Toward a better understanding of front-page and full-text patent references to science: A survey proposal

There is limited understanding of what types of linkages between science and technology (if any) either front page or in-text citations represent. Accordingly, we hope to survey inventors on patents to better understand what types of relationships citations to science in patents represent, and whether there are differences between front-page and full-text references, modeled after the pioneering NBER/Case Western Survey of Patentees (Jaffe et al 2000). We will go beyond this, and use results from the survey to train a computational classification algorithm to classify in-text citations into groups that are more or less informative and according to the specific type of relationship they capture (e.g. tools, background knowledge, concepts that could be improved, etc). Finally, we will extend this backward to the millions of in-text references to science in patents granted since 1984; the resulting dataset will allow for a much deeper understanding of science technology linkages than had been previously been possible. The basic approach is described below, in hopes of getting feedback before we launch the full survey.

As a start, we recently sent out a pilot survey modelled on the NBER survey to try to better understand what types of relationships front-page and in-text citations to science represent. A rough draft is reproduced at the end of this document. The questions aimed to examine operationalize five concepts:

- *familiarity:* how familiar are the inventors with the cited scientific article (q1 and q2)
- *learning:* how and when the inventor learned about the cited article (q3, q4, q5)
- *similarity:* the extent to which the invention and article are similar (q6)
- *relatedness:* the specific ways (if any) in which the invention and cited article are related (q7)
- *cruciality*: how crucial the cited article was for the development of the citing invention (q8)

The first seven questions were adapted from the NBER/Case Western survey. The final one is adapted from Mansfield's work (Mansfield 1995), and helps us get at the counterfactual: but for the cited research would the citing invention have been possible? This is particularly useful to know for exercises that use citation linkages to try to estimate the share of inventions enabled by publicly funded science and the returns to publicly funded research or the rate of return to public funding (Azoulay et al 2018; Sampat and Lichtenberg 2011).

While the final survey will be a mixed mode survey (with first contact by mail) we administered the pilot online via Qualtrics, by emailing inventors. We started with an inventor from a random sample on 1,700 patents issued in 2016 that included at least one science reference (front page or full text). Through Mechanical Turk and hand-searching, we were able to locate an email address for about 70 percent of the inventors, and sent out 1,166 pilot surveys. Of the 1,166 inventor addresses we found, 110 bounced back, leaving a final effective sample of 1,056 inventors. Of these 122 responded, for a response rate of about 12 percent. Not surprisingly the response rate to the online

survey was lower than that for previous mail surveys of patent inventors (Jaffe et al 2000; Nagaoka and Walsh 2009).

Though only a pilot with a small number of responses, the preliminary results are interesting For example, inventors were more likely to be "very familiar" with full-text citations than front-page citations (42 percent vs. 33 percent) learned about these citations before or during development of the invention (42 percent vs 36 percent), actually read the articles or saw them presented (59 percent vs. 47 percent). Front page references were more likely to be described as "very closely" related to the citing invention (14 percent vs 5 percent). While the minority of references were described as crucial for the development of the citing invention, front-page references were much more likely to be than full-text references (19 percent vs 6 percent). Full-text references were more likely to be cited as background knowledge (46 percent vs 40 percent) and front-page references as tools or techniques used to develop the invention (25 percent vs. 20 percent). Few of these differences are statistically significant at conventional levels; this is unsurprising given small sample size. But they point to interesting patterns, and suggest that front-page and full-text references are indeed different in terms of inventor familiarity, learning, relatedness, connections, and cruciality.[1]

Based on feedback from the pilot, we clarified several questions and changed some response scales. (The version below reflects this feedback.) We also contacted non-respondents and learned that many were wary of clicking on unknown links and potential phishing attempts, a common issue in email surveys (Dillman et al 2014), which confirmed to us that for the full survey a mixed-mode survey would be preferable. Finally we learned that there were not large significant response bias patterns across observables, but slightly lower propensity to respond among non-academic and non-U.S. inventors; we will use this information to oversample these groups in the final survey.

Going forward, we hope to administer a mixed-mode survey with multiple modes of contact, which has been shown to yield highest response rates for surveys like ours (Dillman et al 2014). We will administer a survey with four contacts: (1) by mail, with a survey packet; (2) postcard reminder, with option to complete online as well; (3) follow-up email reminder for the non-respondents for who we can locate an email address (approximately 70 percent, based on our pilot); (4) final follow-up mail reminder with survey packet and option to complete online. Based on previous mail surveys of inventors on patents (Nagaoka and Walsh 2009; Jaffe et al 2000) we conservatively expect a response rate for a mixed mode survey to be at least 20 percent, compared to the 12 percent we obtained through our pilot online survey. At a 20 percent response rate, we anticipate a survey of 10,000 inventors would yield approximately 2,000 responses, more than enough to be useful for training the machine learning algorithm, to detect differences between front-page and in-text citations, and to generalize to the target population.

