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 34,940 papers, we find that a paper in a median impact factor journal uploaded to Academia.edu receives 41% more citations after one year than a similar article not available online, 50% more citations after three years, and 73% after five years. We also found that articles also posted to Academia.edu had 64% more citations than articles only 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 networking 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.

We find that a typical article posted on Academia.edu receives approximately 41% more citations compared to similar articles not available online in the first year after upload, rising to 50% after three years, and 73% after five years. We also find that a typical article posted on Academia.edu receives more citations than an article 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 receives 30% more citations than an article uploaded to a non-Academia.edu site after the first year, rising to 42% after three years, and 64% after five years.

Our study is observational, requiring us to carefully account for possible sources of selection bias. We find that the citation advantage persists even after controlling for a number of possible selection biases.

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 disciplines, 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 forth 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

[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).

unclear. The instrumental variables Gaule and Maystre (2011) use are only weakly correlated with citations; as a result their estimate of the open-access advantage is imprecise.[3]

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 underestimate 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 type of article.** Academic journals publish content besides original research or scholarship: book reviews, errata, responses to recently pub-

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

lished articles, conference abstracts, editorials, etc. These other types of content typically receive fewer citations than research articles. If Academia.edu users are less likely to post these other types of content to the site, then we might overestimate the advantage relative to an off-Academia group that contains more "non-research" content.

5. **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 used 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 database 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.

Table 1: Sample size of papers, by cohort.

Year	Off-Academia	On-Academia
2009	4,730	234
2010	5,867	937
2011	7,127	3,627
2012	8,559	3,859
Total	26,283	8,657

Articles in the sample come from 6,312 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 5.4% of the sample, and the top ten journals comprise 11% of the sample.

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

Journal	# Articles	% Total
Analytical Chemistry	1,432	4.10%
PLoS One	445	1.27%
Biological and Pharmaceutical Bulletin	331	0.95%
Analytical Methods: advancing methods and applications	316	0.90%
Analytical Biochemistry	304	0.87%
Bioconjugate Chemistry	287	0.82%
Applied Mechanics and Materials	282	0.81%
Applied Physics Letters	184	0.53%
AAPS PharmSciTech	167	0.48%
Anesthesia and Analgesia	162	0.46%

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 received 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.

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

Journal	Impact factor
Chemical Reviews	45.62
Annual Review of Immunology	43.47
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
Journal of Emergency Medicine	27.1

Journal	Impact factor
Cell	26.85

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.

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.

Division	% All	% On	% Off	Med. Imp. Factor
Medical and Health Sciences	31.7%	17.6%	36.4%	2.66
Engineering	21.7%	11.8%	25.0%	2.66
Biological Sciences	20.0%	18.1%	20.6%	2.63
Chemical Sciences	17.7%	6.4%	21.4%	3.75
Psychology and Cognitive Sciences	8.1%	15.6%	5.6%	2.50
Physical Sciences	7.5%	8.7%	7.1%	2.40

[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
Mathematical Sciences	6.9%	5.0%	7.5%	1.36
Multidisciplinary	5.9%	12.6%	3.7%	3.68
Information and Computing Sciences	4.9%	5.3%	4.8%	2.05
Studies in Human Society	4.5%	10.9%	2.4%	1.13
Earth Sciencess	4.2%	8.2%	2.9%	2.31
Agricultural and Veterinary Sciences	3.7%	4.3%	3.5%	2.15
Environmental Sciences	3.4%	5.0%	2.9%	2.48
Commerce, Management, Tourism and Services	2.9%	4.3%	2.4%	1.34
Technology	2.1%	1.8%	2.3%	2.00
Education	2.0%	4.2%	1.2%	1.06
Language, Communication and Culture	1.8%	5.1%	0.8%	0.70
History and Archaeology	1.7%	5.3%	0.5%	0.98
Philosophy and Religious Studies	1.7%	4.5%	0.8%	0.60
Economics	1.6%	2.0%	1.5%	1.23
Built Environment and Design	1.1%	2.0%	0.8%	1.83
Creative Arts and Writing	0.6%	1.5%	0.3%	0.75
Law and Legal Studies	0.4%	0.9%	0.3%	0.73

Document Types

We include in our analysis only articles with original research, analysis or scholarship, or survey articles. We exclude book reviews, editorials, errata, and other "non-research" content. Our procedure for obtaining on- and off-Academia articles provided 44,689 articles. From this sample, we removed any articles not identified to be original research.

To identify the type of each article, we used Amazon Mechanical Turk (MTurk), a crowdsourcing marketplace. Common uses of MTurk in academic research include collecting survey data, performing online experiments, and classifying data to train and validate machine learning algorithms.

We provided DOI links to articles in our sample to over 300 MTurk workers. The workers were asked to fill out an online form based on information from the abstract or full text at the DOI link. They were finally asked to classify the article as one of the following types:

1. A summary of a meeting or conference
2. An Editorial or Commentary
3. A response to a recent article in the same journal;

4. An article with original research, analysis or scholarship, or a broad survey of research on a topic
5. This is a Book Review, Software Review, or review of some other recent work or performance
6. An Erratum, Correction, or Retraction of an earlier article
7. Something else

Workers might fail to categorize an article, giving one of these reasons: the link was broken, there was no abstract or text available on the site, the article was in a foreign language, or they otherwise couldn't tell. Some workers' results were excluded if they exhibited suspicious patterns, such as giving all articles the same classification, or completing a large number of tasks in an unreasonably short time. Their tasks were then resubmitted so that each article had three independent reviews.

