NBER Digitization Proposal

Christopher Snyder
Joel Z. and Susan Hyatt Professor
Department of Economics
301 Rockefeller Hall
Dartmouth College
Hanover, NH 03755

Tel: 603-646-0642
Fax: 603-646-2122
Email: chris.snyder@dartmouth.edu
Web: http://www.dartmouth.edu/~csnyder/

Overview: A new business model for scholarly journals, open access, has gained wide attention. Under the traditional business model, journals make most of their revenue from library subscriptions and charge authors very little. By contrast, an open-access journal's articles are available over the Internet free of charge to all readers; revenue to cover publication costs comes from authors' fees.

Theoretical analyses of the market (to which I have contributed—see McCabe and Snyder (2005, 2007)—have used two-sided market models in which journals serve as an intermediary between authors on one side of the market and readers on the other. Each side obtains benefits from other, authors from having significant impact through a wide readership, readers from the knowledge gained from the authors' articles. The prices that journals charge to each side balance these externalities. The answer offered by the theory to many policy questions—including whether open access will win out in competition with traditional journals and whether open access is socially efficient—often hinges on the elasticities of demand on the two sides of the market. The elasticity of author demand for open access depends on the benefits authors gain from open access in terms of wider readership and greater impact as captured by citations and other measures.

As explained below, much existing work attempting to measure the benefits of open access has had trouble separating this effect from selection effects, often resulting in huge overstatements of the citation impact of open-access publishing. A promising recent study (Gargouri et al. 2010) found a source of exogenous variation (open-access mandates at scholars' home institutions) to identify the causal effect of self-archiving (a proxy for open access) on citations. Although the authors started with a clever idea, the methods and models applied are flawed, leading the authors to jump to unwarranted conclusions (namely that previous studies are probably free from selection bias).

I propose to construct a model and simulations to clarify the flaws in the existing methodology. Using a similar dataset to the authors' (but updated to make it more current and enlarged), I propose appropriate methods to provide one of the first estimates of the true causal effect of open access on citation rates.

Methods and Analysis: The proposed paper will analyze the citation boost that authors experience from open access. I will build on a clever idea in Gargouri et al. (2010) of using the mandate at certain institutions that their affiliated scholars archive a version of their work in some open-access outlet even if the final journal article appears in a closed-access journal.

Self-archiving has been taken as a proxy of open access by a number of previous studies into the effect of open access on citations. Unfortunately most of these studies suffered from selection bias. Authors who archive their own work (or have the resources to have an administrator do so) are likely different from authors who do not. The former may have better resources and produce higher-impact research. One might try to controlling for such differences by including scholar fixed effects, which amounts to comparing citations of self-archived to non-self-archived articles within the body of a given scholar's work. This approach may make matters worse, as articles a given author does not bother to post on a website are likely to be inferior to ones which he/she decides to highlight.

The idea in Gargouri et al (2010) is to use an institutional open-access mandate as an exogenous source of variation in self-archiving. Unfortunately, the study is flawed in methodology and interpretation. The most serious is the use of the mandate variable as a regressor rather than an instrument, making the interpretation of the uninstrumented self-archiving variable problematic.

I propose to develop an economic model of self-archiving behavior in a world in which archiving takes effort and articles may differ in quality. I will use this model to explain the empirical results in the previous paper and indicate that the conclusions drawn from the way the regressions are specified are unwarranted. I plan to demonstrate that their results do not prove that there is small or no selection bias in earlier studies.

I will then use an update of the authors' own data to run regressions to measure the selection bias, then purge the selection bias and estimate an unbiased, causal effect of self-archiving on citations. Technically, I will estimate a regression with citations as the dependent variable in which the institutional mandate is used as an "instrument" (in the sense of instrumental variables) for the endogenous self-archiving variable. I plan to develop specifications that will allow the estimation of a self-archiving variable along with institution fixed effects. This will likely require a more complicated instrument than simply the institutional mandate, because the latter would be wiped out by the institutional fixed effects. Interactions with time trends or time fixed effects may be required, essentially exploiting variation in the stringency of the mandate over time. Of course such a specification will require a detailed analysis of the evolution of the institutional mandate, understanding whether the mandate took time to spread among the faculty or whether it was stronger when it was novel, but then was gradually ignored by the faculty afterwards.

Another methodological detail regards how the count nature of citations data is treated. I propose to improve this treatment by using Wooldridge's (1990) Poisson quasi-maximum-likelihood method). There are GMM routines which allow estimation of this sort of count data models in a panel setting also allowing for instrumental variables.

Data Availability: The data should be available because of the PLoS policy of data availability. This does not guarantee availability: Savage and Vickers (2009) document a mere 10% success rate in obtaining data from PLoS authors. However, I have contacted the authors and have an agreement that they will share an updated dataset.

References:

Gargouri Y. et al. (2010) "Self-Selected or Mandated, Open Access Increases Citation Impact for Higher Quality Research," *PLoS ONE* 5(10): e13636.

McCabe, M. and C. Snyder. (2005) "A Model of Academic Journal Quality, with Applications to Open-Access Journals," *American Economic Review Papers and Proceedings* 95(2): pp. 453-458.

McCabe, M. and C. Snyder. (2007) “Academic Journal Pricing in a Digital Age: A Two-Sided-Market Model,” *B.E. Journal of Economic Analysis & Policy (Contributions)* 7(1): article 2.

Savage, C and A. Vickers. (2009) “Empirical Study of Data Sharing by Authors Publishing in PLoS Journals,” *PLoS ONE* 4(9): e7078.