---

[1] Though this was not the primary aim of the pilot, or of our current work, it is interesting to note if we compare to results on patent-article citations to those from the NBER patent-patent survey, inventors are much more likely to be "very familiar" with report patent-article citations (whether front-page or full text), much more likely to have learned about science references before the patent application process, much more likely to have read or saw the references, and much more likely to report they actually learned something from the cited reference. Our survey could be easily extended to specifically compare patent-article references to patent-patent references (both front-page and full-text) in future work.

We will use the resulting survey data (with weights to account for non-response bias) to provide basic descriptive data on how familiar the inventors are with the cited articles, how and when they learned about the work, the relatedness of the citing invention and cited article, the types of articles that are cited, and how crucial the cited article is for the research. This will be similar to the analysis in Jaffe et al (2000), but we will also examine differences between front-page and full-text citations, to better understand what types of linkages (if any) each represent.

We also plan to use the survey responses for in-text citations as a training set for a computational classification algorithm based on responses from questions 7 and 8 of the survey, on why a patent was referenced in the patent text, and which references are particularly important. Specifically, we hope to use modern machine learning methods---convolutional neural nets (CNNs) with word embedding---to attempt to categorize all of the citations which were not surveyed, i.e. extend the categorization to all full-text references to science in all patents issued since 1984 creating a new database of science citations with fields indicating likely relationships between and importance of the cited article and citing patent. Our experiments (in consultation with Kory Mathewson, an expert in computational text classification at the University of Alberta) suggest that with 2000 survey responses, we will be able to classify between 80 and 90 percent of out-of-sample citations correctly.

The combination of a machine learning tool applied to the full corpus of patent text, *in conjunction* with survey responses, allows statistical inference *as if* the survey size was far larger. While our application here is patent references, we anticipate this approach may be more broadly useful in economic and legal research as well.

Survey Draft

Start of Block: Default Question Block

Background This survey aims to help understand the relationships between scientific articles and patents that reference them. Its goal to help improve understanding of how to measure connections between science and technology.

We are surveying inventors from a large random sample of patents that were issued in 2016, and asking questions about a scientific article referenced by the article, also chosen at random.

This should take only 3 minutes to complete.

Your recently issued U.S. Patent **PATENT NUMBER HERE** included a reference to the following article **ARTICLE HERE.** If we made a mistake and you are NOT an inventor on this patent, please scroll to the bottom to let us know.

Q1 Please indicate the degree to which you are familiar with the research in the article, ranging from 1 (not familiar) to 5 (very familiar)

- 1 (1)
- 2 (2)
- 3 (3)
- 4 (4)
- 5 (5)

Q2 How would you rate your familiarity with the article?

- I am not familiar with the article (1)
- I have an idea of what the article is about (2)
- I know the topic and key finding (3)
- I know the method and details of the article (4)
- I would be comfortable giving a talk about the article (5)

Q3 When did you learn about the research in the article?

- Before I began working on the patented invention (1)
- During the time I was working on the patented invention (2)
- After I finished working on the patented invention (3)
- Not until now (4)

Q4 Who included the reference to this article in your patent? (Check all that apply)

- [ ] Me (1)
- [ ] Another listed inventor (2)
- [ ] A patent attorney (3)
- [ ] A patent examiner (4)
- [ ] Don't know (5)

Q5 How did you learn about the research in the article? (Check the one statement that best applies)

- O Word of mouth (1)
- O Direct communication with the author of the article (2)
- O Presentation of the article (3)
- O Demonstration or viewing of a product or prototype of the research described in the article (4)
- O Read the article (5)
- O Became aware of the article during the patent application process (6)
- O I (or someone in my lab) wrote the article (7)
- O Other (Please specify) (8) _____
- O Don't remember (9)

Q6 Indicate the degree to which the patented invention and article are related, ranging from 1 (not related) to 5 (closely related)

- O 1 (not related) (1)
- O 2 (2)
- O 3 (3)
- O 4 (4)
- O 5 (closely related) (5)

Q7 <How would you characterize the link between the information in the article and your patented invention? Check the one statement that best applies.

- O It is a technique or method used in researching or developing the invention (1)
- O It is a tool or input used in developing the invention (2)
- O It contains facts or background knowledge related to the general field of the invention (3)
- O It contains facts or background knowledge motivating why the problem we set out to solve is novel (4)
- O It contains facts or background knowledge suggesting the technical feasibility of our invention (5)
- O It shows potential uses of our invention (6)
- O It shows that previous similar inventions exist (7)
- O I don't know (8)

Q8 How important was the information provided by the article for the development of your invention?

- O 1 (Not important) (1)
- O 2 (5)
- O 3 (6)
- O 4 (2)
- O 5 (Very important) (3)

Open ended We are attempting to understand how and why scientific work is referenced in patents. If there are any relationships that our questions above do not cover please let us know here:

_____

_____

_____

_____

_____

Mistake If we made a mistake and you are NOT an inventor on this patent, please let us know below:

_____

Thank you for your time. If you would like a copy of the analyses based on this survey please check the box below and we'll send to you once complete.

- O Yes, please send me a copy of any analyses based on this survey (1)