Open Access Meets Discoverability: Citations to Articles Posted to Academia.edu

Yuri Niyazov
Academia.edu
yuri@academia.edu

Carl Vogel
Polynumeral
carl@polynumeral.com

Richard Price
Academia.edu
richard@academia.edu

Ben Lund
Academia.edu
ben@academia.edu

David Judd
Academia.edu
david@academia.edu

Adnan Akil
Academia.edu
adnan@academia.edu

Josh Schwartzman
Academia.edu
josh@academia.edu

Max Shron
Polynumeral
max@polynumeral.com

Abstract

Using matching and regression analyses, we measure the difference in citations between articles posted to Academia.edu and other articles from similar journals, controlling for field, impact factor, and other variables. Based on a sample size of 44,689 papers, we find that a paper in a median impact factor journal uploaded to Academia.edu receives 37% more citations after one year than a similar article not available online, 58% more citations after three years, and 83% after five years. We also found that articles posted to Academia.edu had 75% more citations than articles posted to other online venues, such as personal and departmental home pages, after five years.

Introduction

Academia.edu is a website where researchers can post their articles and discover and read articles posted by others. It combines the archival role of repositories like ArXiv, SSRN, or PubMed with social network features, such as profiles, news feeds, recommendations, and the ability to follow individuals and topics. The site launched in 2008 and as of April 2015 has approximately 20 million registered users who have uploaded approximately five million articles. Registration on the site is free and users can freely download all papers posted to the site.

There is a large body of research on the citation advantage of open access articles, and researchers are still debating the size and causes of the advantage. Some studies have found that open access articles receive substantially more citations than pay-for-access articles, even after controlling for characteristics of the articles and their authors (Eysenbach 2006; Gargouri et al. 2010). Other studies using experimental and quasi-experimental methods have concluded that any measured citation advantage is mostly due to selection bias and other unobserved differences between free and paid articles (Davis et al. 2008; Davis 2011; Gaule and Maystre 2011).

Both the supportive and critical studies have focused on the accessibility of articles: once found, can the article be obtained for free? They have given less consideration to the discoverability of articles: how easily can the article be found? This makes sense; the methods researchers often use to find articles don't privilege open access over paid sources or vice versa. Google Scholar, for example, returns both free and paid sources, as do many library databases.

Academia.edu, on the other hand, has unique features for discovering articles, making it an interesting venue for analyzing a citation advantage. Users are notified when authors they follow post articles to the site. They can then share those articles with their followers. A user can tag an article with a subject like "High Energy Physics" and users following that subject will be notified about the paper.

A number of users have reported to the Academia.edu team that they observed increased citations after posting their articles to the site (Academia.edu 2012; 2013). Motivated by those anecdotal reports, a formal statistical analysis was conducted of the citation advantage associated with posting an article to the site.

We find that a typical article posted on Academia.edu recieves approximately 37% more citations compared to similar articles not available online in the first year after upload, rising to 58% after three years, and 83% after five years. We also find that a typical article posted on Academia.edu receives more citations than an article that is available online on a non-Academia.edu venue, such as a personal homepage, a departmental

homepage, or a journal site. A typical paper posted to Academia.edu received 29% more citations than an article uploaded to a non-Academia.edu site after the first year, rising to 51% after three years, and 75% after five years.

While our study is observational, and it is difficult to conclude a causal effect, we do find the citation advantage is substantial even after controlling for some potential sources of selection bias.

Background

The Open Access Citation Advantage

Even though Academia.edu differs from traditional venues for open access, the hypotheses and methods in this paper overlap with research on the open access citation advantage.

The term "open access" typically refers to articles made freely available according to specific Open Access policies of academic journals: for example "Gold Open Access" policies where authors or institutions pay the journal to make an article freely available, or "Green Open Access" where an author may archive a free version their article online. Sometimes, though, "open access" is used more loosely to refer to any manner by which articles are made freely available online. Some authors use the term "free access" for this broader definition, to distinguish it from Green and Gold Open Access policies. Our study does not rely on these distinctions, and we will use the terms "open access" and "free access" interchangeably to refer to the broader definition of freely downloadable articles.

Many researchers, beginning with Lawrence (2001), have found that free-access articles tend to have more citations than pay-for-access articles. This citation advantage has been observed in a number of studies, spanning a variety of academic fields including computer science (Lawrence 2001), physics (Harnad and Brody 2004), and biology and chemistry (Eysenbach 2006).

The estimated size of the citation advantage varies across and even within studies, but is often measured to be between 50% and 200% more citations for open access articles. The variety of estimates is unsurprising, since both open access and citation practices vary widely across discipline, and citations accumulate at different rates for different articles published in different venues. Different statistical methods also lead to different estimates. Some studies have simply compared unconditional means of citations for samples of free and paid articles, while others, such as Eysenbach (2006) measured

the advantage in a regression analysis with a battery of controls for characteristics of the articles and their authors.

Critiques of the Citation Advantage

Other studies have presented evidence against an open access citation advantage, arguing that although there is correlation between open access and more citations, open access does not cause more citations.[1]

Kurtz et al. (2005) — in a framework adopted by several subsequent authors[2] — put for three postulates to explain the correlation between open access and increased citations:

1. **The Open Access postulate.** Since open access articles are easier to obtain, they are easier to read and cite.
2. **The Early view postulate.** Open access articles tend to be available online prior to their publication. They can therefore begin accumulating citations earlier than paid-access articles published at the same time. When comparing citations at fixed times since publication, the open-access articles will have more citations, because they have been available for longer.
3. **The Selection Bias postulate.** If more prominent authors are more likely to provide open access to their articles, or if authors are more likely to provide access to their "highest quality" articles, then open access articles will have more citations than paid-access articles.

Kurtz et al. (2005), and later Moed (2007), concluded that the Early View and Selection Bias effects were the main drivers of the correlation between open-access and increased citations. A lack of causal open-access effect was further supported in other studies, such as the randomized trials in Davis et al. (2008) and Davis (2011), and the instrumental variables regressions in Gaule and Maystre (2011).

But even these studies are not conclusive. For example, Kurtz et al. (2005) point out that their conclusions may be specific to their sample: articles published in the top few astronomy journals. The experimental treatment in Davis et al. (2008) and Davis (2011) was to make randomly-chosen articles free to download on the publisher's website. How easily researchers could determine these articles were available for free is unclear. The instrumental variables Gaule and Maystre (2011) are only weakly correlated with citations; as a result their estimate of the open-access advantage is imprecise.[3]

[1] See, e.g., Craig et al. (2007) and Davis and Walters (2011) for critical reviews of the citation advantage literature.

[2] See, e.g., Craig et al. (2007); Moed (2007); and Davis et al. (2008).

[3] See, e.g., Angrist and Pischke (2008), chapter 4.

Regardless of the validity or generality of their conclusions, these studies do establish that any citation advantage analysis must take into account the effects of time and selection bias on citation differentials.

Sources of Selection Bias in Academia.edu Citations

Like most citation advantage studies, ours is observational, not experimental. Articles are not uploaded to Academia.edu randomly. Authors choose to register as users on the site, and then choose which of their articles to upload. When making comparisons to articles not posted to the site, this creates several potential sources of bias in unconditional citation comparisons.