Each article was reviewed by three different workers. Our sample only includes articles that all three articles were "original research" (option 4). Of the original 44,689 articles, this left 37,093 "original research" articles. Relying on a majority, 2-of-3 vote to classify articles would have resulted in 42,174 "original research" articles. Unanimity is a conservative classificaiton rule, but given that false positive classification of "original research" articles could upwardly bias our result, we consider it appropriate.

Appendix A provides more detail on the MTurk classification process, including the instruction form provided to workers.

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 sample 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.

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

		Off-Academia	On-Academia
a.	Full-text available elsewhere	6,798	3,917
b.	Articles searched	26,283	8,657
c.	Share *(a ÷ b)*	25.9%	45.2%

Analysis and Results

Our general empirical strategy is to estimate the distribution of the citation count of article i, published in journal j at time 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

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

or medians,

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

One approach would be to directly estimate these summary statistics 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.

Table 6: Citations summary statistics

Sample	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
off-Academia	0	2	5	10.18	12	1237
on-Academia	0	3	7	12.65	15	721

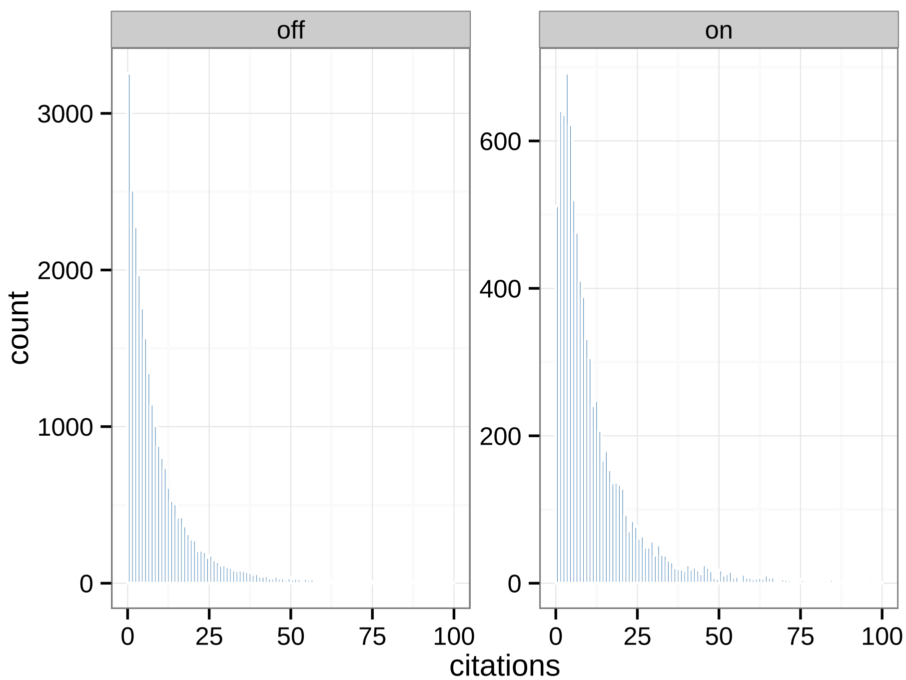

Figure 1: Distributions of citations (x-axis is truncated at 100)

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, each 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 24 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.

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

Year	Impact Factor Bin	Off-Academia	On-Academia	Abs. Diff	% Diff.
2009	[0,0.54]	1	5.0	4.0	400
	(0.54,0.982]	3	7.5	4.5	150
	(0.982,1.41]	6	14.0	8.0	133
	(1.41,1.86]	8	14.5	6.5	81
	(1.86,2.31]	9	19.0	10.0	111
	(2.31,2.76]	10	15.0	5.0	50
	(2.76,3.3]	12	22.0	10.0	83
	(3.3,3.76]	13	23.5	10.5	81
	(3.76,5.3]	17	20.0	3.0	18
	(5.3,45.6]	21	45.0	24.0	114
2010	[0,0.54]	1	3.0	2.0	200
	(0.54,0.982]	3	6.0	3.0	100
	(0.982,1.41]	4	8.0	4.0	100
	(1.41,1.86]	6	11.0	5.0	83
	(1.86,2.31]	7	11.0	4.0	57

[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.
	(2.31,2.76]	9	15.0	6.0	67
	(2.76,3.3]	11	16.0	5.0	45
	(3.3,3.76]	12	18.0	6.0	50
	(3.76,5.3]	13	23.0	10.0	77
	(5.3,45.6]	17	25.0	8.0	47
2011	[0,0.54]	0	2.0	2.0	-
	(0.54,0.982]	2	4.0	2.0	100
	(0.982,1.41]	3	7.0	4.0	133
	(1.41,1.86]	4	7.0	3.0	75
	(1.86,2.31]	5	9.0	4.0	80
	(2.31,2.76]	6	9.0	3.0	50
	(2.76,3.3]	7	10.0	3.0	43
	(3.3,3.76]	8	12.0	4.0	50
	(3.76,5.3]	9	15.0	6.0	67
	(5.3,45.6]	13	24.0	11.0	85
2012	[0,0.54]	0	1.0	1.0	-
	(0.54,0.982]	1	2.0	1.0	100
	(0.982,1.41]	2	3.0	1.0	50
	(1.41,1.86]	3	4.0	1.0	33
	(1.86,2.31]	3	5.0	2.0	67
	(2.31,2.76]	4	6.0	2.0	50
	(2.76,3.3]	4	7.0	3.0	75
	(3.3,3.76]	5	7.0	2.0	40
	(3.76,5.3]	6	9.0	3.0	50
	(5.3,45.6]	9	14.0	5.0	56

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 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 24% of the variance in citations.

