to data on $n = 20,264$ private university scholars between 1993 and 2015. The private university data contain race and gender information alongside scholar names, which allows us to train a threshold algorithm to estimate race and gender based on names. Using these thresholds, we classify advisees in the ProQuest dissertation data into one of three race categories and to assign a gender (40). The race categories are white, Asian, and underrepresented minorities. Underrepresented minorities combines Hispanics, African Americans, Native Americans, and any racial categories not captured by the first three (*SI Appendix*). To further improve recall on genders and races, we focus on uncategorized genders and races and label them based on additional methods for gender (see refs. 41–43) and race (with full names, refs. 44, 45), thus combining the strength of several methods to help increase coverage and precision for gender and race labels.

We then measure the fraction of students in a discipline-year carrying the same gender or race, e.g., the percentage of women in education in 1987 when a student is a woman, the percentage of underrepresented minority scholars when a student is an underrepresented minority, and so forth (% Same-gender, mean = 0.576; SD = 0.180; % Same-race, mean = 0.625; SD = 0.258). We also measure whether a student is part of an underrepresented gender or race in an academic discipline, i.e., whether a student is member of a group smaller than the largest group in a discipline-year (Gender minority mean = 0.336; Racial minority mean = 0.246, see *SI Appendix*, Fig. S4).

To model novelty, impactful novelty, and distal novelty, we use the percentage of same gender or race and whether scholars are white or nonwhite to find to what extent innovation relates to different notions of group representation in science. We then model careers through minority status in disciplines (results are similar for binary gender and race indicators).

Note that the results here do not take into account cases of gender and race that were not classified according to these methods, although the gender and race distinctions such as those shown in Fig. 2 *C* and *F* do not qualitatively change if we do include "unknown" genders and races in the analyses. Our main substantive conclusions and inferences are robust if we only consider those students whose names overwhelmingly occur within one rather than multiple races. Additionally, finer-grained notions of race or even degrees of identity association with gender or race may be desirable as an indicator. However, underrepresented races appear often in small proportions, which provide little statistical power despite likely sharing a common pattern of associations. As such, we render them into coarser indicators of "underrepresented racial minority." We recognize that in reality, individuals and names have varying degrees of gender and racial associations; as such our named-based metric is a simplified signal of gender and racial identity that may better capture how an individual is perceived by others and can be only a coarse proxy for authors' self-identification with certain genders or races.

Confounding Factors. When dissertation metadata did not include a department, we identified academic discipline for theses filed with ProQuest through a random forest classifier (RFC) based on a list of features from the dissertation with 96% precision ($N_{DISCIPLINE} = 84$; see *SI Appendix*). Dissertations that are filed to ProQuest contain metainformation about the institution where the doctorate was awarded. We classify the student into the first institution that students filed to ProQuest ($N_{UNIVERSITY} = 215$). We infer the graduation year in which students obtain their doctorate as the year in which the dissertation was filed to ProQuest (range = 1977 to 2015).

Analytical Strategy. We model each of our dependent variables tailored to their statistical distributions. Scientific novelty (# new links) and impactful novelty (uptake per link), are right-skewed counts of events or rates. For these outcomes, we employ negative binomial regression analyses, where the overdispersion in the outcomes is modeled as a linear combination of the covariates (46). Distal novelty is relatively normally distributed, and we model it through linear regression. Academic careers such as becoming research faculty (yes/no) and sustaining a research career (yes/no) are both binary outcomes, so we use logistic regression analyses for these (*SI Appendix*). The whiskers and shaded lines in Figs. 2–4 represent upper and lower bounds of 95% CIs, and the *P* values we report here are two-sided tests. Figs. 2 *A*, *B*, *D*, and *E*, 3*D*, and 4 all represent average marginal effects considering all other values of the other independent variables. Fig. 2 *C* and *F* reports the incidence rate differences between groups from the negative binomial regressions.

Apart from the main covariates, we include three sets of fixed effects in our models to better isolate our main predictors from confounding factors. We keep institution, academic discipline, and graduation year constant throughout. These fixed effects account for university differences in prestige and the resources they make available to students (33), for the differences across academic fields and disciplinary cultures (34), and for "older" scholars who have had more time to make career transitions or to get recognized.

We weigh the data by the total number of doctorates awarded by an institution in a given year (*SI Appendix*) to account for possible selectivity between universities in years when filing their doctorates' theses in the ProQuest database and to render our results generalizable to the US scholarly population. These survey weights are based on the relative number of PhD recipients in the ProQuest data vis-à-vis the US PhD population per year for each university.