1. **Self-selection of disciplines.** Academia.edu users may be more likely to come from particular disciplines. Since the citation frequency differs across disciplines, a citation advantage estimate that doesn't control for academic discipline might over- or under-estimate the true advantage.
2. **Self-selection of authors.** Researchers who post papers on Academia.edu might differ from those who do not. Users might skew younger, or be more likely to work at lesser-known institutions. If so, we would expect to find that papers posted to the site tend to have fewer citations than those not. Or users might skew in the other direction—having more established reputations, or coming from better-known institutions, in which case we could overestimate the actual advantage. Furthermore, users who post papers may also be generally more proactive about distributing and marketing their work, both through Academia.edu and other venues online and off. If this were true, it would also cause us to overestimate the actual advantage.
3. **Self-selection by article quality.** Even if Academia.edu users were not systematically different than non-users, there might be systematic differences between the papers they choose to post and those they do not. As Kurtz et al. (2005) and others have hypothesized, users may be more likely to post their most promising, "highest quality" articles to the site, and not post articles they believe will be of more limited interest.
4. **Self-selection by article availability.** A user may be more likely to post a paper to the site if they have already made it available through other venues, such as their personal website or institutional or subject-specific repositories. In this case, a citation advantage estimated for Academia.edu papers might be measuring in part or whole, a general open access effect from the articles' availability at these other venues.

Many of these factors cannot be observed directly or completely, and their aggregate effect on citation advantage estimates is difficult to predict. We have collected data and employed matching and regression strategies to mitigate each of the above potential biases, and continue to find a substantive citation advantage to articles posted to Academia.edu.

Data Collection

We rely on data from several sources: (1) articles the Academia.edu website, (2) citation counts and free-access status from Google Scholar, (3) journal rankings from SCIMago/Scopus, and (4) journal research fields from the Australian Research Council. All data and code use in the analysis are available for download at https://github.com/polynumeral/academia-citations.

On-Academia and Off-Academia Articles

Our analysis is a comparison of citations between articles posted to Academia.edu to articles not posted. We refer to these two samples as the "On-Academia" sample and the "Off-Academia" sample. Articles comprising each sample were selected in the following way.

On-Academia Sample: The articles in our analysis were uploaded to the Academia.edu between 2009 and 2012, inclusive. We chose to start at 2009 because this was the first full year that the site was active. We stopped at 2012 so that all articles in the sample are at least two-years old and have had time to accumulate citations. We restrict our sample to articles that were posted to the site in the same year they were published. We refer to this as the "P=U" (Published=Uploaded) restriction. This ensures that all of the articles are exposed to any citation advantage effect starting from their publication. It also mitigates bias from authors favoring their, *ex post*, most-cited articles when uploading to the site.

Our analysis relies on information from Google Scholar and CrossRef. The latter is a databse containing journals, articles, authors, and Digital Object Identifiers (DOIs). Therefore, we restricted the on-Academia sample to articles that could be matched by title and author to both Google Scholar results and CrossRef entries.

Off-Academia Sample: Using the CrossRef database, we selected a random subset of articles published in the same journals and years as articles in the on-Academia sample, but which had not been posted to Academia.edu.

Citation Counts

For all articles in both the on- and off-Academia samples, we obtained citation counts from Google Scholar between April and August 2014.

Table 1 shows the number of articles in each cohort and sample. The on-Academia sample each year is a subset of papers posted to the site that year. We excluded papers uploaded to the site that were published in an earlier year, and papers that could not be matched to a Google Scholar search result or a CrossRef entry based on their titles and authors. Users manually enter a paper's title when they upload it to the site, and what they enter may differ from the paper's canonical title. (For example, a user may add "forthcoming in PLoS" to the title.) This sort of discrepancy was a common reason for a failure to match. We do not believe that failure to match a paper is related to its citations, and therefore these exclusions should not bias our results.

Year	On-Academia	Off-Academia
2009	6,336	310
2010	7,734	1,205
2011	9,039	4,640
2012	10,670	4,755
Total	33,779	10,910

Table 1: Sample size of papers, by cohort.

Articles in the sample come from 7,176 different journals, but there is a concentrated representation of journals. Table 2 lists the ten journals with the highest number of articles in our sample. Analytical Chemistry and PLoS One comprise 4.5% of the sample, and the top ten journals comprise almost 10% of the sample.

Journal	# Articles	% Total
Analytical Chemistry	1,537	3.44%
PLoS One	492	1.10%
Anesthesia and Analgesia	430	0.96%
Biological and Pharmaceutical Bulletin	362	0.81%
Analytical Methods: advancing methods and applications	339	0.76%
Analytical Biochemistry	317	0.71%

Journal	# Articles	% Total
Applied Mechanics and Materials	303	0.68%
Bioconjugate Chemistry	299	0.67%
Applied Physics Letters	190	0.43%
BioEssays	183	0.41%

Table 2: Journals with the most number of articles in the sample.

Journal Impact Factors and Divisions

We used the 2012 impact factor of an article's journal as a matching variable and regression predictor. Journal impact factors were obtained from SCIMago Journal and Country Rank, which uses citation data from Scopus (SCImago 2007). The metric we refer to as the "impact factor" is the "Cites per Doc, 2 year" metric on the SCIMago site. A journal's impact factor is is calculated as the average number of citations recieved in 2012 by papers that were published in the journal in 2010 and 2011. The journals in our sample with the highest impact factors are listed in Table 3.

Journal	Impact factor
CA: A Cancer Journal for Clinicians	113.3
Chemical Reviews	45.62
Annual Review of Immunology	43.47
Nature Reviews: Genetics	37.52
Chemical Society Reviews	30.61
Lancet Oncology	27.96
Nature Materials	27.54
Progress in Polymer Science	27.51
Nature Reviews Neuroscience	27.34
Annual Review of Biochemistry	27.15

Table 3: Top ten journals in sample, by impact factor.

We also obtained data on the journals' fields of research from the Australian Research

Council's *Excellence in Research for Australia* report (Australian Research Council 2012). The report contains data on academic journals that includes labels for their Fields of Research, defined using a hierarchical taxonomy from the Australian New Zealand Standard Research Classification (Australian Bureau of Statistics 2008). Field of Research is the second level of taxonomy, and the journals in our sample cover around 200 different Fields.

We instead rely on the first level of the taxonomy, the "Division" of the journal, which describes broad disciplines of research. There are 22 Divisions in the taxonomy and a journal can be labelled with up to three different Divisions. Multidisciplinary journals, which cover more than three Fields of Research, are labelled with a 23rd Division label of "Multidisciplinary."[4]

Table 4 provides summary data about the Divisions in our sample: the share of articles in the full and on- and off-Academia samples in each discipline, and the median impact factor of journals in our sample in each Division. Nearly a third of articles in our sample are in Medical and Health Sciences journals, while Engineering and Biological Sciences each represent a fifth of articles. The columns add up to more than 100% because journals can be labeled with up to three disciplines.