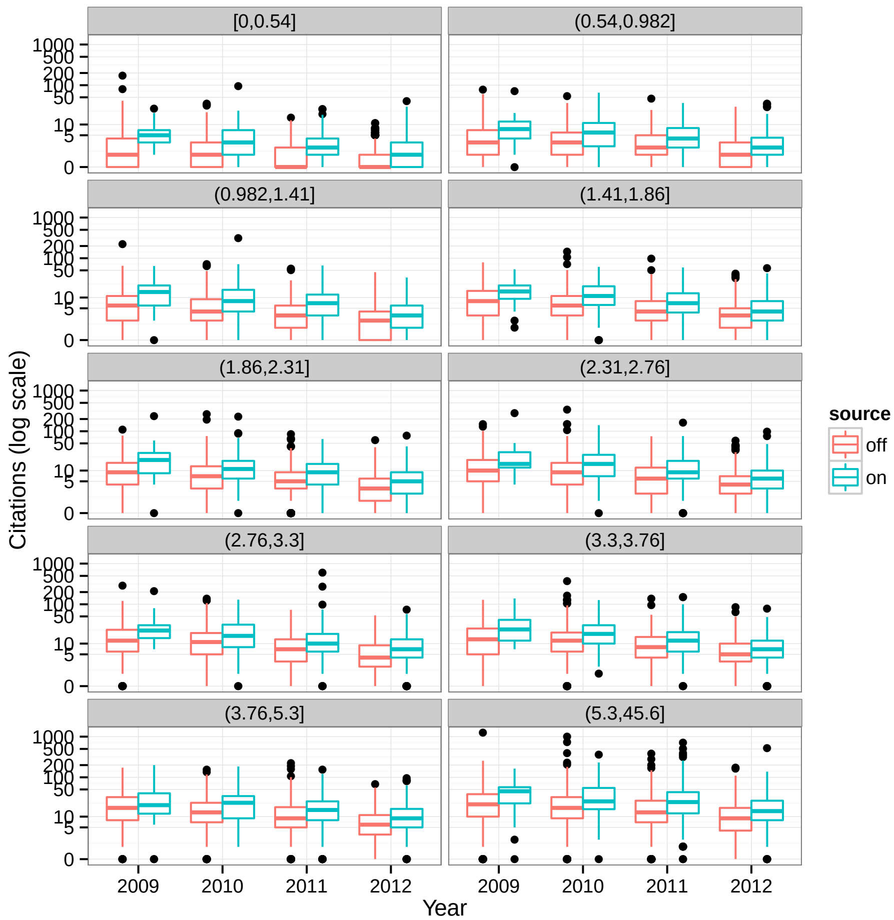

Figure 2: Boxplots of off- and on-Academia article citations, by cohort and impact factor bin.

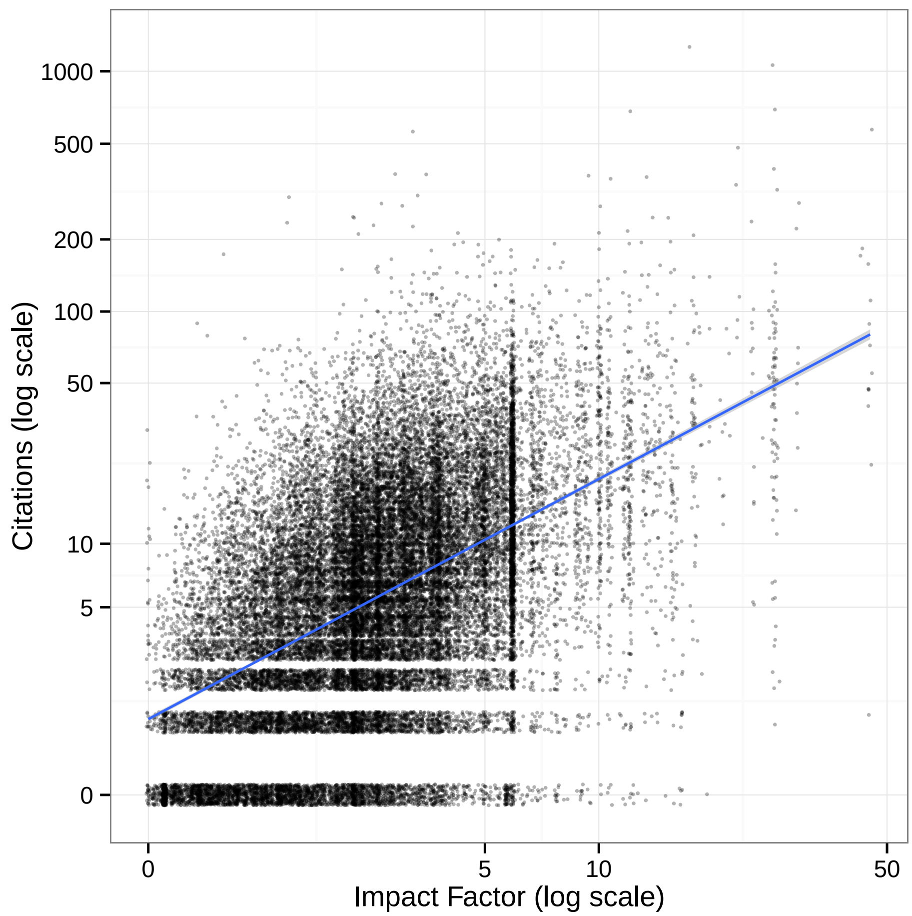

Figure 3: Article citations against Impact Factor (log scale).

	Citations (log scale)
Impact Factor (log scale)	0.956
	(0.009)
Intercept	0.723
	(0.012)
Observations	34,940
R^2	0.240

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 impact factors.

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 is 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.