Finally, novelty (# new links) is modeled for students with nonmissing values on all features (n = 808,375), impactful novelty (uptake per new link) is modeled for those with nonzero novelty and nonzero uptake given its best fit with the negative binomial model (n = 345,257), and distal novelty is modeled for the students with nonzero novelty (n = 630,971). Careers are modeled for those for whom there are no constant successes or failures within the fixed effects and for those who introduce at least one link (n = 805,236) or whose novelty is nonzero for impactful novelty (n = 628,738).

Data and Materials Availability. The data used in this study were obtained according to protocol 12996, approved by Stanford University. We acquired written permission from ProQuest to scrape and analyze their US dissertation data for scientific purposes. The full dissertation corpus can be requested via ProQuest (20), and the Web of Science can be requested via Clarivate Analytics (36). Code to replicate our key metrics is found on GitHub (https://github.com/bhofstra/diversity_innovation_paradox). Top terms from the K = 500 structural topic model that equally balances frequency and exclusivity are also found there.

ACKNOWLEDGMENTS. We acknowledge Stanford University and the Stanford Research Computing Center for providing computational resources and support that contributed to these research results. This paper benefited from discussions with Lanu Kim, Raphael Heiberger, Kyle Mahowald, Mathias W. Nielsen, Anthony Lising Antonio, James Zou, and Londa Schiebinger. This paper was supported by three NSF grants [NSF Grant 1633036, NSF Grant 1827477, and NSF Grant 1829240] and by a grant from the Dutch Organization for Scientific Research [NWO Grant 019.181SG.005].

