Querying the Library Genesis Database: Using *Polymer Science* and *Colloid and Polymer Science* to Assess Relationships Between Piracy and the Citation Record

Frances Corry, Carolina Alvarado
Piracy Lab, Columbia University, Spring 2013

INTRODUCTION

Piracy's cultural legacy is primarily marked by its negative economic effects on various media industries, from MP3s to ebooks. The academic publishing world, while engaged in this monetary economy, also engages with an economy of *citation*. Often we measure the influence of a work through its citation history and its presence in the market—piracy might unsettle these traditional modes of evaluation. This citation record also is of great importance for universities when hiring or deciding tenure (Adler, Ewing, Taylor).

To date, there is scarce literature directly analyzing piracy's engagement with academic publishing; in this study, then, we seek to probe this connection and establish a model for further research. This study is part of Piracy Lab, an academic research collective exploring the impact of piracy on global literary culture, founded in 2013 and affiliated with Columbia University. In this early iteration, the Lab's research has focused on Library Genesis. Library Genesis is an online database of academic texts, primarily from the natural science and engineering fields.

Using metadata from the Library Genesis catalog, this examination sought to determine whether there is a significant difference in citation rates between works uploaded to the Library Genesis database versus comparable works not uploaded to the database. (Descriptions of the dataset, among other materials, can be found on Piracy Lab's website: http://piracylab.org/category/research/.) This study uses *Advances in Polymer Science*, whose entire catalog was uploaded to the database over the course of 2008, and *Colloid and Polymer Science*, a journal that did not exist in the database.

DATA COLLECTION

The Library Genesis database holds several million documents in a variety of languages. This set of documents includes journal issues, individual journal articles, magazine articles, books, and more. Because English-language academic journal's citation records are well documented and easy to access, particularly through resources like *Web of Science*, (via Columbia University's subscription), we narrowed this dataset to fit this criteria.

While the time it takes for an article to be cited by others varies greatly among the disciplines, most fields calculate a journal or an article's impact factor based on the 2-4 years after it was published (Gavin Hall). Therefore, we sought a journal whose files had been uploaded to the database in 2008 or before.

Advances in Polymer Science, begun in 1959, is a well-established English language academic journal that publishes several times per year (volumes per year vary). It is ranked 3 of 79 in the category of Polymer Science in the Thomson Reuters ISI Web of Knowledge. Its entire catalog was uploaded to Library Genesis database over the course of 2008.

Given this record, we sought a comparable journal - that is, a well-established English language journal in the field of Polymer Science, whose issues had not been uploaded to the Library Genesis database. *Colloid and Polymer Science* is an English language journal in this field, which had not been uploaded to the Library Genesis database. Begun in 1906, it is published bi-monthly and is ranked 28 of 79 in the category of Polymer Science.

Given these two journals, bibliometric data was collected from 1980-2008 using the Thomson Reuters ISI Web of Knowledge database. Data was organized by year, and included the total number of citations for that year's articles, the number of citations prior to and including 2008, and the number of citations following 2008. For instance, articles published in *Advances in Polymer Science* during 1980 have been cited in other journals a total of 339 times; 311 of those citations were between 1980 and 2008; 28 of these citations were between 2009-2013.