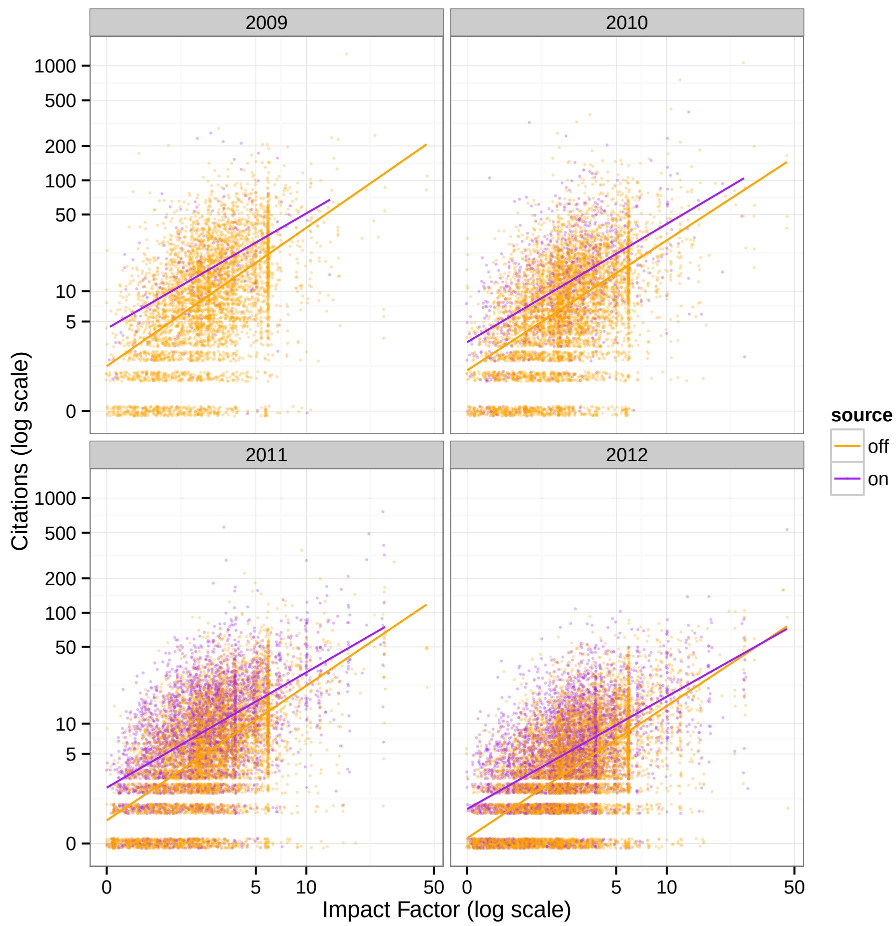

Figure 4: Citations against impact factors, with lines for conditional medians by off- and on-Academia sources.

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 article'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.

Table 9: Summary statistics of regression model covariates.

	Mean	Median	Std. Dev.
Age	3.05	2.90	1.04
Impact factor	2.96	2.34	2.81
Online	0.31	0.00	0.46

Linear Regression

We fit a 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 statistically significant 0.489. 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 69% more citations.

The coefficient on "on-Academia $\times$ Online" is -0.045, indicating that, for articles already freely available online, posting to Academia.edu is associating with a smaller dif-

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

ference in citations. At the sample average (see Table 9), the on-Academia coefficient for an online article is $0.489 - 0.045 = 0.44$. The on-Academia $\times$ Age coefficient is positive, implying that the on-Academia effect increases with time.

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.

Table 10: Regression results. Topic keyword coefficients omitted; t-statistics in parentheses.

	Linear	Neg. Binom.
(Intercept)	1.712	1.942
	(106.598)	(106.841)
On-Academia	0.489	0.529
	(14.617)	(14.211)
Impact factor (log, centered)	1.028	1.262
	(91.184)	(97.725)
Article age (centered)	0.596	0.828
	(16.216)	(19.907)

	Linear	Neg. Binom.
Article age squared (centered)	-0.048	-0.070
	(-8.825)	(-11.412)
Available online	0.127	0.155
	(10.134)	(10.965)
On-Academia × Impact factor	-0.085	-0.176
	(-3.653)	(-6.783)
On-Academia × Age	0.043	-0.122
	(0.483)	(-1.242)
On-Academia × Age Squared	0.004	0.028
	(0.294)	(1.797)
On-Academia × Available online	-0.045	-0.045
	(-1.899)	(-1.721)
N	34,940	34,940
R-squared	0.340	
Deviance	26,841.265	39,305.221
Log-likelihood	-44,971.015	-110,790.733
AIC	90,056.031	221,695.467

Zero-Inflated Negative Binomial Regression

The modal number of citations for an article in our sample is zero or one, and approximately 10.5% 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.

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

Year	Off-Academia	On-Academia
2009	7.86%	2.56%
2010	9.03%	3.52%
2011	12.38%	4.41%
2012	17.21%	8.11%

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 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 coefficients. 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 coefficient on the on-Academia dummy. This indicates, consistent with Table 11, that being posted on-Academia 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]

Table 12: Coefficients from ZINB model.

	Count	Zero
(Intercept)	2.032	-5.591
	(126.506)	(-14.682)
On-Academia	0.343	-17.403
	(18.796)	(-1.598)
Impact factor (log, centered)	1.151	-5.235
	(82.132)	(-17.329)

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

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

	Count	Zero
Article age (centered)	0.800	-0.640
	(19.020)	(-1.384)
Article age squared (centered)	-0.068	0.049
	(-10.851)	(0.700)
Available online	0.154	-0.338
	(10.813)	(-1.933)
On-Academia × Impact factor	-0.089	-12.704
	(-3.669)	(-1.496)
On-Academia × Age	-0.093	4.516
	(-0.957)	(0.424)
On-Academia × Age Squared	0.025	-0.846
	(1.631)	(-0.453)
On-Academia × Available online	-0.051	1.055
	(-2.004)	(0.709)
Log(theta)	0.231	
	(24.531)	
N	34,940	
Log-likelihood	-110,513.5	
AIC	221,161.0	

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

5.01 citations for such articles not available on-Academia or elsewhere online, and 7.51 citations for such an article available only on Academia.edu—a difference of 2 citations or 50%. For a five-year-old article in a median impact factor journal, the linear model predicts 8.12 citations for a paper not available on Academia.edu or elsewhere online, and 14.03 citations for a paper available on Academia.edu—a difference of 5.91 citations, or 73%.

