The incidence and role of negative citations in science

Christian Catalini[a], Nicola Lacetera[b], and Alexander Oettl[c,1]

[a]MIT Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02142; [b]Institute for Management and Innovation, University of Toronto, Mississauga, ON, Canada L5L 1C6; and [c]Scheller College of Business, Georgia Institute of Technology, Atlanta, GA 30308

Edited by Dean Keith Simonton, University of California, Davis, and accepted by the Editorial Board September 4, 2015 (received for review February 4, 2015)

Citations to previous literature are extensively used to measure the quality and diffusion of knowledge. However, we know little about the different ways in which a study can be cited; in particular, are papers cited to point out their merits or their flaws? We elaborated a methodology to characterize "negative" citations using bibliometric data and natural language processing. We found that negative citations concerned higher-quality papers, were focused on a study's findings rather than theories or methods, and originated from scholars who were closer to the authors of the focal paper in terms of discipline and social distance, but not geographically. Receiving a negative citation was also associated with a slightly faster decline in citations to the paper in the long run.

social studies of science | citation analysis | bibliometric techniques | natural-language processing | negative citations

Scientific knowledge is a key input for economic prosperity (1–3) and evolves thanks to the complementary contributions of different scientists. The norms that regulate the scientific community coordinate this endeavor (4–6).

Citation of previous work is one such norm and a major means of documenting the collective and cumulative nature of knowledge production. Citations allow for the establishment of credit and the identification of scientific paradigms and their shifts (7, 8); they measure the impact and quality of discoveries and, by extension, of a researcher, an institution, or a journal (9, 10). Studies rely on citation data also to analyze the diffusion of scientific ideas, the creation and evolution of scientific networks, and the role of top scientists and inventions (11–17).

Less attention has been devoted to the different intentions behind a citation. In particular, although papers may often be cited because a current study is consistent with past work or builds upon it, a reference can sometimes be made to point out limitations, inconsistencies, or flaws that are even more serious. These "negative" citations may question or limit the scope and impact of a contribution, a scholar, or an entire line of research. Criticisms expressed through citations could also be part of the "falsification" process that, according to Karl Popper, characterizes science and could be a signal of the solidity of a field (18). For example, the recent criticisms and eventual dismissal of the evidence of gravitational waves and ultrafast expansion of the universe in the "big bang" were interpreted as developments in the study of the origins of the universe (19). Even for findings that are eventually confirmed, critiques may be beneficial in the process. For instance, the Copernican revolution benefited from and was refined by Tycho Brahe's observations about inconsistencies in the heliocentric view, despite the eventual falsification of Brahe's theory (20).

A thorough classification and understanding of different types of citations, and in particular of negative citations—their incidence, distribution within research fields and across time, their location within a paper, and the connections that they establish between studies and scholars—is therefore a valuable exercise to understand the evolution of science. This enhanced classification may also offer current repositories of scholarly work (such as Google Scholar, PubMed, ISI Web of Science, and Scopus) an opportunity to improve their search and ranking algorithms; by extracting more information from citations, we can uncover more information, reject false knowledge more rapidly, and ultimately enhance the scientific discourse.

Such a classification, however, is difficult to perform and would have been impossible just a few years ago. Recent advancements in natural-language processing (NLP) (21) and in the ability to parse and analyze large bodies of text, however, now allow us to reconstruct the context in which a citation was made, and therefore to understand why a given study was cited in the first place.

We developed a method to identify citations that question the validity of previous results and to analyze their incidence and patterns to determine their role, relevance, and impact using bibliometric data, NLP techniques, and domain experts. In this study we provide evidence of (*i*) how negative citations are expressed, (*ii*) their incidence or frequency, (*iii*) the types of papers that receive these critiques and the types of papers that make the critiques, (*iv*) the parts of a study that are negatively cited (e.g., the theory, the results, the implications, etc.), (*v*) the relationships between the citing and cited authors, and (*vi*) the consequences of a negative citation in terms of future citations. To guarantee homogeneity of the analysis and define a feasible testing ground, the analysis in this paper was based on 15,731 full-text articles in the *Journal of Immunology* (1998–2007) and the 762,355 citations contained in those papers. Details of our procedures are in *Materials and Methods* and *Supporting Information*.