Division	% All	% On	% Off	Med. Imp. Factor
Medical and Health Sciences	33.9%	17.0%	39.4%	2.72
Engineering	19.9%	10.9%	22.8%	2.72
Biological Sciences	18.4%	16.8%	18.9%	2.66
Chemical Sciences	15.9%	6.1%	19.0%	3.75
Psychology and Cognitive Sciences	8.1%	15.1%	5.8%	2.40
Physical Sciences	6.8%	7.9%	6.5%	2.40
Mathematical Sciences	6.5%	4.6%	7.2%	1.33
Multidisciplinary	5.6%	12.0%	3.5%	3.68
Studies in Human Society	5.2%	12.5%	2.8%	1.01
Information and Computing Sciences	4.7%	5.1%	4.6%	1.93
Earth Sciencess	4.0%	7.7%	2.8%	2.33
Agricultural and Veterinary Sciences	3.3%	4.0%	3.2%	2.15

[4]All of the analyses in the paper were also conducted with the "Field of Research" labels, using text analysis and dimension reduction techniques to account for the large number of labels and high correlations amongst them. These analyses gave nearly identical results to those based on the Division labels, so we use the latter since they are easier to interpret.

Division	% All	% On	% Off	Med. Imp. Factor
Environmental Sciences	3.2%	4.6%	2.7%	2.55
Commerce, Management, Tourism and Services	2.9%	4.3%	2.5%	1.26
Language, Communication and Culture	2.4%	6.1%	1.2%	0.48
Philosophy and Religious Studies	2.2%	6.0%	1.0%	0.52
Education	2.2%	4.3%	1.5%	1.02
Technology	2.0%	1.6%	2.1%	2.00
History and Archaeology	1.9%	5.7%	0.7%	0.67
Economics	1.8%	2.2%	1.7%	1.18
Built Environment and Design	1.0%	2.0%	0.7%	1.76
Creative Arts and Writing	0.8%	2.1%	0.4%	0.34
Law and Legal Studies	0.5%	1.1%	0.4%	0.73

Table 4: Journal Divisions, defined according to the taxonomy in (Australian Bureau of Statistics 2008). Share of articles in the full sample, the on-Academia sample, and the off-Academia sample in each Division, and the median impact factor of sample articles in the Division. Journals can be labelled with between one and three disciplines.

Online Availability

In the last section, we considered several potential sources of selection bias in the on-Academia sample. One was that users might be more likely to upload articles to the site if they have also made those articles available elsewhere online. To examine this possibility, we collected data on whether papers in were freely available from non-Academia sources.

To determine whether a paper was available elsewhere, we searched for its title on Google Scholar, and checked whether the results contained a link to a non-paywalled full-text article. This method is subject to false negatives, but we expect its error rate to be the same for both on- and off-Academia articles.

Table 5 lists the number of articles searched, and the percentage with free-access to full text on non-Academia.edu sites. We find that papers in the on-Academia.edu are more likely to be available online as papers in the off-Academia sample. This indicates that there may be some self-selection by availability in our data.

		Off-Academia	On-Academia
a.	Full-text available elsewhere	8,289	4,564
b.	Articles searched	32,591	10,181
c.	Share $a \div b$	25.4%	44.8%

Table 5: Share of sample articles freely available from non-Academia.edu sites.

Analysis and Results

Our general empirical strategy is to estimate the distribution of the citation count of article i, published in journal j in year t, conditional on it being posted to Academia.edu, and compare this distribution to the same article, but conditional on it not being posted to the site. Denoting the number of citations as a random variable Y, we are interested in the distributions

$$\begin{aligned} P^1_{ijt}(y) &= \text{Prob}\left(Y \leq y \mid j, t, \text{on-Academia}\right) \\ P^0_{ijt}(y) &= \text{Prob}\left(Y \leq y \mid j, t, \text{off-Academia}\right). \end{aligned}$$

We can compute the change in an article's citations associated with posting to Academia.edu, Δ_{ijt}, by comparing summary statistics of these distributions. For example, the difference in means

$$\Delta_{ijt} = \text{E}^1_{ijt}(Y) - \text{E}^0_{ijt}(Y),$$

or medians,

$$\Delta_{ijt} = \text{Med}^1_{ijt}(Y) - \text{Med}^0_{ijt}(Y).$$

One approach would be to directly estimate these summary statistic by computing average or median citations within each journal $\times$ year group. Unfortunately many of these groups contain too few articles to accurately estimate summary statistics. Instead, we use journal-specific covariates to represent journals, most prominently the journal's impact factor. This leads to two approaches: a non-parametric matching analysis, and a regression analysis.

Properties of citation count distributions

Citations counts are non-negative integers with a highly right-skewed distribution. This can be seen in Table 6 and Fig. 1, the latter of which also shows that the modal article has one or no citations. Our matching analysis accounts for this aspect of the data by comparing quantiles of on- and off-Academia citation counts. Our regression analysis applies several parametric models that accommodate right-skewed count data.

Sample	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
off-Academia	0	1	5	9.87	12	1,237
on-Academia	0	3	7	12.54	15	1,267

Table 6: Citations summary statistics

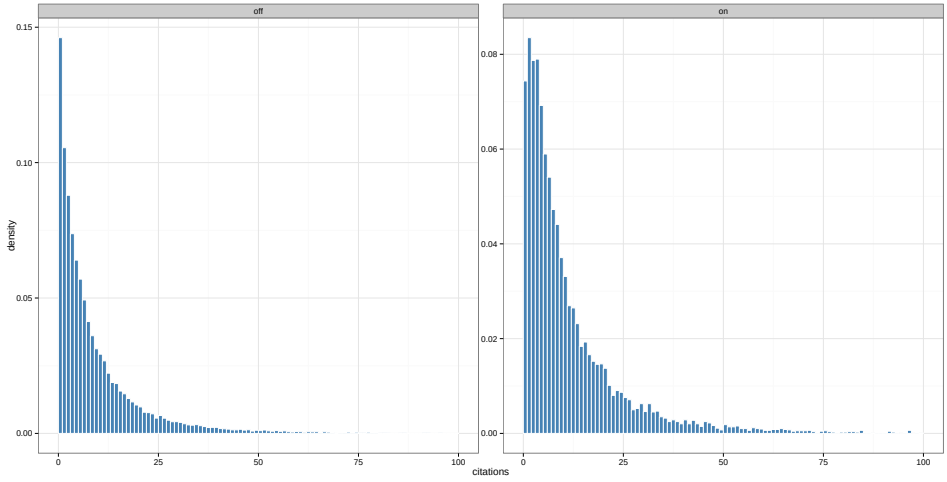

Figure 1: Distributions of citations (x-axis is truncated at 100)

Matching by Impact Factor

Our first analysis compares citations of on- and off-Academia articles grouped by cohort and their journals' impact factors. This is effectively a matching strategy with year and impact-factor as the covariates. To match on-Academia articles to off-Academia articles, we computed decile bins of impact factors amongst the on-Academia articles in a cohort. Therefore, any impact factor bin represents 10% of articles in the on-Academia sample

for that year. We then grouped the off-Academia articles into those bins, and compared samples within each bin.