For articles available online elsewhere, but not on Academia.edu, the predicted number of citations after three years is 5.83. For articles available on Academia.edu and elsewhere online, the predicted number of citations is 8.25—a difference of 2.42 citations, or 42%. This number rises to 64% after five years (9.35 citations for articles available online elsewhere vs 15.33 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 41% in the first year to 73% 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 22% in the first year, rising to 36% in the third year, and 58% 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 65% to 165%, with the largest estimates coming from Divisions with lower median impact factors.

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	0.27	0.99	1.84	2.67	3.31
			Y	0.44	1.27	2.23	3.17	3.89
		Y	N	0.70	1.82	3.28	4.94	6.56
			Y	0.85	2.06	3.65	5.46	7.21
	50th	N	N	1.69	3.22	5.01	6.77	8.12
			Y	2.06	3.79	5.83	7.83	9.35
		Y	N	2.38	4.61	7.51	10.83	14.03
			Y	2.68	5.10	8.25	11.85	15.33
	90th	N	N	4.70	7.95	11.74	15.47	18.33
			Y	5.48	9.17	13.48	17.71	20.95
		Y	N	5.74	10.18	15.96	22.56	28.95
			Y	6.32	11.14	17.43	24.60	31.54
NB	10th	N	N	1.01	1.86	2.99	4.18	5.08
			Y	1.17	2.17	3.50	4.89	5.93
		Y	N	1.65	2.94	4.81	7.22	9.95
			Y	1.84	3.28	5.37	8.06	11.10
	50th	N	N	2.52	4.68	7.52	10.51	12.75
			Y	2.95	5.46	8.78	12.27	14.89
		Y	N	3.64	6.49	10.62	15.94	21.97
			Y	4.07	7.25	11.85	17.80	24.53
	90th	N	N	6.35	11.77	18.93	26.45	32.10
			Y	7.42	13.74	22.11	30.89	37.48
		Y	N	8.06	14.36	23.49	35.27	48.61
			Y	9.00	16.04	26.23	39.38	54.28
ZINB	10th	N	N	0.91	1.81	3.01	4.29	5.27
			Y	1.13	2.20	3.63	5.13	6.27
		Y	N	1.69	3.02	4.94	7.42	10.23
			Y	1.88	3.34	5.47	8.22	11.33
	50th	N	N	2.70	4.92	7.81	10.83	13.10
			Y	3.16	5.74	9.12	12.63	15.29
		Y	N	3.68	6.55	10.72	16.11	22.20
			Y	4.07	7.26	11.88	17.85	24.61
	90th	N	N	6.31	11.46	18.17	25.17	30.45
			Y	7.36	13.36	21.19	29.35	35.51
		Y	N	7.99	14.25	23.32	35.02	48.29

Model	IF Pctile	On-Academia	Online	1 Year	2 Years	3 Years	4 Years	5 Years
			Y	8.86	15.79	25.84	38.81	53.51

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

Model	IF Pctile	On-Academia	Online	1 Year	2 Years	3 Years	4 Years	5 Years
Linear	10th	N	N	–	–	–	–	–
			Y	0.64	0.27	0.21	0.19	0.18
		Y	N	1.58	0.83	0.78	0.85	0.98
			Y	2.12	1.08	0.98	1.04	1.18
	50th	N	N	–	–	–	–	–
			Y	0.22	0.18	0.16	0.16	0.15
		Y	N	0.41	0.43	0.50	0.60	0.73
			Y	0.58	0.58	0.65	0.75	0.89
	90th	N	N	–	–	–	–	–
			Y	0.16	0.15	0.15	0.14	0.14
		Y	N	0.22	0.28	0.36	0.46	0.58
			Y	0.34	0.40	0.48	0.59	0.72
NB	10th	N	N	–	–	–	–	–
			Y	0.17	0.17	0.17	0.17	0.17
		Y	N	0.64	0.58	0.60	0.72	0.96
			Y	0.83	0.76	0.79	0.93	1.19
	50th	N	N	–	–	–	–	–
			Y	0.17	0.17	0.17	0.17	0.17
		Y	N	0.44	0.39	0.41	0.52	0.72
			Y	0.61	0.55	0.58	0.69	0.92
	90th	N	N	–	–	–	–	–
			Y	0.17	0.17	0.17	0.17	0.17
		Y	N	0.27	0.22	0.24	0.33	0.51
			Y	0.42	0.36	0.39	0.49	0.69
ZINB	10th	N	N	–	–	–	–	–
			Y	0.25	0.22	0.20	0.19	0.19
		Y	N	0.87	0.67	0.64	0.73	0.94
			Y	1.07	0.85	0.82	0.91	1.15
	50th	N	N	–	–	–	–	–
			Y	0.17	0.17	0.17	0.17	0.17
		Y	N	0.36	0.33	0.37	0.49	0.69
			Y	0.51	0.48	0.52	0.65	0.88

Model	IF Pctile	On-Academia	Online	1 Year	2 Years	3 Years	4 Years	5 Years
	90th	N	N	-	-	-	-	-
			Y	0.17	0.17	0.17	0.17	0.17
		Y	N	0.27	0.24	0.28	0.39	0.59
			Y	0.40	0.38	0.42	0.54	0.76

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.