Results

Out of 762,355 citations from 15,731 articles in the *Journal of Immunology* (1998–2007), we identified 18,304 as negative (about 2.4% of the total). The 762,355 citations referred to 146,891 unique papers, and of these papers 10,405 (about 7.1%) received at least one negative citation. Thus, although the incidence on a per-citation basis is relatively low, a nontrivial number of papers received at least one negative citation. On the one hand, the low frequency may be evidence of a limited, uninfluential role of negative citations, or of

Significance

Providing a detailed classification of the types of citations that an article receives is important to establish the quality of a study and to characterize how current research builds upon prior work. The methodology that we propose also informs how to improve the citation process, for example by having scientists attach additional metadata to their citations. The approach is scalable to other fields and periods and can also be used to identify other types of citations (e.g., reuse of methods, materials, empirical tests of theory, and so on). Finally, our methods provide online repositories such as Google Scholar, PubMed, ISI Web of Science, and Scopus with a way to improve their search and ranking algorithms.

Author contributions: C.C., N.L., and A.O. designed research, performed research, contributed new reagents/analytic tools, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. D.K.S. is a guest editor invited by the Editorial Board.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. Email: Alexander.Oettl@scheller.gatech.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1502280112/-/DCSupplemental.

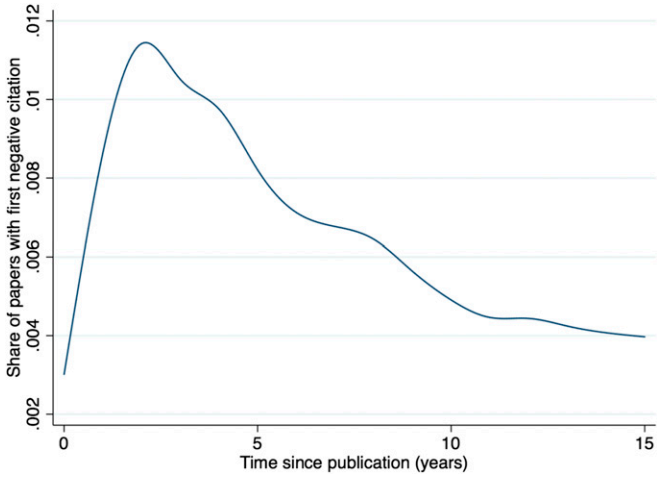

Fig. 1. Share of articles receiving their first negative citation at a given age (years).

the high social cost of making them. On the other hand, these citations represent a nonnegligible share of the total in our data and, as such, could play an important role in limiting and correcting previous results, thus helping science progress (5, 18). Several features of these citations, described below, lend support to the latter hypothesis, that is, that negative citations have unique functions and deserve consideration.

First, Fig. 1 shows that the likelihood of receiving a negative citation was higher in the first few years after a paper was published; this is arguably the period in which the underlying science was potentially more novel, untested, and worthy of more attention and scrutiny.

Second, and consistent with negative citations emerging when enough attention is given to a paper, negatively cited studies were of high quality and prominence; the median number of citations to these papers was higher than to papers that never received negative citations throughout their full citation "life cycle" (Fig. 2). Thus, as scientists pay more attention to a study (potentially also because of its novelty and quality), they are also more likely to provide criticisms, extensions, and qualifications to it. Negative citations can be therefore seen as a way to track where scientists place attention at a certain time within a field.