Fig. 2 shows boxplots of citations to on- and off-Academia articles in each cohort and impact factor bin.[5] Evident in the figure are that older papers have more citations, and that articles published in higher impact factor journals have more citations. Furthermore, we find that median number of citations to on-Academia articles is consistently higher than off-Academia articles across cohorts and impact factor bins. Table 7 provides the medians and citation advantages for each of the comparison groups. The on-Academia citation advantage ranges from 1 extra citation for low impact factor bins to 27 for high impact bins. For low impact factor bins, the advantage is large in percentage terms-2 or 3 extra citations is a 200% increase.

Year	Impact Factor Bin	Off-Academia	On-Academia	Abs. Diff	% Diff.
2009	[0,0.43]	1	3	2	200
	(0.43,0.81]	3	6	3	100
	(0.81,1.25]	4	11	7	175
	(1.25,1.73]	6	14	8	133
	(1.73,2.14]	8	13	5	62
	(2.14,2.63]	10	19	9	90
	(2.63,3.25]	12	18	6	50
	(3.25,3.7]	12	25	13	108
	(3.7,5.28]	15	20	5	33
	(5.28,45.6]	19	46	27	142
2010	[0,0.43]	1	2	1	100
	(0.43,0.81]	2	5	3	150
	(0.81,1.25]	4	7	3	75
	(1.25,1.73]	5	9	4	80
	(1.73,2.14]	7	10	3	43
	(2.14,2.63]	8	11	3	38
	(2.63,3.25]	10	15	5	50
	(3.25,3.7]	9	17	8	89

[5] Bornmann et al. (2008), among others, advocate using boxplots to compare citation differences across samples.

Year	Impact Factor Bin	Off-Academia	On-Academia	Abs. Diff	% Diff.
	(3.7,5.28]	12	23	11	92
	(5.28,45.6]	15	24	9	60
2011	[0,0.43]	0	2	2	-
	(0.43,0.81]	2	4	2	100
	(0.81,1.25]	3	5	2	67
	(1.25,1.73]	3	7	4	133
	(1.73,2.14]	5	8	3	60
	(2.14,2.63]	5	9	4	80
	(2.63,3.25]	6	11	5	83
	(3.25,3.7]	7	12	5	71
	(3.7,5.28]	9	14	5	56
	(5.28,45.6]	12	25	13	108
2012	[0,0.43]	0	1	1	-
	(0.43,0.81]	1	2	1	100
	(0.81,1.25]	1	3	2	200
	(1.25,1.73]	2	4	2	100
	(1.73,2.14]	3	5	2	67
	(2.14,2.63]	4	5	1	25
	(2.63,3.25]	4	6	2	50
	(3.25,3.7]	5	8	3	60
	(3.7,5.28]	6	8	2	33
	(5.28,45.6]	8	14	6	75

Table 7: Median citations by cohort and impact factor bin for off- and on-Academia.edu samples.

Using impact factors to match on- and off-Academia articles serves a few purposes. First, a journal's impact factor provides a baseline estimate for the expected number of citations an article will receive in a year. This isn't a precise estimate; within a journal of a given impact factor, the citations of its articles can vary widely. As Fig. 3 shows, despite the skew of citation distributions, high impact factors are not driven by outliers. Second, using impact factor as a matching covariate should help to account for some

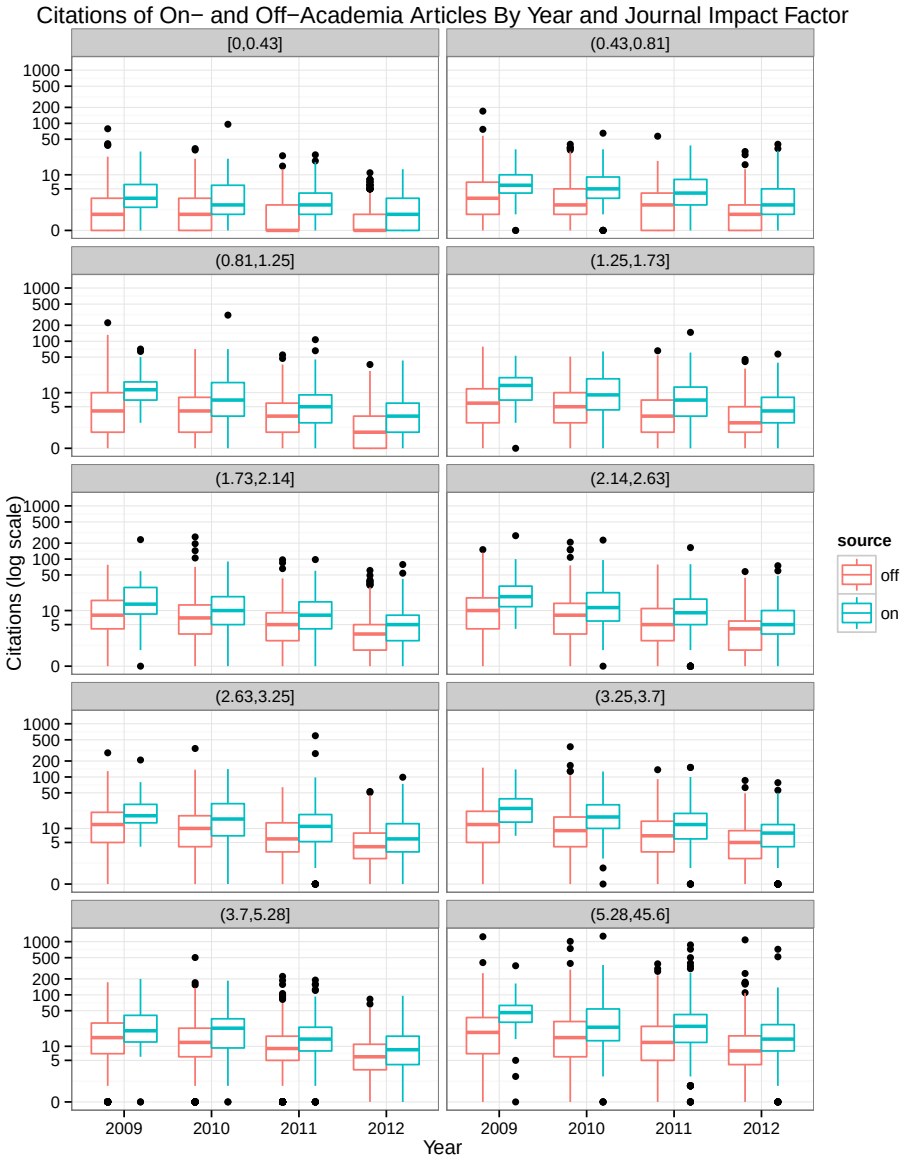

Figure 2: Boxplots of off- and on-Academia article citations, by cohort and impact factor bin.

self-selection of authors and articles. Authors typically want to publish their articles in more prestigious, higher-impact journals; the more prestigious and high-impact the journal, the more selective it can be about publishing articles it expects to be highly cited. In our sample, as seen in Table 8, impact factor is strongly correlated with citations, and explains about 22% of the variance in citations.