Division	Med. IF	% Off	% On	Cites Off	Cites On	Diff.	% Adv.
History and Archaeology	0.98	0.5%	5.3%	2.72	7.27	4.54	167%
Education	1.06	1.2%	4.2%	5.20	11.36	6.16	119%
Creative Arts and Writing	0.75	0.3%	1.5%	4.00	8.36	4.37	109%
Physical Sciences	2.40	7.1%	8.7%	7.35	15.19	7.84	107%
Language, Comm'n and Culture	0.70	0.8%	5.1%	3.78	7.76	3.98	105%
Commerce, Mgt., Tourism and Svcs.	1.34	2.4%	4.3%	6.71	13.43	6.73	100%
Law and Legal Studies	2.05	4.8%	5.3%	7.21	14.16	6.94	96%
Information and Comp. Sci.	1.36	7.5%	5.0%	5.43	10.57	5.14	95%
Psychology and Cognitive Sci.	2.31	2.9%	8.2%	7.84	15.22	7.38	94%
Studies in Human Society	2.50	5.6%	15.6%	8.82	17.02	8.21	93%
Earth Sciences	0.73	0.3%	0.9%	3.78	7.28	3.50	93%
Medical and Health Sci.	2.15	3.5%	4.3%	8.05	15.42	7.36	91%
Technology	2.00	2.3%	1.8%	6.41	11.95	5.54	86%
Economics	1.23	1.5%	2.0%	6.45	11.67	5.22	81%
Mathematical Sci.	2.66	25.0%	11.8%	8.82	15.79	6.96	79%
Agricultural and Vet. Sci.	2.66	36.4%	17.6%	8.46	15.07	6.61	78%
Engineering	1.13	2.4%	10.9%	5.75	9.99	4.24	74%
Biological Sciences	3.75	21.4%	6.4%	11.17	19.39	8.22	74%
Environmental Sciences	1.83	0.8%	2.0%	6.97	12.08	5.11	73%
Chemical Sciences	2.48	2.9%	5.0%	8.56	14.73	6.18	72%
Philosophy and Relig. Studies	2.63	20.6%	18.1%	8.78	15.07	6.28	72%
Built Environment and Design	0.60	0.8%	4.5%	3.57	6.04	2.46	69%
Multidisciplinary	3.68	3.7%	12.6%	11.85	19.29	7.44	63%

Issues and Topics for Further Research

Our results raise several questions that warrant further research. One area to consider is which properties of Academia.edu are responsible for papers receiving 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. Further work could be done to measure the effect of these distributional properties.

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 receive 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 73% more citations than a similar paid-access article, not available elsewhere online, after five years. We find that a typical article that is also posted to Academia.edu has 64% more citations than one that is only 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): 239–48.

Davis, Philip M. 2011. "Open Access, Readership, Citations: A Randomized Controlled Trial of Scientific Journal Publishing." *The FASEB Journal* 25 (7). FASEB: 2129–34.

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). Medical Library Association: 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. BMJ Publishing Group Ltd. doi:10.1136/bmj.a568.

Eysenbach, Gunther. 2006. "Citation Advantage of Open Access Articles." *PLoS Biology* 4 (5). Public Library of Science: 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). Public Library of Science: e13636. doi:10.1371/journal.pone.0013636.

Gaule, Patrick, and Nicolas Maystre. 2011. "Getting Cited: Does Open Access Help?" *Research Policy* 40 (10). Elsevier: 1332–38.

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). Elsevier: 1395–1402.

Lawrence, Steve. 2001. "Free Online Availability Substantially Increases a Paper's Impact." *Nature* 411 (6837). Nature Publishing Group: 521–21.

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). Wiley Online Library: 2047–54.

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.

Appendix A

Document Type Classification with Amazon Mechanical Turk

We relied on Amazon Mechanical Turk to classify documents by type. Mechanical Turk (MTurk) is an online crowd-sourcing marketplace. *Requesters* submit tasks (called "Human Intelligence Tasks," or *HITs*) to *workers* who complete them for a fee. Each of the 44,689 documents in our sample were submitted to three separate workers, for a total of 134,067 HITs. These were completed by approximately 300 workers. Workers were paid $0.10 per HIT.

Workers were provided with a DOI link leading to a journal web page for the document. To complete the HIT, workers were asked to skim the abstract or full text of the article if it was available at the link, and answer a series of short questions. Figs. 5-7 show the instructions provided to workers.

During the process, some workers' results were removed and their HITs resubmitted. This occurred if workers returns suspicious results: for example if they completed too many HITs too quickly, or returned a large amount of identical results.[9]

Classification Rules

We consider two rules for classifying documents: "majority vote" and "unanimous vote." Under the majority vote rule, a document is classified if at least two of three workers select the same document type. Under the unanimous vote rule, a document is classified only if all three workers selected the same document type. Under each rule, there will be some share of documents left unclassified: e.g., for the majority vote rule documents where all three workers selected different types will be left unclassified.

Table 16 shows the distribution of document types classified under each rule. The "No Agreement" row shows the share of unclassified documents. Under unanimous rule, the share of unclassified documents increases dramatically, and the share of each classified document type falls.

[9]Identifying suspicious worker behavior is complicated. Workers can preview tasks before accepting them and requesters cannot observe this previewing behavior. This has two implications. First, the tasks they accept may not be a random sample of all tasks, so long runs of similar results may not indicate inaccurate work, but just "cherry picking" of similar-looking HITs. Second, the amount of time they're observed to have worked on a task (measured as submit time minus accept time) may be underestimated.

