Earlier Web Usage Statistics as Predictors of Later Citation Impact

Tim Brody, Stevan Harnad, and Leslie Carr
Intelligence, Agents, Multimedia Group, School of Electronics and Computer Science, University of Southampton, SO17 1BJ, United Kingdom. E-mail: {tdb01r, harnad, lac} @ecs.soton.ac.uk

The use of citation counts to assess the impact of research articles is well established. However, the citation impact of an article can only be measured several years after it has been published. As research articles are increasingly accessed through the Web, the number of times an article is downloaded can be instantly recorded and counted. One would expect the number of times an article is read to be related both to the number of times it is cited and to how old the article is. The authors analyze how short-term Web usage impact predicts medium-term citation impact. The physics e-print archive—arXiv.org—is used to test this.

Introduction

Peer-reviewed journal article (or refereed conference article) publication is the primary mode of communication and record for scientific research. Researchers—as authors—write articles that report experimental results, theories, reviews, and so on. To relate their findings to previous findings, authors cite other articles. Authors cite an article if they (a) know of the article, (b) believe it to be relevant to their own article, and (c) believe it to be important enough to cite explicitly (i.e., there is both a relevance and an importance judgment inherent in choosing what to cite). It is probably safe to assume that the majority of citations will be positive, but even negative citations (where an author cites an article only to say it is wrong or to disagree with it) will refer to articles that the author judges relevant and important enough to warrant rebuttal (Borgman and Furner, 2002 provide a review of many studies that debate the motivations for and influences on citing). Citations can therefore be used as one measure of the importance and influence of articles; they indirectly reflect the importance of the journals they are published in and the authors that wrote them as well. The total number of times an article is cited is called its *citation impact*.

Received April 4, 2005; revised April 28, 2005; accepted May 12, 2005

 ● Published online 25 April 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20373

The time that it takes—from the moment an article is accepted for publication (after peer review) until it is (a) published, (b) read by other authors, (c) cited by other authors in their own articles, and then (d) those citing articles are themselves peer-reviewed, revised, and published—may range anywhere from 3 months to 1–2 years or even longer (depending on the field, the publication lag, the accessibility of the journal, and the field's turnaround time for reading and citation). In physics, the "cited half-life" of an article (the point at which it has received half of all the citations it will ever receive) is around 5 years (*ISI Journal Citation Reports* shows most physics-based journals having a cited half-life between 3 and 10 years, see Thomson ISI 2003). Although articles may continue to be cited for as long as their contents are relevant (in natural science fields this could be forever), citation counts using the ISI Journal Impact Factor (Garfield, 1994) use only 2 years of publication data in a trade-off between an article being recent enough to be useful for assessment, and allowing sufficient time for it to make its impact felt.

Is it possible to identify the importance of an article earlier in the read–cite cycle, at the point when authors are *accessing* the literature? Now that researchers access and read articles through the Web, every download of an article can be logged. The number of downloads of an article is an indicator of its *usage impact*, which can be measured much earlier in the reading–citing cycle.

We use download and citation data from the UK mirror of arXiv.org (Warner, 2001)—an archive of full-text articles in physics, mathematics, and computer science that have been self-archived by their authors since 1991—to test whether early usage impact can predict later citation impact (Perneger, 2004 [and comments on—Harnad & Brody, 2004] has performed a similar study based on the *British Medical Journal* [*BMJ*], to which we have previously publicized). For a time-period of 2 years of cumulative download and citation data the correlation between download and citation counts is found to be 0.440 (for high energy physics, $n = 14200$, $p = .000$). When this overall 2-year effect is tested at shorter intervals, it turns out that the asymptotic 2-year correlation is

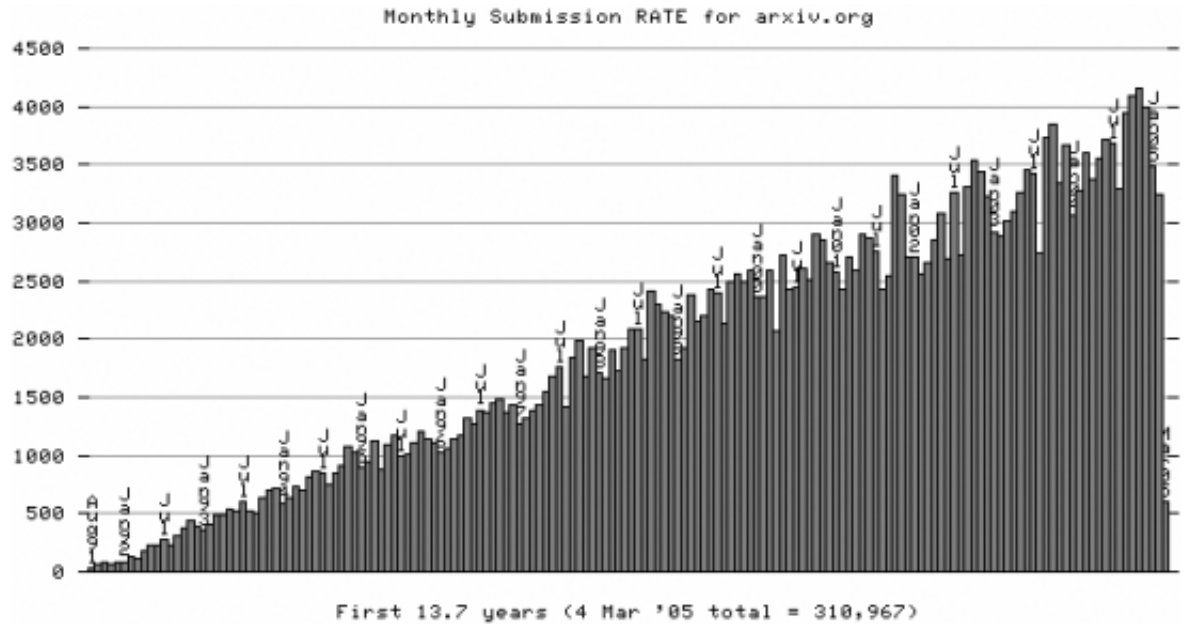


FIG. 1. The monthly number of full-text deposits to arXiv has grown linearly since its creation, to its current level of 4000 deposits per month. (Graph from http://arxiv.org/show_monthly_submissions)

already reached by 6 months. (Web log data are available only from 2000 onwards, so to derive a 2-year window of subsequent data, only articles deposited between 2000 and 2002 are included.) The correlation $r = 0.462$ ($n = 14917$, $p = .000$) is found for articles deposited in 2000 for all subsequent citations and downloads up to March 2005, i.e., from 5 years of data for an article from January 2000 to 4 years of data for an article from December 2000.)

The following section describes the arXiv.org e-print archive and the data used from its UK mirror for this study. We describe how the citation data is constructed in Citebase Search, an autonomous citation index similar to CiteSeer (Bollacker, Lawrence, & Giles, 1998). We introduce the *Usage/Citation Impact Correlator*, a tool for measuring the correlation between article download and citation impact. Using the Correlator, we have found evidence of a significant and moderately large correlation between downloads and citations. We accordingly conclude that downloads can be used as early predictors of citation impact.

Where available the URLs have been provided for the location of the scripts used to generate the graphs, as most of the data and tools described in this articles are available on the Web for general use.