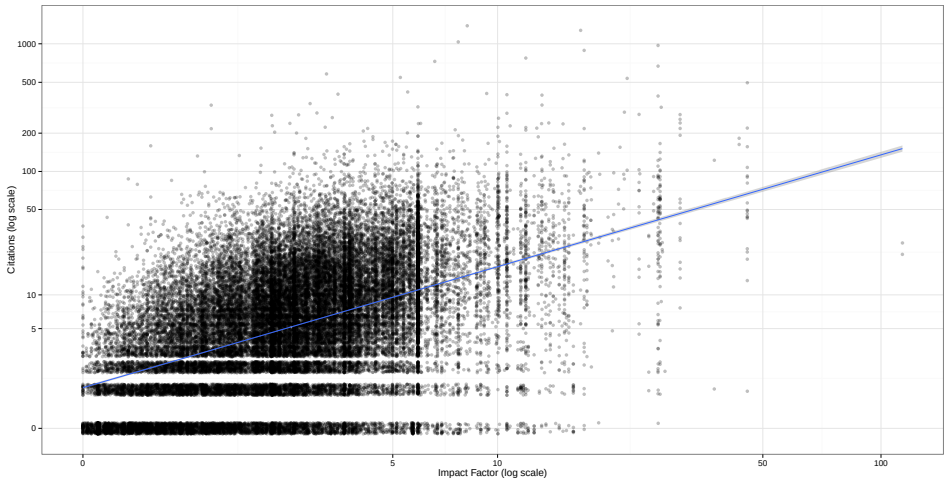

Figure 3: Article citations against Impact Factor (log scale).

	Citations (log scale)
Impact Factor (log scale)	0.906
	(0.008)
Intercept	0.731
	(0.011)
Observations	44,689
R^2	0.221

Table 8: Regression of citations against journal impact factors. (t-statistics in parentheses)

Similar results can be seen in Fig. 4, which shows scatter plots of article citations against journal impact factors (both on a log scale). The lines in the figure are predictions from separate median regressions for the on- and off-Academia group. Here we see the same result: a consistent citation advantage for on-Academia articles across cohorts and im-

pact factors.

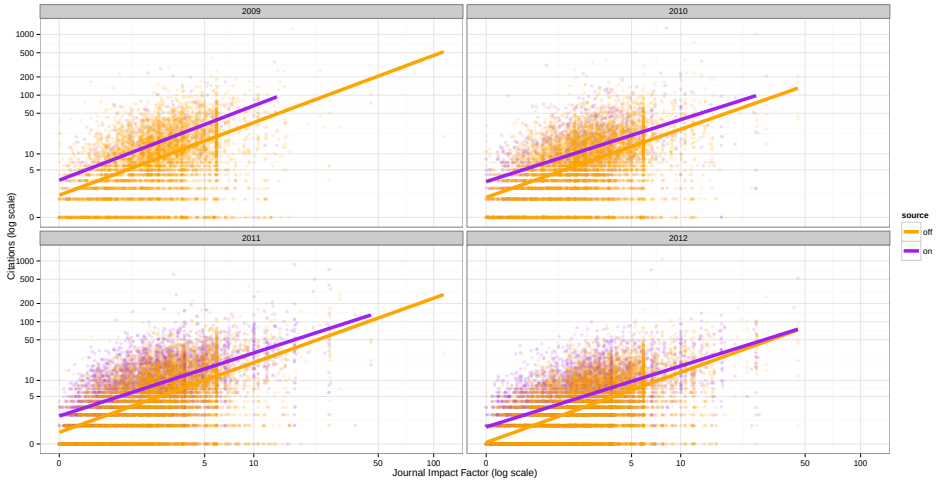

Figure 4: Citations against impact factors, with lines for conditional medians by off- and on-Academia sources.

Regression Analysis

We perform regression analyses with three different models:

1. A *linear regression* of log-scaled citation counts.
2. A *negative binomial regression* that explicitly models citations as (over-dispersed) counts.
3. A *zero-inflated negative binomial regression*. Motivated by the prevalence of uncited articles in our sample, we consider a mixture model of two negative binomial distributions. The first if a "zero" distribution that is degenerate at zero citations. Articles from this population will be uncited with probability one. The second is a "count" distribution, for articles that have a positive probability of being cited. The model estimates both the probability that an article belongs to the "zero" population, and conditional on coming from the "count" population, the probability distribution of an article's citations given its features.

Covariates

We use the same covariates in all three regression models: (1) a dummy variable equal to one for articles posted to Academia.edu, (2) the artcle's age, and squared-age, on the

date its citation data was collected, (3) the impact factor of the article's journal (on a log scale), (4) a dummy variable indicating whether the full-text of the article could be downloaded online from a non-Academia site, and (5) 23 dummy variables for the ANZSRC Divisions, indicating whether the article's journal was labelled with each Division. Variables in (2)-(4) are interacted with the on-Academia dummy to allow for varying effects by age, impact factor, and online availability. In the linear and negative binomial models, the Division dummies in (5) are also interacted with the on-Academia dummy to obtain field-specific estimates of the on-Academia effect.[6]

Summary statistics for the age, impact factor, and online-availability variables are shown in Table 9.

	Mean	Median	Std. Dev.
Age	3.07	2.92	1.04
Impact factor	2.96	2.30	3.03
Online	0.30	0.00	0.46

Table 9: Summary statistics of regression model covariates.

Linear Regression

We fit a multivariate linear regression of log citations on the covariates described above. The coefficients on the on-Academia, year, online, and impact-factor covariates are listed in Table 10, in the column labelled "Linear." For brevity, we exclude the 46 Division covariate and interaction coefficients. The age, age-squared, and impact factor coefficients have the expected signs and magnitudes. The coefficient of the on-Academia indicator is a statis 0.522. Since the age and impact factor covariates are centered to have mean zero, the coefficient implies that for an article of the mean age and journal impact factor that is not available elsewhere online, posting to Academia.edu is associated with approximately 50% more citations.

The coefficient on "on-Academia $\times$ Online" is -0.03, indicating that, for articles already freely available online, posting to Academia.edu is associating with a smaller difference in citations. At the sample average (see Table 9), the on-Academia coefficient for an online article is $0.52 - 0.03 = 0.49$. The on-Academia $\times$ Age coefficient is postive, implying that the on-Academia effect increases with time, .

[6]These interactions were excluded from the zero-inflated negative binomial model as the model typically failed to converge when they were included.

The actual effect size of being on-Academia depends on the Division of the article, so is difficult to infer directly from the coefficient. We provide effect sizes based on typical values of the covariates in the next section.

Negative Binomial Regression

The negative binomial regression uses the same covariates as the linear regression, explicitly models citations as count data. The negative binomial distribution is a common choice for modeling over-dispersed count data.