Updated Instructions (Click to Expand/Collapse)

- An issue of an academic journal contains different kinds of documents. A list of correct categorizations is in the "Examples" expandable panel below
- The purpose of this task is to help us categorize academic documents correctly.
- We are providing you with the title of an academic document, and a link to that document's webpage.
- Please categorize the document by skimming the abstract, extract or full text.
- Most of the time just skimming the abstract is enough. You only need to skim the full text if it was available, and the abstract was not, or if you couldn't figure out the categorization from the abstract.
- Sometimes the link will go to a "gateway" page where you will be presented with a few publication databases to choose from. Like this one (http://doi.org/10.1213/ANE.0B013E318263C924). Follow the first link from the left or the top. Sometimes the left or the top will require a registration (the example we just linked to does exactly that), in which case try the next one.
- Be especially careful if something calls itself a "review". The word "review" is used in two separate senses in academic documents:
  - One sense of the word "review" is when it describes a Book Review or a Performance Review or other types of reviews of a single, individual work by someone else. In this case, choose option (H).
  - Another sense of the word "review" is a synonym for a "broad survey of research on a topic". This happens when someone wants to tell the readers about recent developments in the entire academic field of study. Like this one (http://doi.org/10.1586/ERS.09.16). In this case, choose option (G), *but tell us in the comment field that this was a review or a survey*
- Except for the first HIT, when you have to read these instructions and look at the examples below, these HITs shouldn't take more than 1 minute on average, and are priced accordingly; if you plan to do a few hundred of these, it could be fairly worthwhile. Some of them will be obvious and will only require 20 seconds. Some will require a little bit more thought and will take 2 minutes. It is OK to select option (C) "I can't tell what this is" if you are really unsure and are beginning to spend a lot of time on any individual item.

Examples (Click to Expand/Collapse)

Here are some examples of correct categorization of academic documents

- A summary of a meeting or conference.
  1. Example (http://doi.org/10.2217/FNL.11.61) **2.** Example (http://doi.org/10.1111/J.1444-0938.2008.00339.X) **3.** Example (http://doi.org/10.1586/17512433.2.2.163)
- An Editorial or Commentary.
  1. Example of Commentary (http://www.sciencedirect.com/science/article/pii/S1053810011000079) **2.** Example of Commentary (http://ehp.niehs.nih.gov/120-a106/) **3.** Example of Editorial (http://doi.org/10.1021/AC200805H) **4.** Example of Editorial (http://doi.org/10.1002/BIES.201290005)
- A response to a recent article in the same journal.
  1. Example (http://doi.org/10.1097/ALN.0B013E3182134F68) **2.** Example (http://doi.org

Figure 5: Mechanical Turk instruction page, 1 of 3.

/10.1378/CHEST.10-3195)

- An article with original research, analysis or scholarship, or a broad survey of research on a topic.
    1. Example of original research (http://pubs.acs.org/doi/abs/10.1021/ie2013086) **2.** Example of original research (http://link.springer.com/article/10.1007%2Fs10578-012-0283-4) **3.** Example of analysis (http://doi.org/10.1111/J.1468-0432.2009.00503.X) **4.** Example of a broad survey of research on a topic (http://doi.org/10.4304/JMM.7.3.223-230)
- A Book Review, Software Review, or review of some other recent work or performance.
    1. Example of a book review (http://www.tandfonline.com/doi/abs/10.1080/03643100903173057) **2.** Example of a book review (http://www.tandfonline.com/doi/abs/10.1111/j.1469-5812.2009.00521.x)
- An Erratum, Correction, or Retraction of an earlier article.
    1. Example of an Erratum (http://scitation.aip.org/content/aip/journal/jcp/136/14/10.1063/1.3693966) **2.** Example of a Correction (http://iopscience.iop.org/0953-8984/23/31/319501/) **3.** Example of a Retraction (http://scripts.iucr.org/cgi-bin/paper?S1600536809049976)

Document Title and Link

**${title}** (${doi_url})

1. Please skim the page, abstract, extract or full text. What type of document this is?

- (A) I can't categorize this document because the link is broken or wrong.
- (B) I can't categorize this document even though the link works fine because a payment- or registration- free version of an abstract, extract or full-text in English is not available.
- (C) I can't tell what this is, even though a registration- and payment-free abstract, extract or full-text is available in English, and I skimmed them.
- (D) This is a summary of a meeting or conference.
- (E) This is an Editorial or Commentary.
- (F) This is a response to a recent article in the same journal.
- (G) This is an article with original research, analysis or scholarship, or a broad survey of research on a topic.
- (H) This is a Book Review, Software Review, or review of some other recent work or performance.
- (I) This is an Erratum, Correction, or Retraction of an earlier article.
- (J) This is something else. *(Make sure to answer question 2 if you choose this.)*

Figure 6: Mechanical Turk instruction page, 2 of 3.

2. If none of the categories above seem correct, what would you label this document?

(If you chose (J) "this is something else" in question 1, you must answer this question. If you chose something else question 1, then question 2 is optional: you may leave it blank, or you may answer it to help us categorize documents better.)

3. What information did you use to categorize this document?

- (A) I couldn't categorize the document and chose options (A), (B) or (C) above. *(Please only choose this option if you really did pick (A), (B) or (C) in question 1. We will have to throw the categorization away if you were able to categorize the document but didn't tell us how.)*
- (B) I found and skimmed the abstract or extract.
- (C) I found and skimmed the full text.
- (D) I did something else. *(Make sure to answer question 4 if you choose this.)*

4. If you chose (D) "I did something else" in question 3: What information did you use to categorize this document?

Submit

5. How could this HIT be improved so that you can do it quicker?

(This question is optional. Please only answer this question after you've done a few of our HITs, and only answer it once.)