The arXiv.org Database

ArXiv.org is an online database of self-archived (Ginsparg, 2003; Harnad, 2001) research articles covering physics, mathematics, and computer science. Authors deposit their articles as preprints (before peer review) and postprints (after peer review—both referred to here as *e-prints*) in source format (often Latex), which can be converted by the arXiv.org service into postscript and PDF. In addition to depositing the full-text of the article, authors provide metadata. The metadata include the article title, author list, abstract, and optionally a journal reference (where the article has been or will be published). Articles are deposited into "sub-arXivs," subject categories for which users can receive periodical e-mail alerts listing the latest additions.

The number of new articles deposited in arXiv have been growing at an unchanging linear rate since 1991 (Figure 1). Hence, in the context of all the relevant literature (and assuming that the total number of articles written each year is relatively stable), arXiv's *total annual coverage*—i.e., the proportion of the total annual published literature in physics, mathematics, and computer science that is self-archived in arXiv—is increasing linearly. The subareas of arXiv are experiencing varying rates of growth. The high energy physics (HEP) subarea is growing least (because most of the material within that arXiv subject is already being self-archived), whereas condensed matter and astrophysics (ASTRO) are still growing considerably (Figure 2). Kurtz et al. (2004) found 74% of articles published by the *Astrophysical Journal* in 2003 also had a version deposited in arXiv. As HEP is an older subject area than Astro, it is likely those journals whose articles fall within arXiv's HEP field will have similarly high percentages self-archived in arXiv.

In addition to being aided by the wide coverage of the HEP sub-arXiv, Citebase's ability to link references in the HEP field is strengthened by the addition of the journal reference to arXiv's records by SLAC/SPIRES High Energy Physics Literature Database (http://www.slac.stanford.edu/spires/hep/). SLAC/SPIRES indexes HEP journal articles, and links the published version to the self-archived e-print version in arXiv. Where an author cites a published article without providing the arXiv identifier, Citebase can use the data provided indirectly by SLAC/SPIRES to link that citation, thereby counting it in the citation impact.

DOI: 10.1002/asi

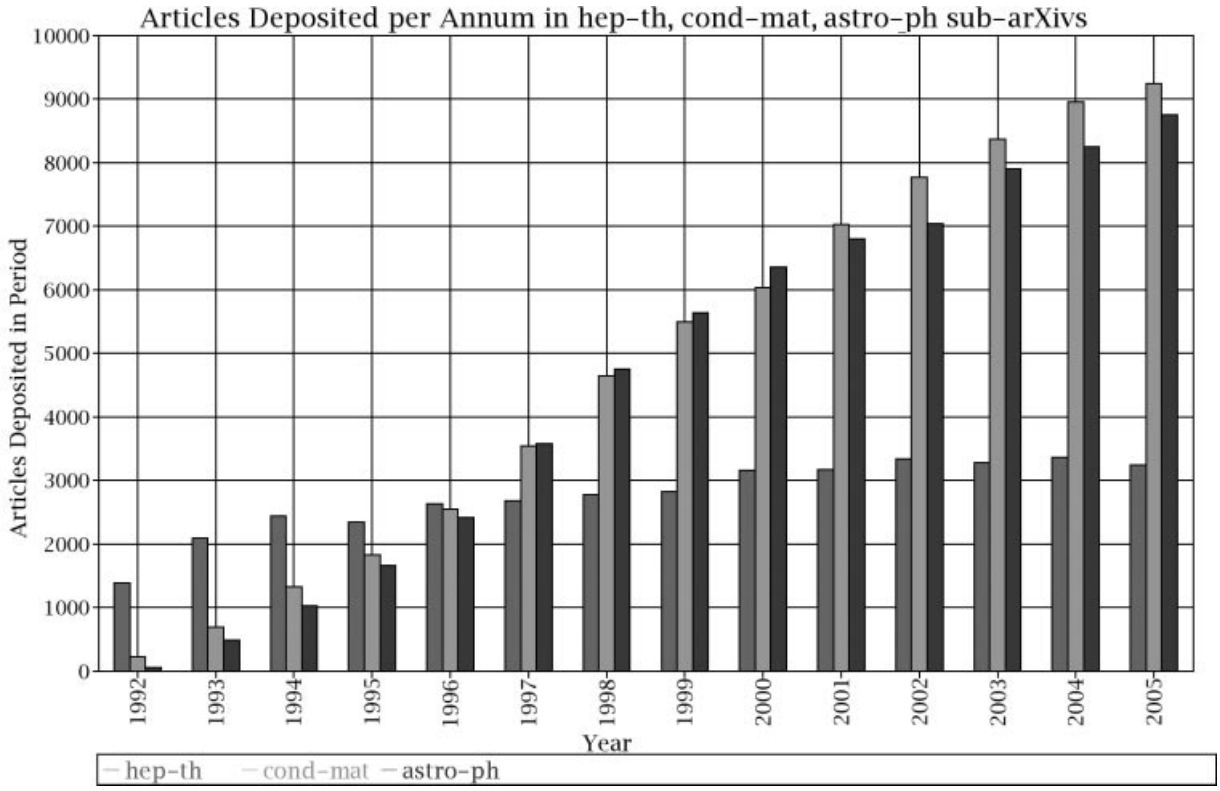


FIG. 2. Deposits in three of arXiv.org's subfields. HEP-TH (theoretical high energy physics) seems to have reached an asymptote, with little annual growth since the mid-1990s. In contrast, in COND-MAT (condensed matter) and ASTRO-PH (astrophysics) self-archiving rates are still growing substantially each year. (See http://citebase.eprints.org/cgi-bin/analysis/statistics?type=papers_per_field&format=graph&field=hep-th&field=cond-mat&field=astro-ph)

With 300,000 articles self-archived over 12 years, arXiv is the largest self-archived *centralized* e-print archive. There exist bigger archives, such as Citeseer whose contents are computationally harvested from *distributed* sites. The Astrophysics Data Service (ADS) by scanning back catalogs and in collaboration with publishers provides comprehensive free-access to the astrophysics literature. arXiv is an essential resource for research physicists, receiving 10,000 downloads per hour from the main mirror site alone (there are a dozen mirror sites).

Over the lifetime of arXiv there is evidence that physicists' citing behavior has changed, probably as an effect of arXiv's rapid dissemination capability. Figure 3 shows that the average latency between an article being deposited and later being cited has been reduced substantially. What (in 1992) used to be a citation peak 12 months after deposit has today shrunk to almost zero delay between the deposit date and the citation peak (Figure 3). The advent and growth of electronic publishing has certainly reduced the time between when an author submits a preprint and when the postprint is published, but the evidence from arXiv is that authors are also increasingly citing very recent work, both pre- and postrefereeing. This raises some interesting questions about the role that peer-review—as quality-controller and gatekeeper for the literature—plays for arXiv.org authors. There is no doubt, however, that the rapid dissemination model of arXiv has accelerated the read–cite–read cycle substantially.

Harvesting From the arXiv.org Database

ArXiv provides access to its metadata records through the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) in Dublin Core format. As the full-texts are available without restriction, these are harvested by a Web robot (which knows how to retrieve the source and PDF versions from arXiv's Web interface). Both metadata and full-text are stored in a local cache at the University of Southampton.

Web logs in Apache "combined" format are sent from the UK arXiv mirror (also at Southampton) via e-mail and stored locally. Web logs for the other arXiv mirror sites (including the main site in the United States) are currently not made available to us. The Web logs are filtered to remove common search engine robots, although most Web crawlers are already blocked by arXiv.[1] Requests for full-texts are then extracted, e.g., URLs that contain "/pdf/" for PDF requests.