In a negative binomial regression, the number of citations to article i, y_i is modeled as a function of covariates $\mathbf{x}_i$ according to:

$$y_i \sim \text{NegBin}(\phi_i, \theta) \tag{1}$$

$$\phi_i = e^{\mathbf{x}_i \beta} \tag{2}$$

Fitting the regression provides estimates of the coefficients β and the scale parameter θ. Results for the entire sample are show in Table 10 in the column "Neg. Binom." We find similar results to the linear regression model-a large on-Academia coefficient that diminishes somewhat for articles available online elsewhere, but remains substantial. Again, due to the effects of Divisions, but also because the model is nonlinear, effect sizes are difficult to infer from the model coefficients, but we provide some effect size estimates in the next section.

	Linear	Neg. Binom.
(Intercept)	1.619	1.872
	(105.797)	(106.587)
On-Academia	0.522	0.527
	(16.387)	(14.641)
Impact factor (log, centered)	0.952	1.257
	(92.278)	(104.832)
Article age (centered)	0.558	0.798
	(16.053)	(19.986)
Article age squared (centered)	-0.046	-0.068
	(-8.939)	(-11.425)
Available online	0.144	0.180
	(12.061)	(13.151)

	Linear	Neg. Binom.
On-Academia × Impact factor	-0.028	-0.157
	(-1.303)	(-6.401)
On-Academia × Age	0.103	-0.040
	(1.217)	(-0.416)
On-Academia × Age Squared	-0.003	0.017
	(-0.238)	(1.121)
On-Academia × Available online	-0.032	-0.049
	(-1.431)	(-1.917)
N	41,891	41,891
R-squared	0.316	
Deviance	34,874.836	47,245.242
Log-likelihood	-55,601.329	-131,268.944
AIC	111,316.659	262,651.889

Table 10: Regression results. Topic keyword coefficients ommitted. t-statisics in parentheses.

Zero-Inflated Negative Binomial Regression

The modal number of citations for an article in our sample is zero or one, and approximately 15% of articles in our sample are uncited. Table 11 shows the share of uncited articles in each cohort of the off- and on-Academia samples. As expected, articles in newer cohorts are more likely to be uncited. But off-Academia articles are also much more likely to be uncited than on-Academia articles.

Year	Off-Academia	On-Academia
2009	10.45%	2.90%
2010	11.73%	5.56%
2011	14.37%	5.58%
2012	19.18%	9.91%

Table 11: Share of uncited articles in off- and on-Academia samples, by cohort.

To model these two aspects of the data, we fit a *zero-inflated* negative binomial model.

This model assumes that an article comes from one of two populations: A "zero" population of articles that will be uncited with probability one, and a "count" population of articles whose citations will be drawn from negative binomial distributions conditioned on the articles' features.

To represent the mixture of these two distributions, we add a second stage to the negative binomial model: a model of z_i, which is equal to one when article i is from the "zero" distribution. z_i is assumed to come from a Bernoulli distribution that depends on the features of the article x_i. We can write this as:[7]

$$y_i \sim \text{NegBin}(\phi_i, \theta) \tag{3}$$

$$\phi_i = (1 - z_i)e^{x_i\beta} \tag{4}$$

$$(1 - z_i) \sim \text{Bern}\left[\text{logit}^{-1}(\mathbf{x}_i\gamma)\right]. \tag{5}$$

When $z_i = 1$, then $\phi_i = 0$, and the negative binomial distribution $\text{NegBin}(0, \theta)$ is degenerate at zero, and article i will have zero citations with probability one. Fitting the model estimates the γ and β coeffients. These are shown in Table 12.

The coefficients in the "count" model are consistent with the linear and negative binomial regression coefficients in Table 10. In the "zero" model, though, we observe a large negative coeffient on the on-Academia dummy. This indicates, consistent with Table 11, that being posted on-Academia makes is associated with a much lower likelihood of being uncited. The on-Academia coefficient in the "count" model is smaller than the same coefficient in the Negative Binomial model. This implies that, compared only with off-Academia articles that have some positive probability of being cited at all, the on-Academia effect is somewhat smaller.[8]

	Count	Zero
(Intercept)	1.956	-5.763
	(126.613)	(-15.381)
On-Academia	0.371	-10.133
	(21.017)	(-2.154)
Impact factor (log, centered)	1.136	-5.578

[7]There are different ways to represent negative binomial distributions. This the mixture-of-Poissons representation.

[8]Though the lack of Division $\times$ on-Academia interactions in the zero-inflated model makes a direct comparisons difficult.

	Count	Zero
	(86.182)	(-17.884)
Article age (centered)	0.762	-0.804
	(18.888)	(-1.929)
Article age squared (centered)	-0.064	0.069
	(-10.702)	(1.100)
Available online	0.178	-0.353
	(12.996)	(-2.218)
On-Academia × Impact factor	-0.064	-7.070
	(-2.766)	(-1.688)
On-Academia × Age	-0.008	-1.092
	(-0.081)	(-0.394)
On-Academia × Age Squared	0.014	0.164
	(0.924)	(0.395)
On-Academia × Available online	-0.053	0.532
	(-2.151)	(0.718)
Log(theta)	0.114	
	(13.280)	
N	41,891	
Log-likelihood	-130,908.395	
AIC	261,950.790	

Table 12: Coefficients from ZINB model.

Predicted Citation Advantages

Table 13 shows the predicted number of citations from the models above based on different values of the covariates. We predict citations for articles that:

1. Are in journals with impact factors at the 10th, 50th, or 90th percentiles of the sample;

2. Are one to five years old;
3. Are available online somewhere besides Academia.edu or are not;
4. Are either posted to Academia.edu or are not; and
5. Have values for the Division variables set to their sample means, i.e., the proportion of articles in the sample labelled with that Division.

The models give similar results though the linear model tends to predict the lowest number of citations for any combination of covariates. Taking a three-year-old article published in a median impact factor journal as an example, the linear model predicts 4.54 citations for such articles not available on-Academia or elsewhere online, and 7.17 citations for such an article available only on Academia.edu—a difference of 2.63 citations or 58%. For a five-year-old article in a median impact factor journal, the linear model predicts 7.05 citations for a paper not available on Academia.edu or elsewhere online, and 12.87 citations for a paper available on Academia.edu—a difference of 5.82 citations, or 83%.

For articles available online elsewhere, but not on Academia.edu, the predicted number of citations after three years is 5.40. For articles available on Academia.edu and elsewhere online, the predicted number of citations is 8.14—a difference of 2.64 citations, or 51%. This number rises to 75% after five years (8.30 citations for articles available online elsewhere vs 14.52 for articles available on Academia.edu and elsewhere online).

Table 14 calculates the percentage increase in predicted citations, compared to an article not posted on Academia.edu and not available elsewhere online. If we measure the Academia.edu citation advantage as the percentage difference in citations to articles posted to on-Academia but not elsewhere online, then we find a range of estimates for the advantage depending on the age and impact factor, with the linear model predicting 37% in the first year to 83% in the fifth year for articles published in median impact factor journals. Consistent with the coefficients on the interaction term, the table shows that the advantage decreases for higher impact factor journals, which expect more citations just from being published. For example, we find that the Academia.edu citation advantage for a paper published in a high impact factor journal is 25% in the first year, rising to 49% in the third year, and 73% in the fifth.