6. What is a fair price for this HIT?

(This question is optional. Please only answer this question after you've done a few of our HITs, and only answer it once.)

Figure 7: Mechanical Turk instruction page, 3 of 3.

Table 16: Distribution of document type classifications amongst total sample of articles (N=44,689). Majority classification is defined as at least two of three Mechanical Turk workers agreeing on a classification; unanimous classification as all three.

Document Type	Majority	Unanimous
Original Research	94.4%	83.0%
No Agreement	1.5%	16.0%
No English Abstract or Full Text	1.4%	0.3%
Broken DOI Link	1.1%	0.5%
Editorial/Commentary	0.9%	0.1%
Response or Comment	0.3%	0.0%
Book Review, etc.	0.2%	0.0%
Errata/Correction	0.1%	0.0%
Other	0.1%	0.0%
Conference Summary	0.0%	0.0%
Unknown	0.0%	0.0%

In the final sample, we only include documents that were unanimously classified as "Original Research." Table 17 shows the share of documents classified as "Original Research" in the off- and on-Academia sets under each rule (with unclassified documents included in the total). There are slightly fewer original research documents in the on-Academia set under both rules.

Table 17: Share of documents unanimously classified as "Original Research",in off- and on-Academia samples.

Sample	Majority	Unanimous
off-Academia	93.7%	82.4%
on-Academia	96.3%	84.8%

Fig. 8 shows the distribution over workers of the rate of non-"Original Research" documents, by sample. Workers typically found more non-"Original Research" documents amongst the off-Academia documents they reviewed.

Agreement with a Gold Standard Set

To assess the quality of MTurk worker classifications, we compared them to a "Gold Standard" set of 100 documents classified by in-house researchers. Each documents

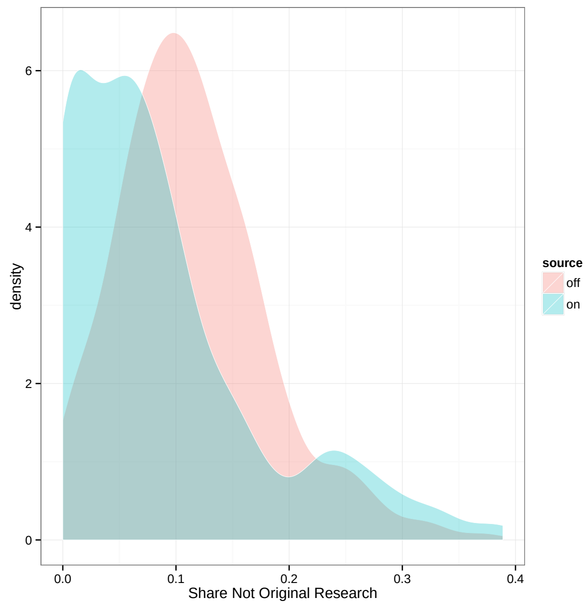

Figure 8: Share of non-"Original Research" documents, distribution over Mechanical Turk Workers.

was classified by two independent researchers, with disagreements settled by a third.

The documents in the Gold Standard set were not randomly selected. They over-represent documents in the on-Academia sample (46% in the Gold Standard vs. 24% in the total). Documents were also selected to include documents that appeared difficult to classify based on their titles or subject matter. For example, to see if workers could distinguish a Book Review of a recent title from a survey or "review" article of a research topic, or a scholarly or literary "review" of older work. Table 18 show that about 30% of documents in the Gold Standard set were classified as something other than "Original Research," which we expect overestimates the share in the population.

Table 18: Distribution of document types in the Gold Standard set. Documents were independently classified by two in-house researchers. Disagreements were settled by a third.

Document Type	Num. Documents	% Documents
Original Research	71	71%
Editorial/Commentary	5	5%
No English Abstract or Full Text	5	5%
Other	5	5%
Book Review, etc.	4	4%

Document Type	Num. Documents	% Documents
Response or Comment	4	4%
Errata/Correction	3	3%
Conference Summary	2	2%
Unknown	1	1%

Table 19 is a confusion matrix comparing the agreement of in-house and MTurk classifications of the Gold Standard set. The classifications are grouped into "Original Research" and "Other" categories. Unclassified documents (where workers did not agree on a type) are included in the "Other" category.

Table 19: Confusion matrices for in-house vs. Mechanical Turk classifications of the Gold Standard set.

In-house	Mechanical Turk	Majority	Unanimous
Other	Other	19	22
Other	Original Research	10	7
Original Research	Other	2	20
Original Research	Original Research	69	51

Table 20 provides various agreement statistics based on the confusion matrix. Because many documents were included in the Gold Standard set because of their perceived difficulty to classify, we expect these statistics to be worse than those for a random sample. The accuracy and precision of the MTurk worker classifications are high. Using a majority-vote classification rule, the false positive rate—the share of documents classified in-house as "Other," but as "Original Research" by MTurk workers—is about 34%. The false negative rate—the share of documents classified in-house as "Original Research," but as "Other" by MTurk workers—is only about 3%. Using a unanimous-vote classification rule, the false positive rate falls to 24%, while the false negative rate rises to 28%. A high rate of false positives could potentially cause an upward bias in an Academia citation advantage estimate. Therefore, despite the fact that the unanimous-vote rule has worse overall agreement statistics with the Gold Standard set, we rely on it to filter our sample, since it minimizes the false positive rate.

Table 20: Agreement statistics for in-house vs. Mechanical Turk classifications of the Gold Standard set.

Statistic	Majority	Unanimous
Accuracy	0.88	0.73
Precision	0.87	0.88

Statistic	Majority	Unanimous
Recall	0.97	0.72
F1 Score	.92	0.79
False Positive Rate	0.34	0.24
False Negative Rate	0.03	0.28