Google Scholar and Free or Open Access Scholarly Content: Impact on Academic Libraries

C. Sean Burns, sean.csb@gmail.com, @burnscsb, http://web.missouri.edu/~csbc74

School of Information Science and Learning Technologies, 111 London Hall, University of Missouri, Columbia, MO 65211

Take Home Message

If we have a measure of retrieved, relevant journal articles and add to that what we know about information seeking preferences, while taking into consideration the decentralization of scholarly content, we can begin to ask:

> *What's the probability that scientists used the academic library (or Google) as a research starting point given that they retrieved a relevant, full text journal article?*

This research suggests that there's an 82% chance that an academic library provided the content and an 18% chance that the academic library was bypassed due to the availability of services such as Google Scholar in combination with the availability of content such as open access journal articles.

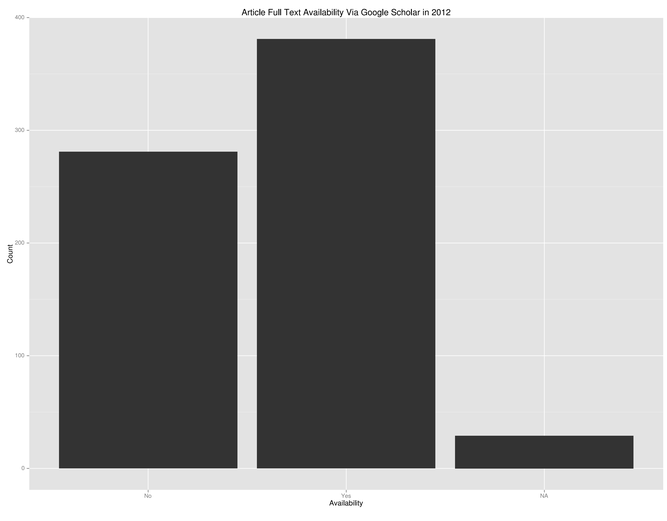

Fig. 1: CiteULike.org users collect more articles with greater potential full text availability through Google Scholar.

Arguing that academic libraries risk "disintermediation", Ithaka S+R (2010) shows that 38% of scientists claim to use Google as a research starting point. Per my data, as of 2012, Google Scholar provides full text access to 58% (p = 0.0001, 95% CI [54, 61]) of the articles that CiteULike users collected as of 2010. If scientists use the library as a starting point 38% of the time and have a 58% potential success rate to retrieve a relevant, full text article, then applying Bayes' Theorem (Phillips, 1973) suggests the above take home message.

Problem Statement

If academic libraries do not have a monopoly on providing scholarly documentation, then academic libraries are in competition to provide such access.

Purpose & Theoretical Framework

Is it rational to use Google Scholar to search for and retrieve open access content instead of, for example, Scopus to search for and retrieve licensed content? This question directs the purpose of this study, which is to provide insights into the strategic future of academic libraries. As such, it employs a decision and game theoretical framework, underpinned by Herbert Simon's concept of *bounded rationality* and George Zipf's *principle of least effort*.

Research Questions

I. Are academic libraries being disintermediated from the search process and are its collections becoming irrelevant?

a. What is the probability that someone can use Google Scholar to access the full text of a relevant journal article without the benefit of a proxy?

b. What bibliometric or publishing characteristics drive full text access to journal articles?

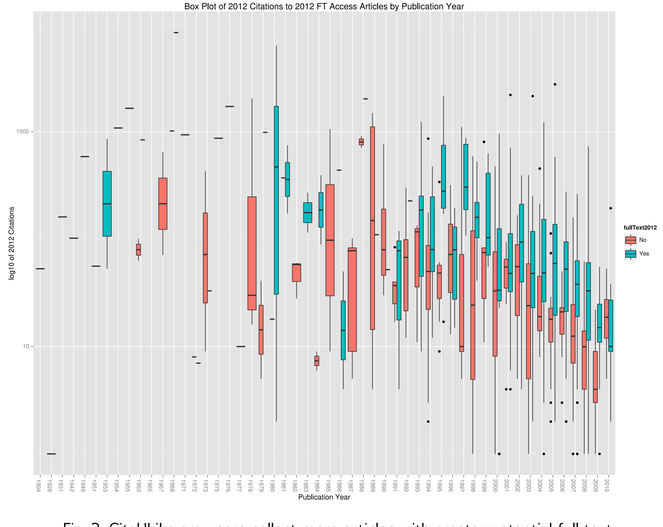

Fig. 2: CiteULike.org users collect more articles with greater potential full text availability by publication year and by citation count. However, citation count does not seem to have an effect on full text availability (OR = 1.000, p = 0.5170).

Data Collection & Method

1. May 18, 2010: Took a systematic random sample of 999 out of 2,419,452 unique bibliographic references from CiteULike.org. Sample includes 691 article references.

2. July 2010, July 2011, July 2012: Gathered bibliometric data for these references from Google Scholar.

3. Conducted a bibliometric and citation analysis on the collected data.

4. Performed a logistic regression to model the relationship: full text availability ~ author count + publication year + post year + citation count.

5. Assessed the probability, using Bayes' Theorem, of the research starting point of CiteULike.org users.

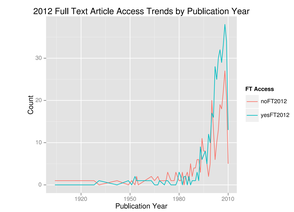

Fig. 3: CiteULike.org users exhibit a trend in collecting journal articles with greater potential full text availability through Google Scholar if those articles have been published more recently.

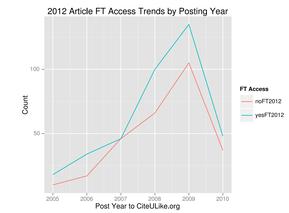

Fig. 4: CiteULike.org users have always collected journal articles that have had greater potential full text availability through Google Scholar.

Table 1: Most influential sources for full text articles in 2012, includes only sources with greater than 3 instances. There were 230 unique sources, in total, providing access to 381 articles via Google Scholar without need of a proxy, up from 177 for 345 articles in 2010.

NIH	40
arXiv	26
Oxford Journals	12
PNAS	11
BioMed Central	11
PLoS	5
CiteSeerX	5
Harvard University	5
Rockefeller University	4
Am. Meteor. Society	4

Conclusion

The key interpretation is that we cannot argue that academic libraries are risking disintermediation based on data about research starting points alone. Rather, it is helpful to take into consideration the success of the starting point, where success is retrieving and collecting a relevant, full text article, as well as the source location.

Academic libraries are very important stakeholders in the scholarly communication system. Given that researchers do use and increasingly prefer services such as Google Scholar, a strategic reply, leading to game equilibrium, should include access to source material via institutional repositories, discoverable through such services.

Bibliography

http://goo.gl/lHqTC