The second row of each model/impact-factor panel in Table 14 gives an advantage estimate for article available online but not on Academia.edu. These are estimates of the general Open Access advantage in our data, and are about 20% for three year-old articles.

Citation Advantages by Division

In Table 15, we predict the citation advantage for three year old articles published in the median impact factor journal within each Division. The advantage estimates range from 50% to 150%, with the largest estimates coming from Divisions with lower median impact factors.

Model	IF Pctile	On-Academia	Online	1 Year	2 Years	3 Years	4 Years	5 Years
Linear	10th	N	N	0.27	0.93	1.68	2.39	2.90
			Y	0.47	1.23	2.10	2.91	3.50
		Y	N	0.60	1.67	3.04	4.53	5.86
			Y	0.79	1.99	3.52	5.19	6.67
	50th	N	N	1.63	3.00	4.54	6.00	7.05
			Y	2.04	3.62	5.40	7.08	8.30
		Y	N	2.23	4.40	7.17	10.18	12.87
			Y	2.62	5.04	8.14	11.51	14.52
	90th	N	N	4.34	7.11	10.24	13.20	15.35
			Y	5.17	8.37	11.99	15.40	17.88
		Y	N	5.43	9.73	15.23	21.23	26.57
			Y	6.19	11.01	17.16	23.87	29.85
NB	10th	N	N	0.94	1.71	2.71	3.75	4.52
			Y	1.13	2.05	3.24	4.48	5.41
		Y	N	1.50	2.74	4.54	6.78	9.15
			Y	1.71	3.13	5.17	7.73	10.44
	50th	N	N	2.46	4.46	7.06	9.76	11.79
			Y	2.94	5.33	8.44	11.68	14.11
		Y	N	3.46	6.34	10.49	15.67	21.15
			Y	3.94	7.23	11.96	17.87	24.11
	90th	N	N	6.26	11.34	17.96	24.83	30.00
			Y	7.49	13.57	21.49	29.72	35.90
		Y	N	7.83	14.35	23.74	35.48	47.89
			Y	8.93	16.36	27.07	40.45	54.59

Model	IF Pctile	On-Academia	Online	1 Year	2 Years	3 Years	4 Years	5 Years
ZINB	10th	N	N	0.81	1.63	2.72	3.86	4.71
			Y	1.05	2.06	3.38	4.74	5.77
		Y	N	1.52	2.79	4.62	6.90	9.32
			Y	1.72	3.16	5.23	7.82	10.56
	50th	N	N	2.66	4.72	7.35	10.07	12.14
			Y	3.19	5.65	8.79	12.05	14.51
		Y	N	3.46	6.33	10.47	15.65	21.14
			Y	3.93	7.18	11.86	17.73	23.96
	90th	N	N	6.23	11.02	17.15	23.48	28.28
			Y	7.45	13.17	20.50	28.07	33.81
		Y	N	7.69	14.06	23.24	34.73	46.93
			Y	8.71	15.93	26.34	39.36	53.19

Table 13: Predicted citations. Impact factor percentiles are based on the entire sample of articles. The Division variables are set to their sample means, which correspond to the share of articles labelled with that Division.

Model	IF Pctile	On-Academia	Online	1 Year	2 Years	3 Years	4 Years	5 Years
Linear	10th	N	N	-	-	-	-	-
			Y	0.73	0.32	0.25	0.22	0.21
		Y	N	1.20	0.79	0.81	0.90	1.02
			Y	1.90	1.13	1.09	1.17	1.30
	50th	N	N	-	-	-	-	-
			Y	0.25	0.21	0.19	0.18	0.18
		Y	N	0.37	0.47	0.58	0.70	0.82
			Y	0.61	0.68	0.79	0.92	1.06
	90th	N	N	-	-	-	-	-
			Y	0.19	0.18	0.17	0.17	0.17
		Y	N	0.25	0.37	0.49	0.61	0.73
			Y	0.43	0.55	0.68	0.81	0.94

Model	IF Pctile	On-Academia	Online	1 Year	2 Years	3 Years	4 Years	5 Years
NB	10th	N	N	-	-	-	-	-
			Y	0.20	0.20	0.20	0.20	0.20
		Y	N	0.59	0.60	0.68	0.81	1.02
			Y	0.81	0.83	0.91	1.06	1.31
	50th	N	N	-	-	-	-	-
			Y	0.20	0.20	0.20	0.20	0.20
		Y	N	0.41	0.42	0.49	0.61	0.79
			Y	0.60	0.62	0.69	0.83	1.05
	90th	N	N	-	-	-	-	-
			Y	0.20	0.20	0.20	0.20	0.20
		Y	N	0.25	0.27	0.32	0.43	0.60
			Y	0.43	0.44	0.51	0.63	0.82
ZINB	10th	N	N	-	-	-	-	-
			Y	0.31	0.26	0.24	0.23	0.22
		Y	N	0.88	0.71	0.70	0.79	0.98
			Y	1.13	0.93	0.92	1.03	1.24
	50th	N	N	-	-	-	-	-
			Y	0.20	0.20	0.20	0.20	0.20
		Y	N	0.30	0.34	0.42	0.55	0.74
			Y	0.47	0.52	0.61	0.76	0.97
	90th	N	N	-	-	-	-	-
			Y	0.20	0.20	0.20	0.20	0.20
		Y	N	0.23	0.28	0.36	0.48	0.66
			Y	0.40	0.45	0.54	0.68	0.88

Table 14: Predicted citation advantages relative to paid-access articles, from Table 13.

Division	Med. IF	% On	% Off	Cites Off	Cites On	% Adv.
History and Archaeology	0.67	5.7%	0.7%	1.98	5.45	175%

Division	Med. IF	% On	% Off	Cites Off	Cites On	% Adv.
Education	1.02	4.3%	1.5%	4.05	10.07	149%
Creative Arts and Writing	0.34	2.1%	0.4%	2.24	5.57	148%
Physical Sciences	2.40	7.9%	6.5%	6.51	14.54	123%
Language, Communication and Culture	0.48	6.1%	1.2%	2.76	6.00	118%
Commerce, Management, Tourism and Services	1.26	4.3%	2.5%	5.90	12.43	111%
Law and Legal Studies	0.73	1.1%	0.4%	3.16	6.51	106%
Information and Computing Sciences	1.93	5.1%	4.6%	6.19	12.59	103%
Psychology and Cognitive Sciences	2.40	15.1%	5.8%	7.47	15.12	102%
Studies in Human Society	1.01	12.5%	2.8%	4.45	8.94	101%
Earth Sciences	2.33	7.7%	2.8%	7.04	14.14	101%
Medical and Health Sciences	2.72	17.0%	39.4%	6.96	13.85	99%
Technology	2.00	1.6%	2.1%	5.53	10.91	97%
Economics	1.18	2.2%	1.7%	5.36	10.52	96%
Mathematical Sciences	1.33	4.6%	7.2%	4.90	9.45	93%
Agricultural and Veterinary Sciences	2.15	4.0%	3.2%	7.38	14.11	91%
Engineering	2.72	10.9%	22.8%	8.09	15.10	87%
Biological Sciences	2.66	16.8%	18.9%	8.01	14.30	79%
Environmental Sciences	2.55	4.6%	2.7%	7.80	13.62	75%
Chemical Sciences	3.75	6.1%	19.0%	2.83	4.93	74%
Philosophy and Religious Studies	0.52	6.0%	1.0%	10.19	17.74	74%
Built Environment and Design	1.76	2.0%	0.7%	6.17	10.72	74%
Multidisciplinary	3.68	12.0%	3.5%	10.89	17.57	61%

Table 15: Predicted citations and on-Academia citation advantages by Division for five year old articles. Citations are predicted from the "Linear" model in table 10, and are calculated for five year old articles from journals with the median impact factor of the Division. Articles are assumed to have a single Division.