Third, the 4,888 papers (31% of the citing papers) that made at least one negative citation had a distribution of citations that is statistically indistinguishable from the 10,843 articles not making any negative citation (P value for test of equality of distribution = 0.32; the full analysis is in *Supporting Information*). Negatively citing papers are therefore not just marginal studies, perhaps differentiating themselves through incremental critiques of previous work; they rather appear as "equal" contributors to the overall advancement of a field. Seen in conjunction with the prior observation that negatively cited papers are of higher quality, this result may imply (although more work is needed to test this) that fields where negative citations occur are built on more solid foundations or at least may evolve more quickly because of increased interest by scientists.

Fourth, we took advantage of the relatively standardized structure of scientific articles in immunology to assess whether negative citations disproportionately appeared in certain sections of an article. Fig. 3 shows that about 84% of all negative citations occurred in the "Results and Discussion" section, as opposed to 42% of objective citations (χ^2 test for the equality of distribution of types of citations across sections = 8,700.6, $P < 0.000$). On the one hand, this suggests that negative citations may serve a specific purpose, at least in immunology (i.e., they mostly focus on findings rather than methods or theories, and are therefore different from other citations). On the other hand, it is possible that negative citations to, for example, theories and methods use language and wording that is more subtle, and thus less likely to be identified by our algorithm than negative citations related to results. To address this concern, we compared the keywords used in negative citations across different sections of a paper, and in negative and objective citations within the same section. To calculate how similar two citations were, we used the cosine similarity between the vectors of keywords generated by the relevant paragraphs of text. Perhaps not surprisingly, the analysis revealed that negative citations in a section were more similar to negative citations in other sections than to objective citations in the same section. More interestingly, we found no systematic or sizeable differences in the similarity of negative citations when making pairwise comparisons between sections.

Fifth, negative citations were more likely to come from scientists who were close in discipline and social distance to the cited scholars. Fig. 4 reports the coefficient estimates from regressions of the probability that a paper received a negative citation (conditional on ever being cited) on variables measuring the physical distance between authors on the citing and cited papers, whether the citing and cited papers were both in the field of immunology, and whether authors on citing and cited papers

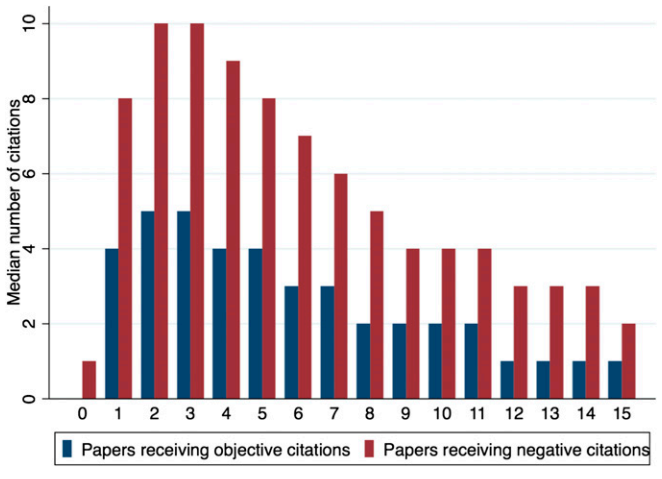

Fig. 2. Median overall citations for papers receiving and not receiving negative citations, by age (years) of the papers.

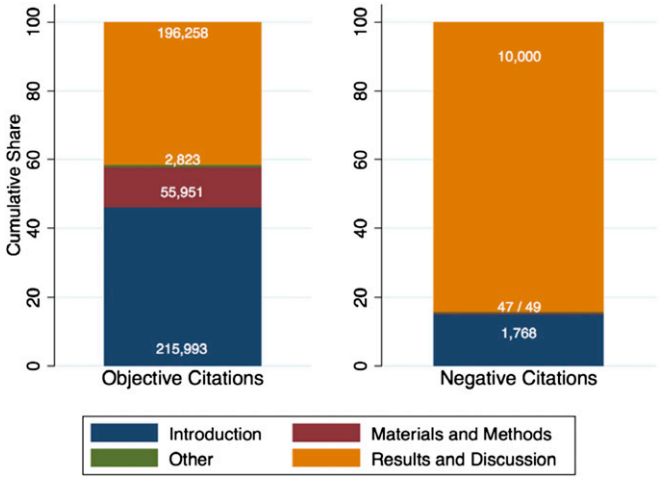

Fig. 3. Distribution of objective and negative citations by sections in an article.