[1]In addition to having a "robots.txt" (that declares full-text downloads off-limits to all sites, with the exception of Google) arXiv watches for rapid downloads from a single site to restricted areas of the Web site, and blocks those sites if they continue after being given a warning.

DOI: 10.1002/asi

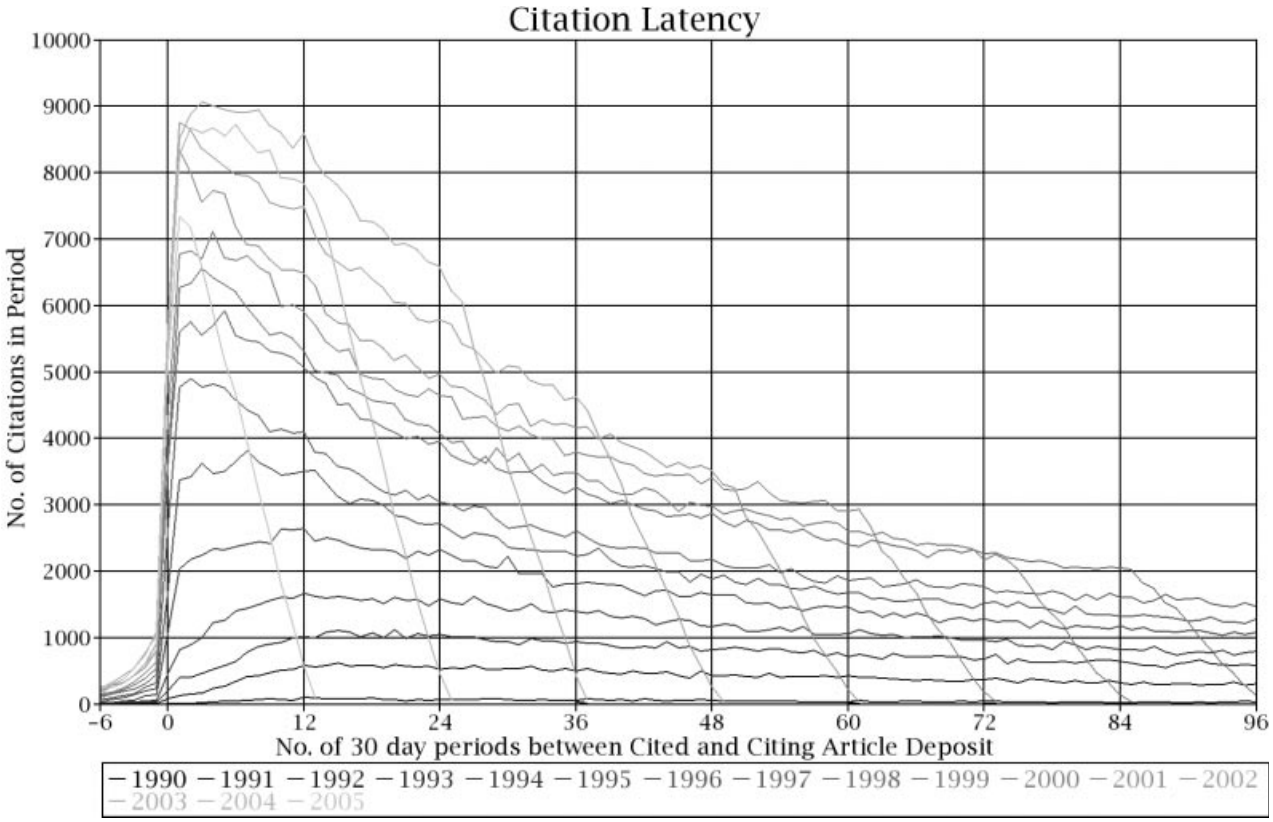


FIG. 3. Citation latency is the amount of time (in 30-day increments) between when an article is deposited and when it is cited (paired-values). This graph plots the frequency of citation latencies by the year the *cited* article was deposited. Each line represents a different sample year, with newer sample years containing more articles, hence a higher line on the graph. The significance of this graph lies in the changing distribution of latencies: For articles deposited in 1992 the peak citation rate was reached only in the following year (+12 months). The citation peak latency has since shrunk to almost zero (see also Figure 4). Negative latencies occur when the citing article has an accession date *after* the cited article. This could happen for three reasons: (a) an article is updated to include references to new articles (a facility supported by arXiv.org), (b) an article cites a published version for which the e-print was later deposited, or (c) an author has cited an article they know will exist but has not yet been published (e.g., they have read a draft elsewhere). (See http://citebase.eprints.org/cgi-bin/analysis/statistics?format=graph&type=citation_latency_per_year.)

On any given day only one full-text download of an article from one host is counted (so one user who repeatedly downloads the same article will only be counted once per day). This removes problems with repeated requests for the same article, but results in undercounts when more than one user requests an article from a single host or from behind a network proxy. This study cannot count multiple readings from shared printed copies, or readings from electronic copies in different distribution channels such as the publisher.

Each full-text request is translated to an arXiv identifier and stored, along with the date and the domain of the requesting host (e.g., "ac.uk"). This corresponds to some 4.7 million unique requests from the period August 1999 (when the UK arXiv.org mirror was set up) to October 2004. Because only one mirror's logs are available, this biases the requests towards UK hosts, and possibly towards UK-authored articles. This bias cannot be tested or corrected unless the logs are made available from other mirrors, and augmented with data from other e-print archives—as we hope these results will encourage them to be.

Citebase

Citebase is an autonomous citation index. Metadata records are harvested from arXiv.org using the OAI-PMH.[2] These records are indexed by Citebase, along with records from several other OAI-PMH compliant archives. In addition to harvesting metadata, full-texts are downloaded from arXiv.org and parsed by Citebase to extract their reference lists. These reference lists are parsed, and the cited articles are looked up in Citebase. Where the cited article is also deposited in arXiv.org, a citation link is created from the citing article to the cited article. These citation links create a citation database that allows users to follow links to cited articles ("outlinks"), and to see what articles have cited the article they are currently viewing ("inlinks"). Citation links are stored as a list of pairs of citing and cited article identifiers.

[2]Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), http://www.openarchives.org/.

DOI: 10.1002/asi

The total number of citation inlinks to an article provides a *citation impact* score for that article. Within Citebase the citation impact—as well as other performance metrics—can be used to rank articles when performing a search.

The citation impact score found by Citebase is therefore dependent upon several systematic factors: whether the cited article has been self-archived, the quality of the bibliographic information for the cited article (e.g., the presence of a journal reference), the extent to which Citebase was able to parse the references from citing articles, and how well the bibliographic data parsed from a reference matches the bibliographic data of the cited article. Citebase's citation linking is based either upon an arXiv.org identifier (if provided by the citing author), or by bibliographic data. Linking by identifier can lead to false positives, where an author has something in their reference that looks like an identifier but is not, or where an author has made a mistake (in either case the reference link goes to the wrong article). Linking by bibliographic data is more robust, as it requires four distinct bibliographic components to match (author or journal title, volume, page, and year), but this will obviously be subject to some false positives (e.g. where two references are erroneously counted as one) and uncounted citations from missed links.