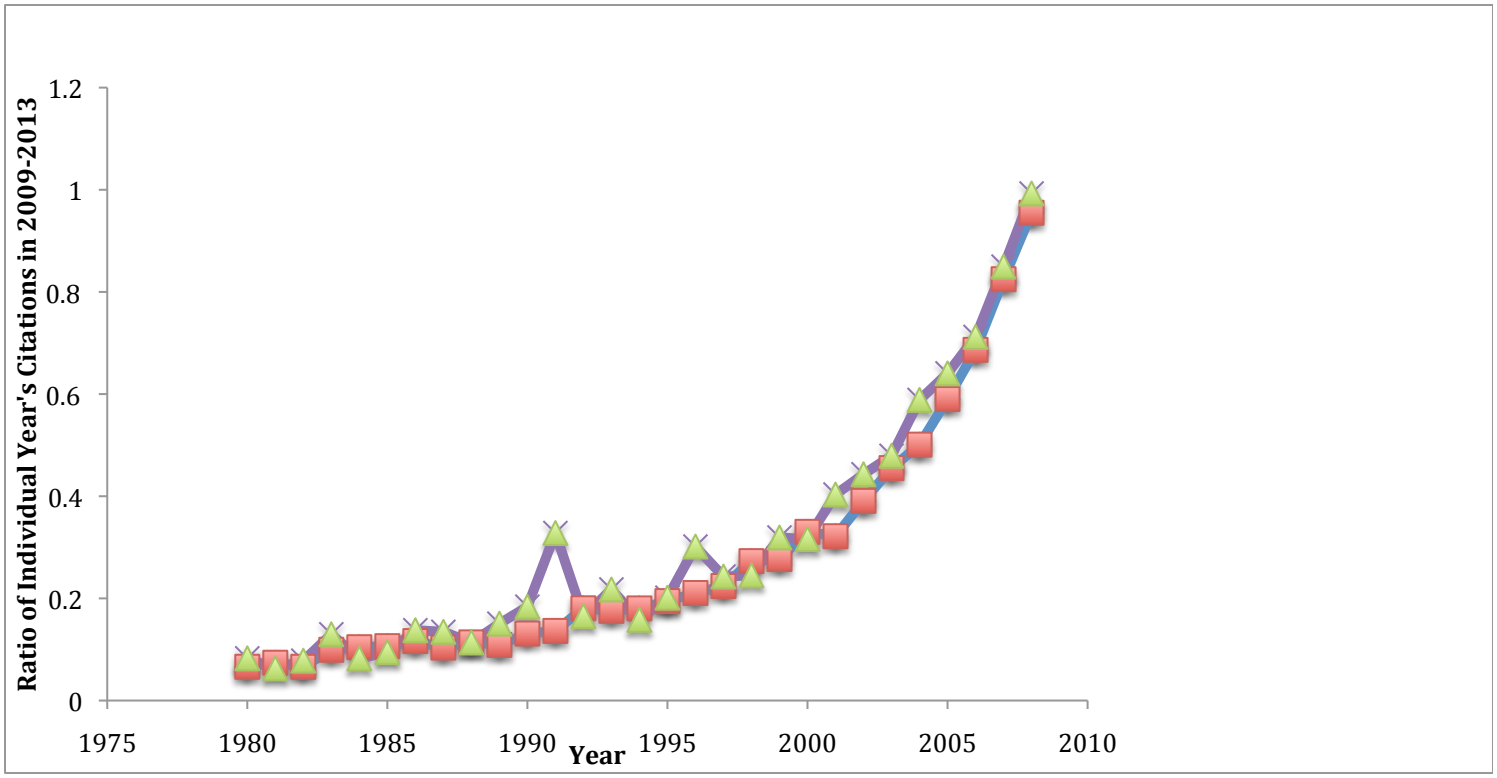
DATA ANALYSIS

This data was used to calculate each year's ratio of citations from 2009-2013 as part of their total citation count. For example, using the 1980 *Advances in Polymer Science,* 28 citations from 2009-2013 represent a ratio of 0.0826 (8.26%) of their total 339 citations. These ratios were then compared against calculated ratios from *Colloid and Polymer Science*.

1. Ratio of 2009-2013 Citations in Journal's Total Citation Count 1980-2013

Year	Colloid and Polymer Science	Advances in Polymer Science
1980	0.066131498	0.08259587
1981	0.074022989	0.062720848
1982	0.066030231	0.078064516
1983	0.099378882	0.129726585
1984	0.104458042	0.082364532
1985	0.106504065	0.094500801
1986	0.116438356	0.137645108
1987	0.103722443	0.134582624
1988	0.113874346	0.113815789
1989	0.108922364	0.150519031
1990	0.131129272	0.183754993
1991	0.136312849	0.327956989
1992	0.180622837	0.165133395
1993	0.175603217	0.217429194
1994	0.18037518	0.158023185
1995	0.194647202	0.201354402
1996	0.210526316	0.301518438
1997	0.224199288	0.242768009
1998	0.272937294	0.244988864
1999	0.277017784	0.319221968
2000	0.330272109	0.316223648
2001	0.321710832	0.403361345
2002	0.39090065	0.443064182
2003	0.455098223	0.479192938
2004	0.500986972	0.588557517
2005	0.590227651	0.640542786
2006	0.686567164	0.71188673
2007	0.82490484	0.849541284
2008	0.954949615	0.992688871

2. 1980-2013: Ratio of Individual Year's Citations in 2009-2013 from Individual Year's Total Citations



3. *Testing for Significant Difference: Unpaired T Test*

An unpaired T test was used to determine whether the data sets are significantly different.

Group	Colloid and Polymer Science	Advances in Polymer Science
Mean	0.27580939693	0.30530153248
SD	0.23456194854	0.24432210071
SEM	0.04355706036	0.04536947512
N	29	29

P value and statistical significance:
The two-tailed P value equals 0.6409

By conventional criteria, this difference is considered to be not statistically significant.

Confidence interval:
The mean of Colloid and Polymer Science minus Advances in Polymer Science equals -0.02949213555
95% confidence interval of this difference: From -0.15548320978 to 0.09649893867

Intermediate values used in calculations:
t = 0.4689
df = 56
standard error of difference = 0.063

CONCLUSION

The p-value determined from the t-test, and the simple visual similarity of the data when graphed, suggests that *Advances in Polymer Science's* upload to Library Genesis did not affect its citation count in the five years that followed its upload.

Why might this be? An English language article's net citations may not be correlated with an article or journal's having been uploaded to the Library Genesis database because those individuals likely to be publishing for and cited in these journals – ie those in the academic world – already have access to these journals via their affiliated institution. To put it simply, those publishing in these journals have legal access to these articles and do not need the pirated version in Library Genesis.

It must be noted that this test is not without flaw. The two journals, while both well-established, published in English and about similar topic, are slightly different in terms of their ranking in the polymer science field as well as their publishing frequency. Furthermore, Web of Knowledge does not include all journals in its reaches, and we must recognize that these journals may be cited in other publications outside of the Web's purview.

In addition, this test represents only a tiny sample of the millions of documents uploaded to the Library Genesis database. This analysis, then, should serve not be looked at as conclusion, but as the very first, very small step in analyzing the impact of piracy on the academic world.

NEXT STEPS

What other questions should be asked of the Library Genesis database? What other questions should be asked about piracy and the academic world? One of Piracy Lab's projects is researching who is using, uploading, and downloading these texts. We suggest that future iterations of Piracy Lab work closely with those doing fieldwork research to ask more nuanced questions of the database. Library Genesis files are largely hosted by participants in Eastern Europe. Analyzing articles and journals published in Eastern Europe would be valuable research, or in locations where comprehensive academic databases are more difficult to access.

WORKS CITED

"About." *Piracy.Lab*. Columbia University, n.d. Web. 10 May 2013.

"Colloid and Polymer Science." *Web of Science*. N.p., n.d. Web.

Adler, R. Ewing, J and Taylor, P. “Citation Statistics.” *Statistical Science.* 24.1 (2009): 1-14. *JSTOR*. Web.

"Advances in Polymer Science." *Web of Science*. N.p., n.d. Web.

Gavin Hall, Peter. "Comment: Citation Statistics." *Statistical Science* 24.1 (2009): 25-26. *JSTOR*. Web.