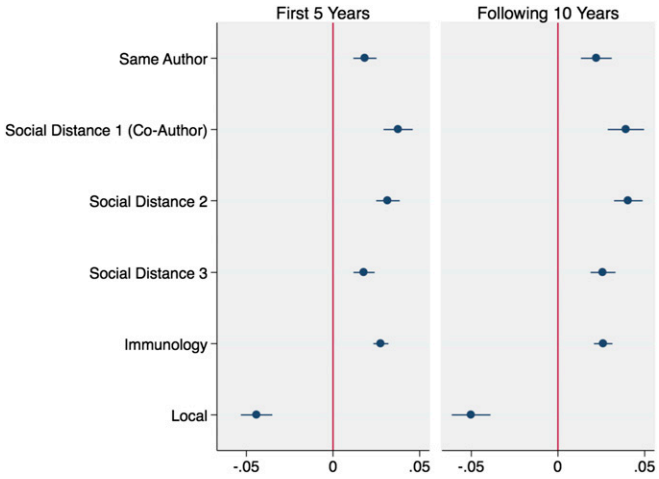

Fig. 4. Estimated changes in the probability of receiving a negative citation as a function of social, discipline, and geographic distance. Estimates are from Logit models where the outcome variable is an indicator for having received a negative citation at a given time, conditional on ever being cited, on the variables indicated on the vertical axis. Same author indicates self-citations. Social distance 1 (2, 3) indicates citation by a coauthor (coauthor of a coauthor, coauthor of the coauthor of a coauthor). Immunology indicates that the citing paper is in an immunology journal; local indicates that the closest author of a citing paper is at less than 150 miles from the closest author of the cited paper.

were connected through previous collaborations. Within the first 5 years after publication, negative citations were more likely to come from articles published in other immunology journals, from authors who were more connected to the authors of the cited paper (e.g., coauthors, and coauthors of coauthors), and from scientists who were affiliated with institutions at a distance greater than 150 miles from the closest author in the focal article. In addition to being better positioned to understand the work of a focal author, a vast literature shows that scientists who are in the same discipline and socially proximate are more likely to interact and exchange information in a plurality of ways (22, 23). This finding is again consistent with the interpretation that awareness and scrutiny are prerequisites for negative citations; moreover, this result hints to the fact that negative citations may be one of the ways in which scientists debate and make progress in their field of research. In contrast, geographic proximity was negatively correlated with the presence of a negative citation. An explanation for this result is that social proximity is a more accurate indicator of closeness in knowledge space than physical proximity (colocated scientists who do not have direct or indirect coauthorship links may well be working in unrelated areas). Another interpretation is that it may be socially costly to negatively cite the work of a local colleague. Thus, for geographically colocated scientists other forms of feedback (e.g., personal interactions) may substitute rather than complement more "formal" negative citations.

Finally, we assessed the impact of receiving a negative citation on the subsequent citation profile of a paper. Previous studies analyzed the effect of a retraction on the future citations to the retracted paper (24–27). We first compared the negatively cited papers to all other papers (Fig. 5). There was a marked relative decline in citations for the negatively cited papers after the first negative citation. However, this analysis may be misleading because the comparison was between potentially very different articles; recall, for example, that negatively cited articles were on average more prominent and might have therefore received more attention. We therefore matched and compared negatively cited articles to other papers with comparable characteristics, such as age and previous citations. Fig. 6 shows a much more similar citation profile over time for these two sets of articles, both before and after a negative citation occurred. The relative decline in overall citations eight or more years after the first negative citation reveals that there is a small, long-term "penalty" for negatively cited papers, although it takes a long time to occur.