Accuracy of Citation Links Within Citebase

The citation impact of articles within this study is dependent upon the number of references found by Citebase to those articles. An absolute limit on the number of citations is the coverage of the body of literature analyzed, i.e., arXiv's holdings. The question of coverage is a general problem for any study of citations, as anything from the references from a single journal, to links from Web pages may be counted. Of course, there could be significant differences in the purpose of making a citation between subjects, journals, or the Web in general. Within the context of this study, we are comparing downloads and citations from the same source, so the biased coverage of citations (only from arXiv articles to arXiv articles) will also be a bias in downloads (arXiv downloads only).

For a certain proportion of articles within arXiv, no references can be successfully extracted. This may arise because of the document format and reference style. It is an ongoing optimization process to keep trying to decrease the number of articles for which no references can be extracted, and to increase the accuracy of that extraction.

Articles that are published in a journal have two manifestations that may be cited by authors—the arXiv e-print and the official version available from the publisher's Web site. The arXiv version may be cited using the arXiv identifier, either in the absence of or in addition to the bibliographic reference to the journal. The publisher's version may be cited only by the bibliographic reference (particularly where an arXiv version may be deposited at a later date as a postprint).

To link a bibliographic journal reference the bibliographic data needs to be available for the cited article; the journal title, volume, etc. Unfortunately, many authors do not provide the journal reference for the published article that they have also deposited as an e-print in arXiv. This will depress the citation impact for that article, as only references that explicitly provide the arXiv identifier can be linked.

To check the accuracy of Citebase's reference linking, a sample of 500 randomly selected articles was chosen, spread across all years of the arXiv (1991+). Ninety articles from 2003–2005 were checked by hand to ascertain the number of reference links missed (references to items in Citebase, but with no link, hence not counted in the citation impact).

Article reference lists from Citebase's abstract pages provide several links—depending on the available data—to help find the cited articles: (a) a query to Google Scholar (http://scholar.google.com/) using author names, journal title and publication year; (b) for some publishers a link to the journal article; and (c) for this study an additional link was created to query Citebase using the author's names and publication year. It was assumed that references to articles older than 1992 would not be in arXiv (arXiv started in 1991), nor would books.

Over time, Citebase's capabilities have been extended, so the initial search for the cited item was made using the latest iteration of the OpenURL link resolver (that will make use of additional rules to tidy up references, which are run only occasionally against the entire database). If the direct Citebase look-up failed, queries to Google Scholar or the publisher were made to attempt to ascertain a fuller citation for the cited item. If a title was forthcoming from either of those sources it was used to query Citebase, otherwise a query based upon only authors and year was made with the cited item possibly being among the matches. Where an unlinked reference couldn't be found in Google Scholar or by the publisher (i.e., an unknown journal), it was assumed that it was not in Citebase.

During this study it became clear that a common cause of link failure was inconsistent formatting of journal and volume, e.g., "Phys.Rev.D65" and "Phys.Rev.D 65", causing the "D" to be picked up as part of the volume or journal title, respectively. Citebase has since had additional rules added to reformat these cases consistently, although to keep this study consistent these additional citation links were not included. (Tables 1–3).

Correlation Between Citations and Downloads

Correlation is a statistical measure of how two variables covary. Two positively correlated variables x and y will tend to have high values of x paired with high values of y, and low

TABLE 1. Summary of sample set used to test Citebase's reference parsing and citation linking accuracy.

Total articles	500
Articles without any references	8
Articles with no linked references (including no refs)	64
Average number of references/article	33
Average number of references linked/article	14

DOI: 10.1002/asi

TABLE 2. Summary of reference links from an example paper. The authors of this paper have already provided arXiv e-print identifiers for all references that could be located by this study within arXiv's collection. One "reference" as counted by Citebase was a single-authored reference split into two where the author had used a semicolon (Citebase treats semicolons as reference separators), so the actual number of authored references is 63. "Misc. material" are what looks like technical reports, while "non-published" is, e.g., "private communication."

Article identifier	oai:arXiv.org:hep-ph/0502036
References	64
References w/arXiv identifier	48
Reference links missed to arXiv articles	0
References unlinked w/journal reference	8
References to misc. material	5
References to non-published items	3

TABLE 3. The most recent 90 articles from the random selection of 500 articles were chosen to look at in detail (in total 1293 references that aren't linked to the cited item, although most are to items older than arXiv or not journal articles). On average 1.8 references were found in each article that were in arXiv, but not linked by Citebase—5% of all references.

2003–2005 Articles	90
Average number of references/article	35.26
Average number of references linked/article	21.46
Average number of references failed to be linked/article	1.80

values of x with low values of y. A negative correlation is where high values of x are paired with low values of y. Correlation is a normalized, scale-independent measure based on standard deviations above and below each variable's mean—the raw values of x and y can be in any number range.

A correlation between x and y may occur because x influences y, y influences x, the influence is in both directions, or a third variable influences both x and y. Intuitively, one would expect citations and downloads to exert a bidirectional influence, cyclical in time: An author reads a paper A, finds it useful and cites it in a new paper B (download causes citation). Another author reads B, follows the citation, reads A (citation causes download) and then perhaps goes on to cite it in another paper, C (download causes citation), etc. The correlation will be less than 1.0, not only because we don't cite everything we read, nor read everything that an article we read cites, but because both downloads and citations are subject to other influences outside this read–cite–read cycle (e.g., from alternative discovery tools, or when authors copy citations from papers they read without reading the cited works–perhaps when they have read the article before). "Reader-only" users contribute to the cite–read effect but not the read–cite effect, adding further noise to the read–cite effect.

Monitoring the correlation between citations and downloads is also informative because although articles can be downloaded and cited for as long as they are available, the peak rates for downloads and citations tend to occur at different times. The articles in arXiv.org that are over a year old show an almost flat rate of download, whereas their citation rate shows a more linear rate of decay over the period of available data (Figure 4). If there is a correlation between citations and downloads, a higher rate of downloads in the first year of an article could predict a higher number of eventual citations later.

We accordingly built a "correlation generator" to analyze the relationship between the citation and download counts for research articles in arXiv.org and to test whether a higher rate of downloads leads to a higher rate of citations (Figure 5).

Correlation Generator

The correlation generator provides a number of filters that can be used to restrict the data going into the correlation. As the correlation is calculated from pairs of citation and download counts corresponding to one article, the filters determine which articles to include and which downloads and citations to count. The values used can be specified symbolically, by entering values into the form, or via icons, by clicking the upper and lower limits on the mini-graphs relevant to the filter (Figure 6).

Articles can be filtered in terms of their arXiv subarea (e.g., high energy physics), the date the article was deposited, and the total number of citations/downloads to each article. This is particularly useful for restricting the analysis to articles for which sufficient data are available: For example, whereas there exist articles that have been deposited since 1991, the download data are only available from 1999. Hence, although the download data cover all of the articles deposited up to that date, the predictive power of downloads can only be tested for articles deposited since 1999.

Each citation and each download has a latency value: (a) the time between when the cited and citing article were deposited, and (b) the time between when an article is deposited and when it is downloaded. The user might choose to include only downloads that occurred up to a week ("7 days") after an article was deposited.

Once all the filtered pairs have been found, the natural logarithm is calculated for both values, to allow a generalized linear correlation to be performed.