1. R. K. Merton, *The Sociology of Science* (University of Chicago Press, Chicago, IL, 1973).
2. M. S. Granovetter, The strength of weak ties. *Am. J. Sociol.* **78**, 1360–1380 (1973).
3. R. S. Burt, Structural holes and good ideas. *Am. J. Sociol.* **110**, 349–399 (2004).
4. M. W. Nielsen *et al*., Opinion: Gender diversity leads to better science. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 1740–1742 (2017).
5. S. E. Page, *The Diversity Bonus: How Great Teams Payoff in the Knowledge Economy* (Princeton University Press, Princeton, NJ, 2009).
6. S. T. Bell, A. J. Villado, M. A. Lukasik, L. Belau, A. L. Briggs, Getting specific about demographic diversity variable and team performance relationships: A meta-analysis. *J. Manage.* **37**, 709–743 (2011).
7. M. de Vaan, D. Stark, B. Vedres, Game changer: The topology of creativity. *Am. J. Sociol.* **120**, 1144–1194 (2015).
8. C. A. Moss-Racusin, J. F. Dovidio, V. L. Brescoll, M. J. Graham, J. Handelsman, Science faculty's subtle gender biases favor male students. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 16474–16479 (2012).
9. W. W. Ding, F. Murray, T. E. Stuart, Gender differences in patenting in the academic life sciences. *Science* **313**, 665–667 (2006).
10. J. D. West, J. Jacquet, M. M. King, S. J. Correll, C. T. Bergstrom, The role of gender in scholarly authorship. *PLoS One* **8**, e66212 (2013).
11. L. A. Rivera, When two bodies are (not) a problem: Gender and relationship status discrimination in academic hiring. *Am. Sociol. Rev.* **82**, 1111–1138 (2017).
12. M. H. K. Bendels, R. Müller, D. Brueggmann, D. A. Groneberg, Gender disparities in high-quality research revealed by Nature Index journals. *PLoS One* **13**, e0189136 (2018).
13. A. Clauset, S. Arbesman, D. B. Larremore, Systematic inequality and hierarchy in faculty hiring networks. *Sci. Adv.* **1**, e1400005 (2015).
14. B. Uzzi, S. Mukherjee, M. Stringer, B. Jones, Atypical combinations and scientific impact. *Science* **342**, 468–472 (2013).
15. J. G. Foster, A. Rzhetsky, J. A. Evans, Tradition and innovation in scientists' research strategies. *Am. Sociol. Rev.* **80**, 875–908 (2015).
16. D. Wang, C. Song, A. L. Barabási, Quantifying long-term scientific impact. *Science* **342**, 127–132 (2013).
17. A. T. J. Barron, J. Huang, R. L. Spang, S. DeDeo, Individuals, institutions, and innovation in the debates of the French Revolution. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 4607–4612 (2018).
18. I. Iacopini, S. Milojević, V. Latora, Network dynamics of innovation processes. *Phys. Rev. Lett.* **120**, 048301 (2018).
19. D. Trapido, How novelty in knowledge earns recognition: The role of consistent identities. *Res. Policy* **44**, 1488–1500 (2015).
20. ProQuest, ProQuest Dissertations & Theses Global. https://www.proquest.com/products-services/pqdtglobal.html. Accessed 14 January 2020.
21. S. Fortunato *et al*., Science of science. *Science* **359**, eaao0185 (2018).
22. M. E. Roberts *et al*., Structural topic models for open-ended survey responses. *Am. J. Pol. Sci.* **58**, 1064–1082 (2014).
23. A. El-Kishky *et al*., Scalable topical phrase mining from text corpora. *Proc. VLDB Endowment* **8**, 305–316 (2014).
24. J. M. Bischof, E. M. Airoldi, "Summarizing topical content with word frequency and exclusivity" in *Proceedings of the 29th International Conference on Machine Learning*, J. Langford, J. Pineau, Eds. (Omnipress, Madison, 2012), pp. 9–16.
25. M. G. Sarngadharan, M. Popovic, L. Bruch, J. Schüpbach, R. C. Gallo, Antibodies reactive with human T-lymphotropic retroviruses (HTLV-III) in the serum of patients with AIDS. *Science* **224**, 506–508 (1984).
26. L. Schiebinger, *The Mind Has No Sex? Women in the Origins of Modern Science* (Harvard University Press, Cambridge, MA, 1991).
27. D. Strickland, G. Mourou, Compression of amplified chirped optical pulses. *Opt. Commun.* **6**, 447–449 (1985).
28. T. S. Kuhn, *The Structure of Scientific Revolutions* (University of Chicago Press, Chicago, Chicago, IL, 1962).
29. M. L. Weitzman, Recombinant growth. *Q. J. Econ.* **113**, 331–360 (1998).
30. D. Stark, *The Sense of Dissonance: Accounts of Worth in Economic Life* (Princeton University Press, Princeton, NJ, 2011).
31. D. Jurgens *et al*., Measuring the evolution of a scientific field through citation frames. *Trans. Assoc. Comput. Linguist.* **6**, 391–406 (2018).
32. H. Small, On the shoulders of Robert Merton: Towards a normative theory of citation. *Scientometrics* **60**, 71–79 (2004).
33. V. Burris, The academic caste system: Prestige hierarchies in PhD exchange networks. *Am. Sociol. Rev.* **69**, 239–264 (2004).
34. A. Abbott, *The System of Professions: An Essay on the Division of Expert Labor* (University of Chicago Press, Chicago, IL, 1998).
35. T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space. arXiv:1301.3782 (7 September 2013).
36. Clarivate Analytics,Web of Science Raw Data (XML). clarivate.libguides.com/rawdata. Accessed 14 January 2020.
37. C. Ramisch, P. Schreiner, M. Idiart, A. Villavicencio, "An evaluation of methods for the extraction of multiword expressions" in *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions 2008* (MWE 2008, Marrakech, 2008), pp. 50–53.

38. US Census Bureau, Frequently occurring surnames from the census 2000. https://www.census.gov/topics/population/genealogy/data/2000_surnames.html. Accessed 27 September 2017.
39. Social Security Administration, Popular baby names. https://ssa.gov/oact/babynames/limits.html. Accessed 27 September 2017.
40. B. Hofstra, R. Corten, F. Van Tubergen, N. C. Ellison, Sources of segregation in social networks: A novel approach using Facebook. *Am. Sociol. Rev.* **82**, 625–656 (2017).
41. C. W. Fox, C. S. Burns, A. D. Muncy, J. A. Meyer, Gender differences in patterns of authorship do not affect peer review outcomes at an ecology journal. *Funct. Ecol.* **30**, 126–139 (2015).
42. C. M. Topaz, S. Sen, Gender representation on journal editorial boards in the mathematical sciences. *PLoS One* **11**, e0161357 (2016).
43. L. Holman, D. Stuart-Fox, C. E. Hauser, The gender gap in science: How long until women are equally represented? *PLoS Biol.* **16**, e2004956 (2018).
44. G. Sood, S. Laohaprapanon, Predicting race and ethnicity from the sequence of characters in a name. arXiv:1805.02109 (5 May 2018).
45. G. Sood, Florida voter registration data. Harvard Dataverse. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/UBIG3F. Accessed 23 January 2020.
46. C. A. Cameron, P. K. Trivedi, *Microeconomics Using Stata* (Stata Press, College Station, TX, 2010), pp. 553–598.