Discussion

Our findings suggest that negative citations may indeed play a special role in science. A possible explanation for the lack of an immediate penalty from negative citations on the subsequent interest for an article is that negative citations may contribute to refining initial findings and help a field evolve. Another explanation is that negative citations simply go unnoticed, and that the information they carry may take a long time to diffuse. Before making a citation, scientists would have to read all of the articles that cite the focal article to verify whether any of its content was updated by a follow-up study: although the focal article might diffuse rapidly, the one carrying the improved or correct information has to start a diffusion process of its own. As such, tracking the evolution of negative citations, and more generally of the different reasons why previous literature is referenced, is

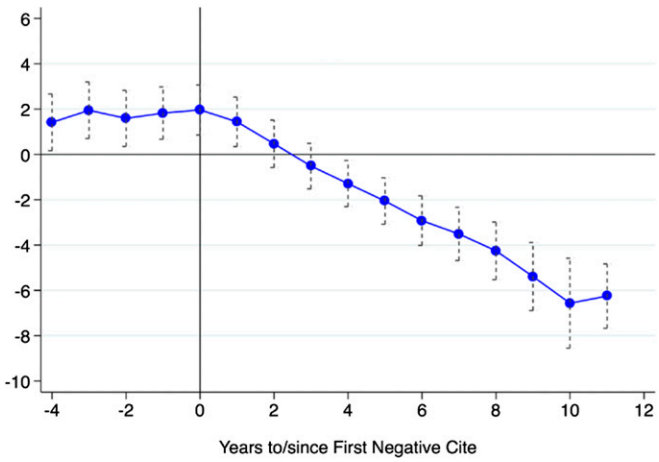

Fig. 5. Citations to the focal article after the first negative citation is made, not controlling for quality. Error bars correspond to 95% confidence intervals.

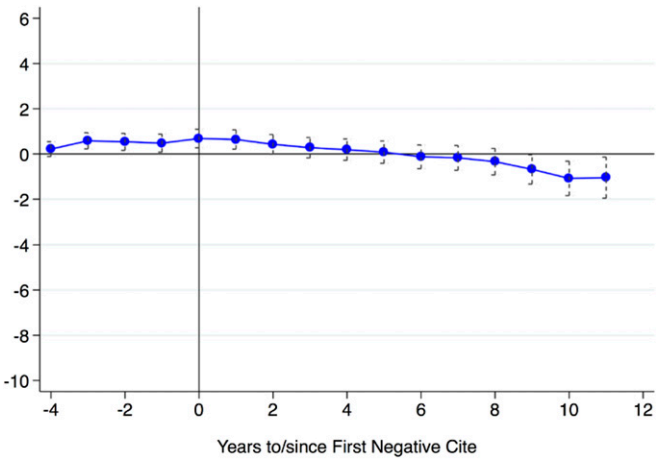

Fig. 6. Citations to the focal article after the first negative citation is made, controlling for quality. Error bars correspond to 95% confidence intervals.

an important exercise. Our approach has the potential to inform how the citation process could be improved, and what kind of metadata scientists should be invited to attach to their citations to facilitate search, discovery, and knowledge recombination.

The findings in this paper are limited by the fact that they concern only the field of immunology. Although a small-scale, manual analysis that we performed on 2,860 citations in the mathematics section of *PLOS ONE* returned a similar rate of negative citations (1.7% of the total), a more comprehensive comparison across disciplines is necessary before results can be generalized. We hope that future research will build on the methodology presented here to cover additional fields and historical periods, and to address some of the conjectures and open questions generated by our results as described above.