Correlation Generator Implementation

The generator is based on Citebase's MySQL database and a combination of Perl server-side scripts and Java client-side Web applications. The tables relevant to calculating the correlation are the citation links table, download "hits" table, and record datestamps (the earliest datestamp is taken as the date of accession—the date used by the generator to determine latency and to filter by date). The downloads and citation links are preprocessed into latency tables that consist of the article's identifier and the number of days since accession, e.g., an article 58432 deposited on May 14th and then downloaded on the May 21st is stored in the downloads latency table as "58432—7," similarly 58432 cited by an article 69710 deposited on May 26th is stored in the citations latency table as "58432—12." At the time of writing, the

DOI: 10.1002/asi

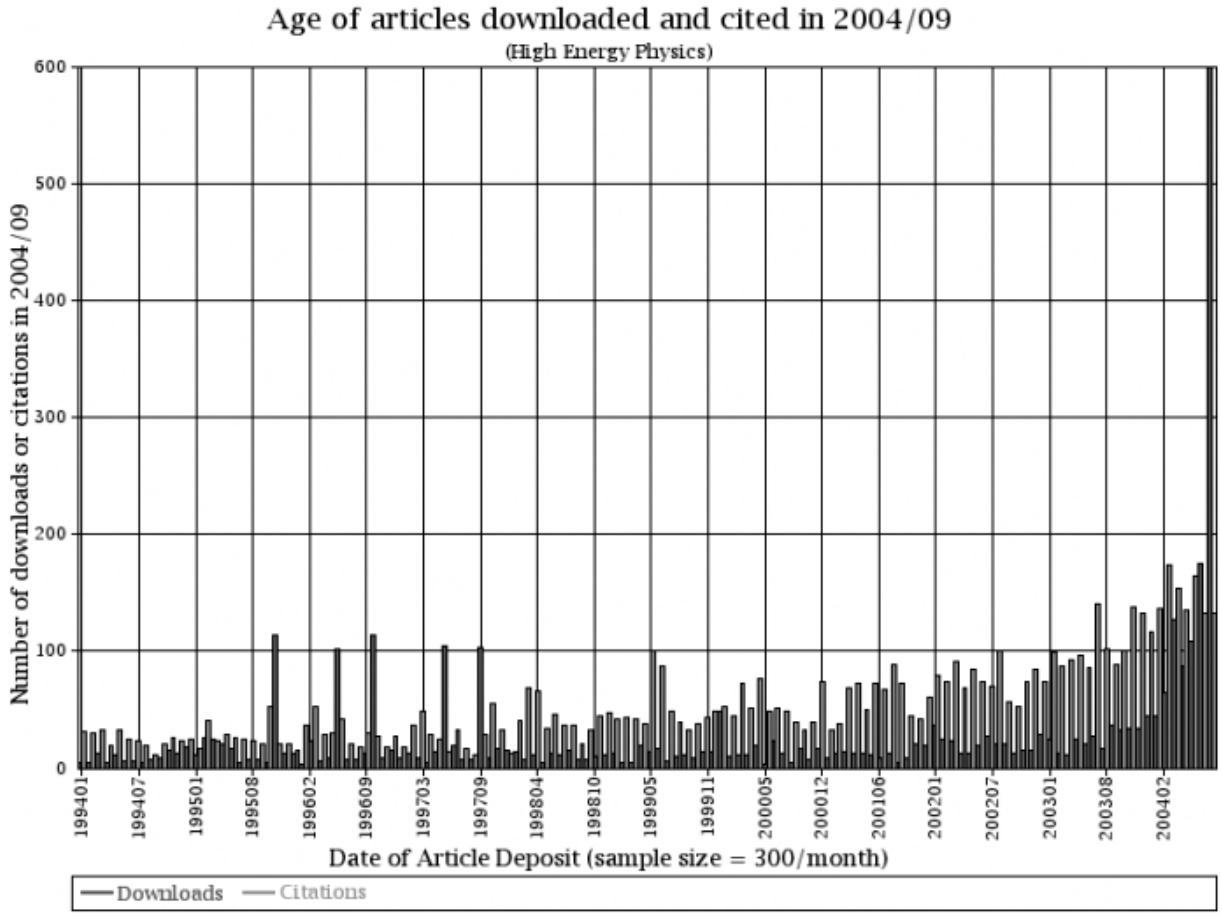


FIG. 4. Download and citation frequencies for all articles in arXiv.org downloaded or cited in September 2004. The most frequently downloaded articles are those deposited in the previous year (2003–) with a steep fall-off during that year; for articles from earlier years downloads are (with a few exceptions) few and equal from year to year. For citations the fall-off looks more gradual and linear, taking about 6 years or more to settle into a flat, constant rate. If higher impact articles account for that higher rate of downloads in the first year, then the initial year of download data could be used to predict citation impact data over the longer term. A random sample of 300 articles was chosen from each preceding month until 1994–2001, with the number of downloads and citations to each month's sample plotted as a bar (downloads as black, citations as gray; http://citebase.eprints.org/cgi-bin/analysis/statistics?type=hits_latency_normalised).

download latency table contained 4 million records and the citation latency table 2.3 million records, which have to be processed in real-time to support an interactive tool. Citebase is updated daily with new articles from the source repositories which, in addition to refinements to Citebase's processes, results in the data set used by the correlation generator changing over time.

To provide the user with a graphical representation of the source data the database tables are rendered by Java graphs (Figure 6), which retrieve the data as a plain-text list of values from a supporting server-side script. The graphs are tied by client-side Javascript into the main submission form (Figure 5) allowing the user to "click" on the graphs, with the appropriate values being filled into the form.

When the user submits the Web form the citations and downloads are first filtered: only those article identifiers within the given arXiv.org subfield, articles whose accession is within the given date range, and only those downloads and citations that occurred within the given latency period. Articles with no citations and/or no downloads are discarded. The citations and downloads are subtotaled for each article, from which the logarithm is taken. The correlation is calculated from the citations and downloads logarithmic values. The server-side script either returns a single image of a scatter graph, a summary table, and the correlation or—if the user chose to output as "table"—a list of paired values allowing the raw citations/downloads values to be imported into a separate statistical package.

Sample correlations. The correlation generator builds a scatter plot, as well as calculating the basic distribution of the citation and download counts, and the correlation between the two. The scatter plot consists of density dots—the darker the color the more pairs exist at the same point. This helps to emphasize where the bulk of the pairs lie.

The basic statistical information on the number of pairs used (n), and the distributions of the two variables (sum, mean, and SD) is shown. Both citations and downloads have some large deviations from the mean; this is due to a small

DOI: 10.1002/asi

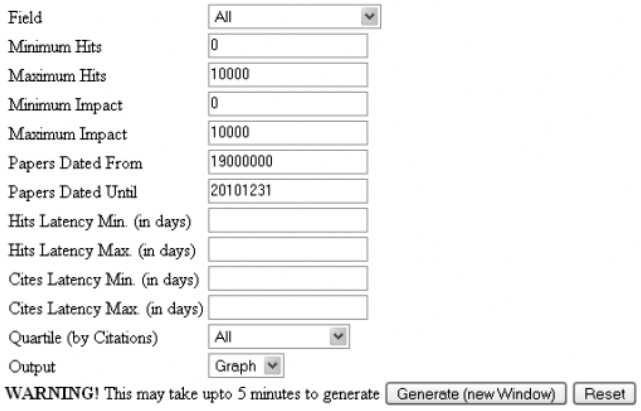


FIG. 5. Form used to generate correlations. This allows the user to choose what data and data-ranges are used to generate the correlation between downloads ("Hits") and citations ("Impact"). This provides filters to delimit which papers (Field, Min/Max Downloads, Min/Max Impact, Date), which downloads and citations (Min/Max downloads latency, Min/Max citation latency), and which citation quartile to include. The citation quartile includes in the result only the bottom, lower, upper, or top 25% of papers after rank-ordering by citation impact. Clicking "Generate" calls a Web script that extracts the data sets from Citebase, generates the correlation, and displays the result as a graph.

number of very high-impact articles accounting for most downloads and citations, while the majority of articles receive few or no downloads or citations—the distribution of citations has been shown to adhere to Zipf's law, meaning that the number of citations to an article is proportional to its rank order. (Figure 6 illustrates that the distributions for downloads and citations show exponential decay.)