Issues and Topics for Further Research

Our results raise several questions that warrant further research. Primarily, our data don't allow us to conclude why papers posted to Academia.edu might receive more citations. We observed that the Academia.edu citation advantage is distinct from a general open access advantage; even amongst papers posted online elsewhere, those that are also posted on Academia.edu receive more citations. One hypothesis is that Academia.edu goes to various lengths to expose posted paper to other users. Academia.edu users are actively notified about papers posted by users they follow and in research topics they follow. This may provide more articles with more exposure than they otherwise would have had, which may lead to more citations. More data and further work would be necessary to measure how articles are discovered on the site and whether more exposure leads to further citations.

Another line of study relates to the dynamics of citations. In this study, we have looked at citation counts at a fixed moment in time. Other studies, notably Schwarz and Kennicutt Jr. (2004), have looked at the accumulation of citations over time. Having longitudinal data on citations would help us answer several questions. For articles uploaded to Academia.edu after they were published—which we exclude from this study—we could test for a change in the rate of citations received after uploading. For articles posted at the same time they're published—which we did study here—we could analyse to what extent there are feedback effects. Is the relatively large citation advantage a result of being more likely to recieve the first one or two citations from posting to the site?

Beyond Academia.edu, our work raises questions about how characteristics of venues matter for open access citations. To our knowledge there has been no research on what features of open access repositories or databases make articles easier to discover, and to what extent that leads to increased citations.

Conclusions

We have analyzed the effect of open access on citations using a novel venue for free-to-access articles, Academia.edu. Using a matching analysis and regression models with covariates to control for potential sources of selection bias, we find a substantial increase in citations associated with posting an article to Academia.edu. We find that a typical article posted to Academia.edu has 83% more citations than a similar paid-access article, not available elsewhere online, after five years. We find that a typical article posted to Academia.edu has 75% more citations than an article that is available elsewhere online through a non-Academia.edu venue: a personal homepage, departmental homepage, journal site, or any other online hosting venue.

While the true effect of open access on citations remains debated in the literature, the effect we find here suggests that features that improve the discoverability, such as the feeds and notifications used on Academia.edu, may be important factors in determining how much open access increases citations. We believe more research along these lines would help improve our understanding of the causal mechanisms behind the open access citation advantage, help researchers make better decisions about how to provide access to their research, and help journals and institutions make their open access policies more effective.

References

Academia.edu. 2012. "User Spotlight: Richard Kahn 'Academia.edu Has Increased Citations of My Work by over 30%'." http://blog.academia.edu/post/25110440121/user-spotlight-richard-kahn-academia-edu-has.

———. 2013. "Rags to Riches, PhD Style: Spotlight on Pramod Kumar, Indian Institute of Science Education & Research, Mohali." http://blog.academia.edu/post/49368089549/rags-to-riches-phd-style.

Angrist, Joshua D, and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton university press.

Australian Bureau of Statistics. 2008. "Australian and New Zealand Standard Research Classification (ANZSRC)." Australian Bureau of Statistics. http://www.arc.gov.au/pdf/ANZSRC_FOR_codes.pdf.

Australian Research Council. 2012. "Excellence in Research for Australia (ERA) 2012 National Report." Australian Research Council. http://www.arc.gov.au/era.

Bornmann, Lutz, Rüdiger Mutz, Christoph Neuhaus, and Hans-Dieter Daniel. 2008. "Citation Counts for Research Evaluation: standards of Good Practice for Analyzing Bibliometric Data and Presenting and Interpreting Results." *Ethics in Science and Environmental Politics* 8 (1): 93-102.

Craig, Iain D., Andrew M. Plume, Marie E. McVeigh, James Pringle, and Mayur Amin. 2007. "Do Open Access Articles Have Greater Citation Impact?" *Journal of Informetrics* 1 (3): 239-248.

Davis, Philip M. 2011. "Open Access, Readership, Citations: a Randomized Controlled Trial of Scientific Journal Publishing." *The FASEB Journal* 25 (7): 2129-2134.

Davis, Philip M., and William H. Walters. 2011. "The Impact of Free Access to the Scientific Literature: a Review of Recent Research." *Journal of the Medical Library Association: JMLA* 99 (3): 208.

Davis, Philip M., Bruce V. Lewenstein, Daniel H. Simon, James G. Booth, and Mathew J. L. Connolly. 2008. "Open Access Publishing, Article Downloads, and Citations: randomised Controlled Trial." *BMJ* 337. doi:10.1136/bmj.a568.

Eysenbach, Gunther. 2006. "Citation Advantage of Open Access Articles." *PLoS Biology* 4 (5) (May): e157.

Gargouri, Yassine, Chawki Hajjem, Vincent Larivière, Yves Gingras, Les Carr, Tim Brody, and Stevan Harnad. 2010. "Self-Selected or Mandated, Open Access Increases Citation Impact for Higher Quality Research." *PLoS ONE* 5 (10) (October): e13636. doi:10.1371/journal.pone.0013636. http://dx.doi.org/10.1371%2Fjournal.pone.0013636.

Gaule, Patrick, and Nicolas Maystre. 2011. "Getting Cited: does Open Access Help?" *Research Policy* 40 (10): 1332-1338.

Harnad, Stevan, and Tim Brody. 2004. "Comparing the Impact of Open Access (OA) Vs. Non-OA Articles in the Same Journals." *D-Lib Magazine* 10 (6).

Kurtz, Michael J., Guenther Eichhorn, Alberto Accomazzi, Carolyn Grant, Markus Demleitner, Edwin Henneken, and Stephen S Murray. 2005. "The Effect of Use and Access on Citations." *Information Processing & Management* 41 (6): 1395-1402.

Lawrence, Steve. 2001. "Free Online Availability Substantially Increases a Paper's Impact." *Nature* 411 (6837): 521-521.

Moed, Henk F. 2007. "The Effect of 'Open Access' on Citation Impact: An Analysis of ArXiv's Condensed Matter Section." *Journal of the American Society for Information Science and Technology* 58 (13): 2047-2054.

Schwarz, Greg J., and Robert C. Kennicutt Jr. 2004. "Demographic and Citation Trends in Astrophysical Journal Papers and Preprints." *arXiv Preprint Astro-Ph/0411275*.

SCImago. 2007. "SJR: SCImago Journal and Country Rank." http://www.scimagojr.com.