Materials and Methods

The analysis was based on 15,731 full-text articles in the *Journal of Immunology* (1998–2007) and the 762,355 citations (to studies published in any journal) contained in those papers, which we extracted, parsed, and linked to the full bibliographic information on the citing articles. The database used for the retrieval of the information on the citing papers was ISI Web of Science. We were able to match 486,600 (64%) of the 762,355 citations to their full bibliographic details. Note that the incidence of negative citations was statistically indistinguishable between all citations (2.40%) and the sample of citations for which we could retrieve the full bibliographic data (2.44%; $P = 0.146$). To identify negative citations, we first developed a training set of 15,000 citations (i.e., sentences in a paper that contained the reference to another paper). A team of immunology PhDs and researchers manually reviewed these citations. These experts classified as negative citations references that pointed to the inability to replicate past results, disagreement, or inconsistencies with past results, theory, and literature. All other references that were simply referring to or building on past work were classified as "objective" citations. The quotes below report examples of negative citations as identified by our experts:

> "The data therefore contrast with reports that Tregs and conventional T cells are equally sensitive to (superantigen-dependent or peptide-dependent) deletion (7, 9)."

> "This conclusion appeared inconsistent with other experiments that indicated that H2-DM mutant animals generate strong CD4 + T cell responses when immunized with synthetic peptides (16)."

> "This finding stands in contrast to the negative results of a previous study that used a similar recombinant Melan-A protein to screen 100 serum samples from melanoma patients (33)."

> "However, our findings differ from those of Yan and colleagues (14), who reported MIP-2 expression (by immunohistochemical analysis) in the corneal epithelium in herpes simplex keratitis."

We used NLP to run the full set of citations and automatically assign them using our training set of labeled data to the two types of citations: objective and negative. We relied on the Python NLTK library to perform the classification; when a citation paragraph was analyzed, the algorithm decomposed it into its grammar components (e.g., adjectives, verbs, etc.). These components were then stored into a presence dictionary and used to build a feature set. The feature set was the basis for a Bayesian model of the data. We used the NLTK Naive Bayes algorithm (www.nltk.org/_modules/nltk/classify/naivebayes.html) and probabilistically assigned the remaining citations to the two types of interest. As the training set grows in size, this approach allows us to accurately process large sets of citations in a relatively short time.

To determine the similarity in the wording used in a negative citation in different sections of an article, we transformed our citation paragraphs into vectors of keywords (after removing symbols and special characters from the text) and calculated the similarity between all pairs of citations coming from the different sections. For example, we calculated the similarity between negative citations coming from the "Results and Discussion" section and the "Materials and Methods" section. We also calculated the similarity between objective and negative citations within the same section (e.g., negative and objective citations in the "Introduction" section). Our measure of similarity comes from the Scikit-Learn Python module (scikit-learn.org/stable/modules/metrics.html), which calculates the L2-normalized dot product of two vectors of keywords (cosine similarity).

For the analyses reported in Fig. 6, finally, we matched negatively cited articles to objectively cited control articles using the coarsened exact matching procedure (28). The control sample that allowed us to compare the effect of being negatively cited consisted of objectively cited papers that were similar to the negatively cited papers in citation profile, cohort, and age. For each of our cited papers, we constructed discrete bins for the year in which the paper was published (cohort), the age of each cited paper when it was first cited by a paper in the *Journal of Immunology* (age), the number of citations received 1 year before the citation in the *Journal of Immunology*, and the cumulative number of total citations received 1 year before the citation in the *Journal of Immunology*. We created individual strata for each bin and selected a random objectively cited paper from each stratum that also contained a negatively cited paper to serve as the control paper. We were able to find suitable control articles for 7,741 of the 10,405 negatively cited articles.

ACKNOWLEDGMENTS. We thank Drs. Melissa Kemp and Jennifer Leavey for their comments on our research and for helping us to better understand the immunology context. Zerzar Bukhari, Edward Kim, and Jack Gao provided excellent research assistance.