The correlation generator calculates the value for Pearson's *r*—the degree to which data points deviate from the line of best fit. Pearson's gives values from −1 to 1: −1 is a perfect negative correlation; 0 is no correlation; and 1 is a perfect positive correlation. Pearson's in effect provides gradient of the "line of best fit," which goes through the mean (the correlator draws this line in red). Pearson's is intended to be used for data that is normally distributed, therefore to normalize the distributions for use with Pearson's *r* the natural logarithm of downloads and citations is taken; hence, the correlator uses a logarithmic axis (Figure 7). The correlator also calculates Spearman's Rank-order correlation, which converts each download and citation count into a rank order, from 1 to *n*, and performs Pearson's *r* on the rank-order values. For multiple points of the same value (e.g., many papers have only one citation) the average of the rank-order values is used for all points (Table 4). The algorithms used by the correlator were checked by entering the same source data into Microsoft Excel (Pearson's only; Microsoft, Redman, WA) and SPSS, a statistical analysis software program (SPSS, Inc., Chicago, IL) to compare the results for Pearson's *r* and Spearman's Rank-order.

When generating the correlation any downloads within the first 7 days of the article appearing were excluded, as these downloads reflect users scanning all new articles (e.g., in response to e-mail alerts), hence those downloads are unlikely to discern between high-impact and low-impact articles and would dampen any predictive effect. The first 7 days of downloads accounted for on average .5 downloads per article—roughly a fifth of all downloads.

While many articles may be downloaded, but not cited, articles that are highly cited are always downloaded. This can be seen in the scatter graphs; as high-download high-impact articles fall closer to the line of best fit, while low-impact articles appear to form a separate distribution above the line of best fit. Some articles that have high-download counts but low-citation counts may be the result of Citebase failing to find the citation links, e.g., where the author has not supplied the journal reference; hence, any citations to

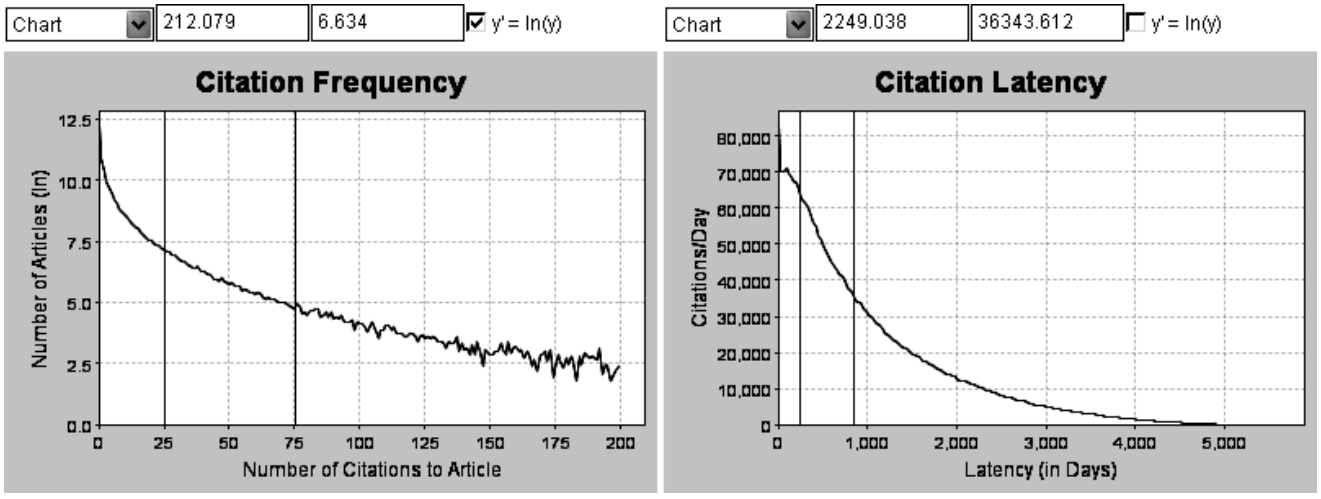


FIG. 6. Auxiliary graphs. These graphs show the distribution of the variables that go into the correlation (note that most use a logarithmic scale). Citation frequency shows the distribution of articles in terms of the number of times they were cited. Citation latency is the time between an article being deposited and later cited (total citations per day latency). Web download frequency is the distribution of articles in terms of the number of times they were downloaded. Web download latency is the time between an article being deposited and later downloaded. The user can click on the graphs to set minimum and maximum values, which are filled into the query form (Figure 5).

