Digitization of Out-of-Print Books:
Estimating Demand and Welfare Impact

Proposal to the NBER Economics of Digitization and Copyright Initiative

Michael D. Smith and Rahul Telang
Carnegie Mellon University

Digitization and digital marketplaces have introduced significant changes for the distribution of media content. One notable change is in the distribution of books. Prior to digitization, the fixed and marginal costs of printing meant that books could only be profitably published in print runs of 2,000 to 5,000 copies. As a result, when expected demand for a book dropped below these levels, the book would typically cease to be commercially available. There are currently roughly 5 million such out of print English languages titles.

Digitization changes the economics of producing and distributing books. Digital books have relatively low fixed costs of production (essentially the cost of putting the book in a digital format), and near zero marginal costs of reproduction and distribution. This means that books that could not be profitably produced in a traditional print format, can be profitably digitized and made available in an electronic format.

However, reintroducing out of print books as eBooks raises significant copyright challenges, as evidenced by the controversy surrounding Google's book digitization project. The debate usually surrounds how the monies should be allocated. However, in the face of this copyright debate, little thought has been given to the social welfare that might be generated by bringing out of print books back to circulation. The goal of this research proposal is to fill this gap by estimating the demand for out of print books, and in turn, the potential social welfare gain from making out of print books available in an electronic format.

Proposed Method:

By construction, the demand for out of print books cannot be observed. In some cases, one may be able to go back in time to observe the demand for these books when they were in circulation. But such historical data would be hard to obtain. Moreover, most books go out of circulation because their demand is too small to justify the cost of additional print runs, or too small to justify the cost of determining copyright when ownership of the book is unclear (e.g., authors may pass away or publishers may go out of business).

However, in spite of the presumed low sales of out-of-print books, we believe that we can use available data to estimate the economic impact of bringing out of print titles back into print. We propose to do this in two ways. Our primary estimation method is to construct an appropriate counterfactual sample for the out of print books from the currently selling books, and then use this sample to estimate demand if these books were in print.

To do this, we will select a random sample of 15,000 out of print titles from Bowker's Books Out of Print database. These books can be characterized by genre, age, date when the

book went out of print, binding type (hardcover or paperback), and "Google search frequency" (which may be correlated with demand). We plan to use this as our control group.

We then plan to collect a large sample (5,000 to 10,000 books) of titles that are out-of-print as physical books, but are currently in eBook circulation from Amazon, as a basis for selecting an appropriate treatment group. We will select these titles after deliberately oversampling for older, low selling books, using sampling techniques based on our prior work with Amazon sales-rank distributions and new data documenting Amazon weekly sales, obtained from a major publisher.

Next we will select an appropriate control group from this sample of low selling titles. To do this, we can observe the characteristics of these (currently out-of-print) books, using the variables described above for the treatment group. We will then use propensity score techniques to find a set of 1,000-2,000 treatment group titles (out-of-print, but available as eBooks) that match the characteristics of our control group (out-of-print in both channels).

Specifically, we will first estimate: $P(\text{treated}) = X\beta + \varepsilon$

Here X's are the covariates (age of the book, genre, hardcover or paperback, frequency of Google search), β are the estimated coefficients and ε is the error term.

With these estimates, we can match books in the control and treatment groups with similar propensity scores. An important point to note is that, unlike other settings where selection raises significant endogeneity concerns, we do not believe that is an insurmountable problem here. Discussions with publishers suggest that heterogeneity in strategy across publishers, authors, and titles means that out-of-print books are brought back into print as eBooks for fairly random reasons. Thus, the propensity score technique should give us reasonable starting point to form the control and treatment groups.

Once we have a matched sample, we can impute the observed demand for the treatment group as the lost demand in the control group. Given that Amazon's Marketplace is a large source of out-of-print sales we also plan to use techniques developed in our research group to monitor marketplace (used) sales for both our control and treatment groups to further validate sales levels and potential unmet demand for out-of-print titles.

Our secondary estimation method is to scrape upcoming eBook releases on Amazon to identify titles that are currently out of print, but will be made newly available in an eBook format. Amazon's site is such that we can identify these titles well in advance of their eBook availability date and then track both initial eBook sales ranks and using the methods discussed above, estimate how eBook availability changes the demand for used print titles.

Social Welfare Estimates:

We have made significant progress in collecting data using both methods described above and are confident that these techniques will provide a sufficiently large and diverse set of books and give us a reasonable estimate of the demand curves for these titles. We plan to experiment with a few curves to generate a demand curve that fits our data the best and based on the variance in our data possibly estimate separate demand curves for different

genres. Once we have the estimated demand curves, we plan to use established economic techniques to estimate both consumer surplus and social welfare gains for out-of-print book in our sample and then generalize this to the total set of out-of-print titles.