1. Romer PM (1990) Endogenous technological change. *J Polit Econ* 98(5):S71–S102.
2. Phelps ES (1966) Models of technical progress and the golden rule of research. *Rev Econ Stud* 33(2):133–145.
3. Arrow KJ (1969) Classificatory notes on the production and transmission of technological knowledge. *Am Econ Rev* 59(2):29–35.
4. Partha D, David PA (1994) Towards a new economics of science. *Res Policy* 23(5): 487–521.
5. Merton RK (1957) Priorities in scientific discoveries: A chapter in the sociology of science. *Am Soc Rev* 22(6):635–659.
6. Stephan PE (2010) The economics of science. *Handbook of the Economics of Innovation*, eds Hall BH, Rosenberg N (North-Holland, Amsterdam), pp 217–273.
7. Merton RK (1968) The Matthew effect in science. *Science* 159(3810):56–63.
8. Kuhn TS (1962) *The Structure of Scientific Revolutions* (Univ of Chicago Press, Chicago).
9. Hall BH, Jaffe A, Trajtenberg M (2005) Market value and patent citations. *RAND J Econ* 36(1):16–38.
10. Trajtenberg M (1990) A penny for your quotes: Patent citations and the value of innovations. *RAND J Econ* 21(1):172–187.
11. Gittelman M, Kogut B (2003) Does good science lead to valuable knowledge? Biotechnology firms and the evolutionary logic of citation patterns. *Manage Sci* 49(4):366–382.
12. Narin F, Hamilton KS, Olivastro D (1997) The increasing linkage between U.S. technology and public science. *Res Policy* 26(3):317–330.
13. Breschi S, Lissoni F (2004) Knowledge networks from patent data: Methodological issues and research targets. *Handbook of Quantitative S&T Research: The Use of Publication and Patent Statistics in Studies of S&T Systems*, eds Glanzel W, Moed H, Schmoch U (Springer, Berlin), pp 613–643.
14. Cowan R, Jonard N (2004) Network structure and the diffusion of knowledge. *J Econ Dyn Control* 28(8):1557–1575.
15. Fleming L, King C, Juda AI (2007) Small worlds and regional innovation. *Organ Sci* 18(6):938–954.
16. Singh J (2005) Collaborative networks as determinants of knowledge diffusion patterns. *Manage Sci* 51(5):756–770.
17. Harhoff D, Narin F, Scherer FM, Vopel K (1999) Citation frequency and the value of patented inventions. *Rev Econ Stat* 81(3):511–515.
18. Popper KR (1959) *The Logic of Scientific Discovery* (Hutchinson, London).
19. Ball P (September 26, 2014) Scientists got it wrong on gravitational waves. So what? *Guardian*. Available at www.theguardian.com/commentisfree/2014/sep/26/scientists-gravitational-waves-science. Accessed September 23, 2015.
20. Sherwood S (2011) Science controversies past and present. *Phys Today* 64(10): 39–44.
21. Cambria E, White B (2014) Jumping NLP curves: A review of natural language processing research. *IEEE Comput Intell Mag* 9(2):48–57.
22. Agrawal A, Goldfarb A (2008) Restructuring research: Communication costs and the democratization of university innovation. *Am Econ Rev* 98(4):1578–1590.
23. Newman ME (2004) Coauthorship networks and patterns of scientific collaboration. *Proc Natl Acad Sci USA* 101(Suppl 1):5200–5205.
24. Azoulay P, Furman JL, Krieger JL, Murray FE (2012) Retractions. NBER Working Paper No. 18499. Available at www.nber.org/papers/w18499.
25. Furman JL, Jensen K, Murray F (2012) Governing knowledge in the scientific community: Exploring the role of retractions in biomedicine. *Res Policy* 41(2):276–290.
26. Lu SF, Jin GZ, Uzzi B, Jones B (2013) The retraction penalty: Evidence from the Web of Science. *Sci Rep* 3:3146.
27. Campanario JM (2000) Fraud: Retracted articles are still being cited. *Nature* 408(6810):288.
28. Iacus SM, King G, Porro G (2012) Causal inference without balance checking: Coarsened exact matching. *Polit Anal* 20(1):1–24.