DOI: 10.1002/asi

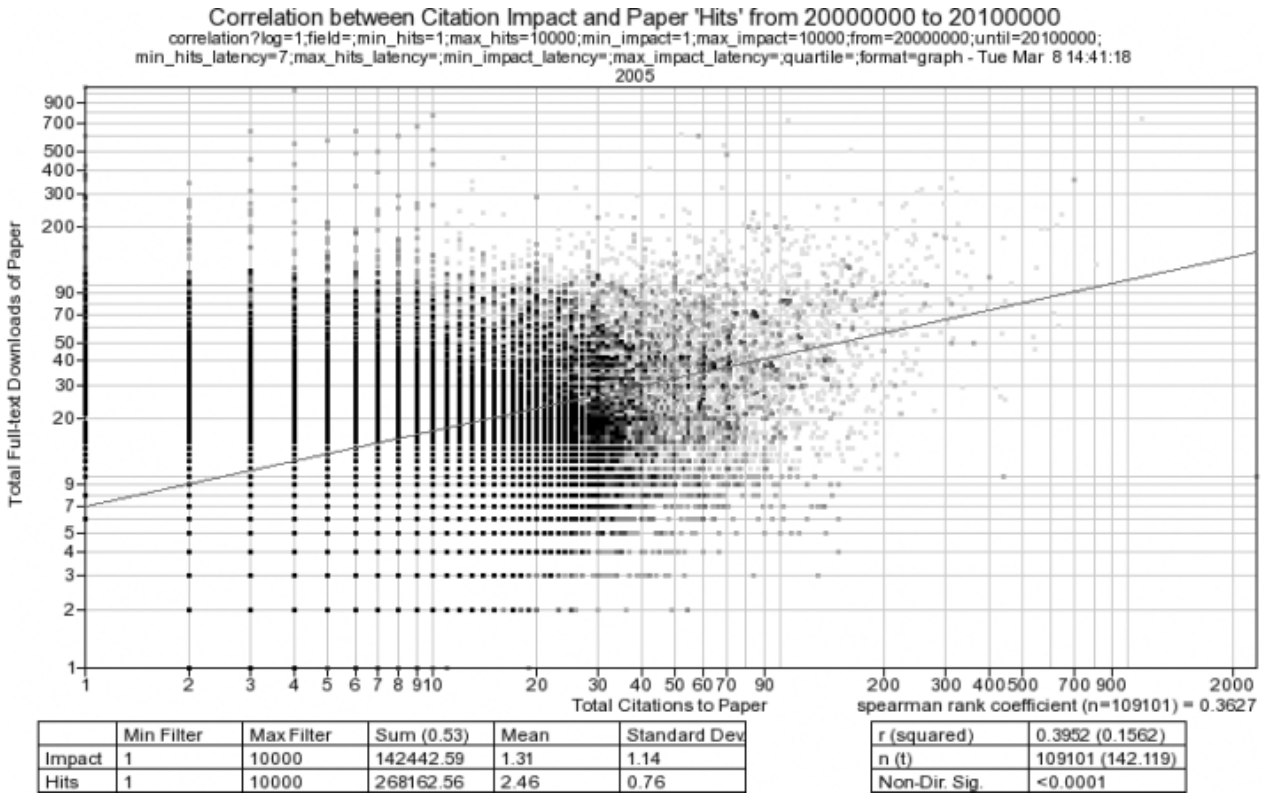


FIG. 7. Download/Citation correlation scatter-plot generated for all papers deposited between 2000–2004. Each dot corresponds to an article. The number of articles with the same values is indicated by shades of grey (black being the highest, with four or more articles having the same number of downloads and citations). The download and citation counts are the cumulative amounts up to March 2005. The correlation for these 109,101 articles is $r = 0.395$. From the distribution in the scatter graph it can be seen that the distribution is very noisy, but that few articles with high citation impact receive low download impact. The ratio of downloads to citations is 2.24:1 (only download statistics for the UK mirror are available), which corresponds to a mean of 1.30 citations for each article, and 2.46 downloads. Non-Direct Significance is the (two-tailed) probability of such a correlation by chance.

TABLE 4. Fictitious data to illustrate the values used to calculate the correlations using Pearson's *r* and Spearman Rank.

Downloads	Citations	Downloads (Log)	Citations (Log)	Downloads (Rank)	Citations (Rank)
3	1	1.10	0	1	1
6	2	1.80	0.69	2	2.5
7	2	1.95	0.69	3	2.5
30	5	3.40	1.61	5	4
12	13	2.48	2.56	4	5

those articles using only a journal reference cannot be linked and counted.

Predicting From Correlations

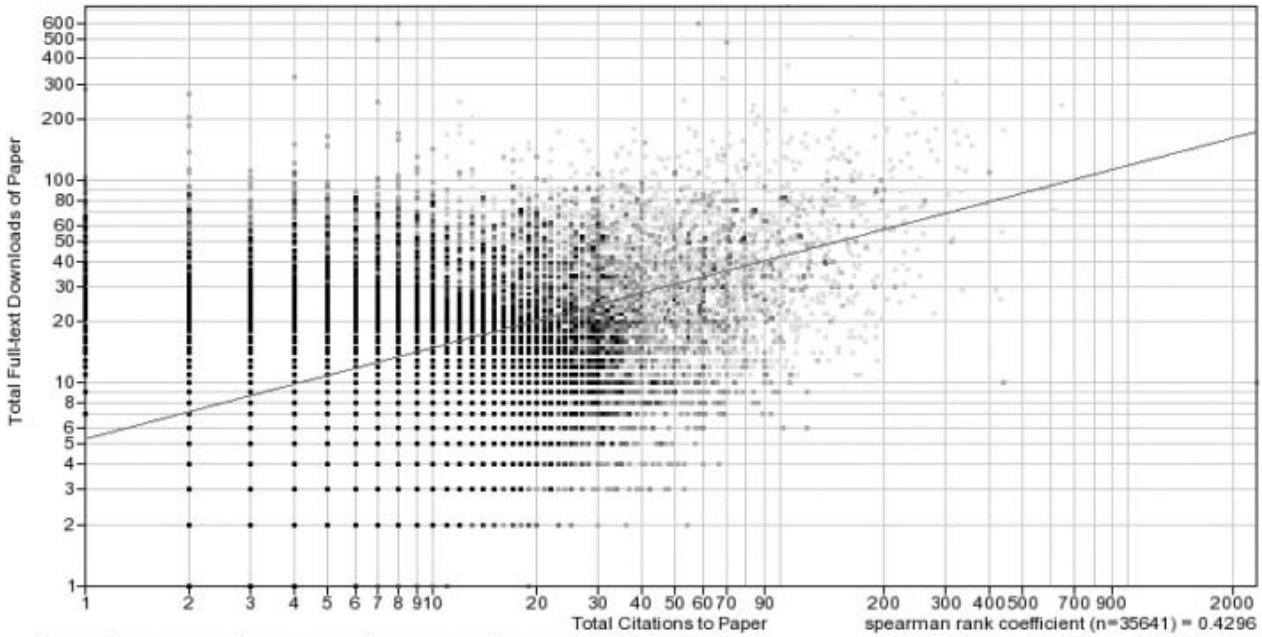
The articles used for testing how well downloads can predict citation impact are from the high energy physics sub-arXivs[3] (Figure 8). Download data are only available from late 1999 onwards and, to allow for 2 years of subsequent download and citation data following deposit, only articles up to 2002 can be included. Figure 9 shows the correlation for those 2000–2002 articles. All articles deposited in arXiv.org receive downloads very soon after deposit (resulting in background "noise" that reduces any predictive signal)—to minimize this effect the first 7 days of downloads are excluded.

Figure 10 shows the downloads and citations history for two articles deposited in April 2001 and December 1999, respectively. The first 2 months of downloads and 2 years of citations are highlighted by two boxes. To test how well downloads predict citations to the total downloads up to the end of 2 months is correlated with the total citations up to the end of 2 years (in relation to the date the article was deposited in arXiv.org). The correlation generator supports this by specifying the maximum number of days to include downloads and citations after the article is deposited. (Thirty days is taken as being 1 month, and 730 days as 2 years.)

Given that citation and download impact for the HEP sub-arXiv has a correlation of 0.462, how long do downloads need to be counted to get close to this correlation? To test this, queries were made to the correlation generator using nine different time periods for download data: 1, 2, 3, 4, 5, 6, 7, 12, and 24 months following the deposit of an article (Table 5). The correlation increases from 0.270 one month after deposit to 0.440 at 24 months. Figure 11 reveals an important finding: This increase is not linear and it approximates the correlation

[3]The hep-th, hep-ph, hep-lat, and hep-ex sub-arXivs are the longest established, hence most comprehensive and well citation-linked parts of arXiv.org.

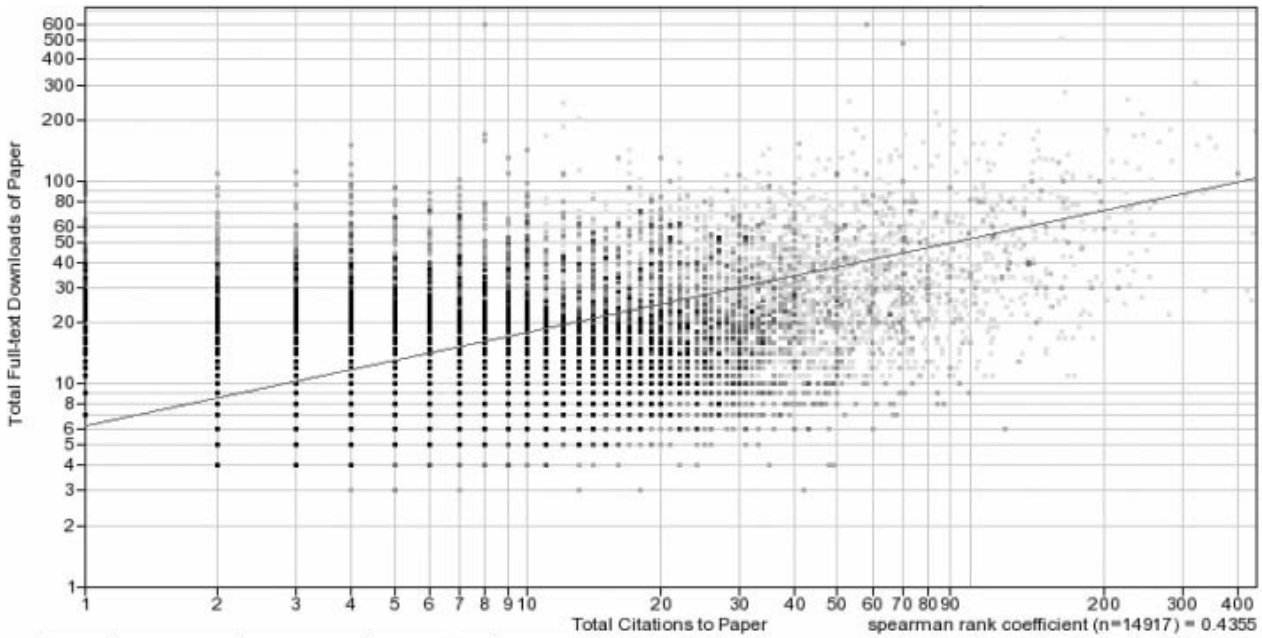
DOI: 10.1002/asi



	Min Filter	Max Filter	Sum (0.7)	Mean	Standard Dev
Impact	1	10000	61209.53	1.72	1.21
Hits	1	10000	86825.37	2.44	0.80

r (squared)	0.452 (0.2043)
n (t)	35641 (95.6655)
Non-Dir. Sig.	<0.0001

FIG. 8. Download/Citation correlation for papers deposited in 2000–current, from the High Energy Physics sub-arXivs. HEP papers have a mean citation impact of 1.72, and download impact of 2.44. The correlation is higher at $r = 0.452$ (due most likely to more accurate citation counts for the HEP sub-field). Non-Direct Significance is the (two-tailed) probability of such a correlation by chance.



	Min Filter	Max Filter	Sum (0.72)	Mean	Standard Dev
Impact	1	10000	29474.73	1.98	1.24
Hits	1	10000	40849.71	2.74	0.64

r (squared)	0.4619 (0.2134)
n (t)	14917 (63.6122)
Non-Dir. Sig.	<0.0001

FIG. 9. Download/Citation correlation for papers deposited 2000–2002 (a 2 year period), in the HEP fields. Given that there is a delay in citation impact differentiating articles, restricting the sample set to papers deposited more than 2 years ago provides a better correlation (0.462). Non-Direct Significance is the (two-tailed) probability of such a correlation by chance.

DOI: 10.1002/asi

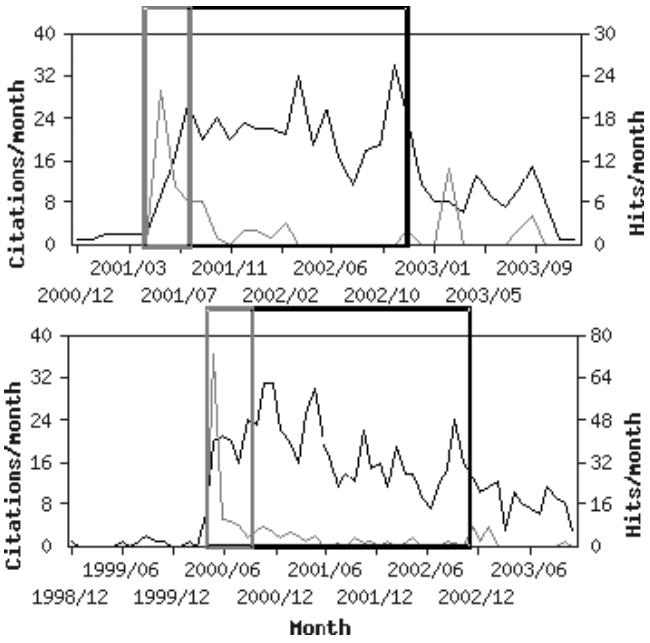


FIG. 10. An early-day window of downloads (gray) may predict a later window of citations (black).

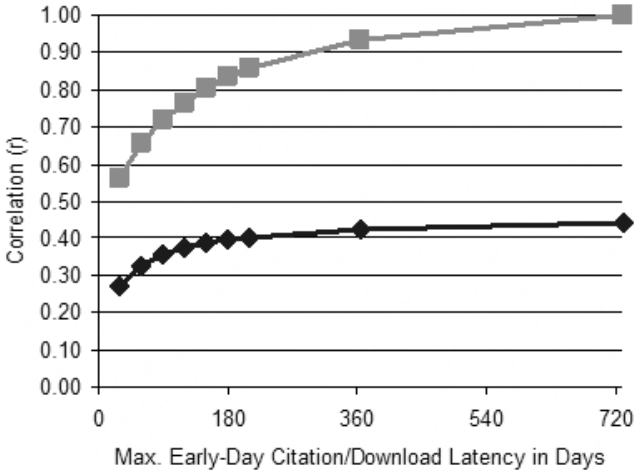


FIG. 11. Download/citation correlation (♦) and hence the power of download counts to predict citation counts reaches an asymptote about 6 months after deposit. Download impact at 6 months can predict citation impact at 2 years. Citation/citation correlation (■) shows a higher predictive correlation, possible due to arXiv's rapid distribution model allowing for citations to appear very soon after an article is deposited. At 2 years citation/citation is measuring the same values, hence $r = 1$.

TABLE 5. Correlation between citations and downloads at different maximum download latency periods. The longer the period for which downloads are counted, the higher the correlation between citation and download impact. After 6 months the correlation increases only by a small amount, suggesting that counting the downloads of an article at 6 months will provide as good a prediction of the citation impact of that article after 2 years as counting the downloads after 2 years.

Max. Download Latency (days)	Mean Downloads/Article (excl. first 7 days)	Correlation (r)
30	0.85	0.270
60	1.10	0.326
90	1.26	0.357
120	1.39	0.373
150	1.49	0.386
180	1.57	0.397
210	1.64	0.402
365	*1.86*	*0.424*
730	*2.20*	*0.440*

found with 2 years of download and citation data using only 6–7 months of download data. This suggests that if the baseline correlation for a field is significant and sufficiently large, the download data could be used after 6 months as a good predictor of citation impact after 2 years.

Figure 3 shows the peak "citation latency" for articles deposited in arXiv has decreased from 12 months in the early 1990s to no delay for new articles. The rapid distribution system of arXiv allows for the citation impact to be identified at the earlier preprint stage in the article's lifecycle (draft, preprint, postprint, published, etc.). Citations to the preprint within arXiv can be identified and tracked as soon as the article appears, and those citations could be a predictor of future citation impact.

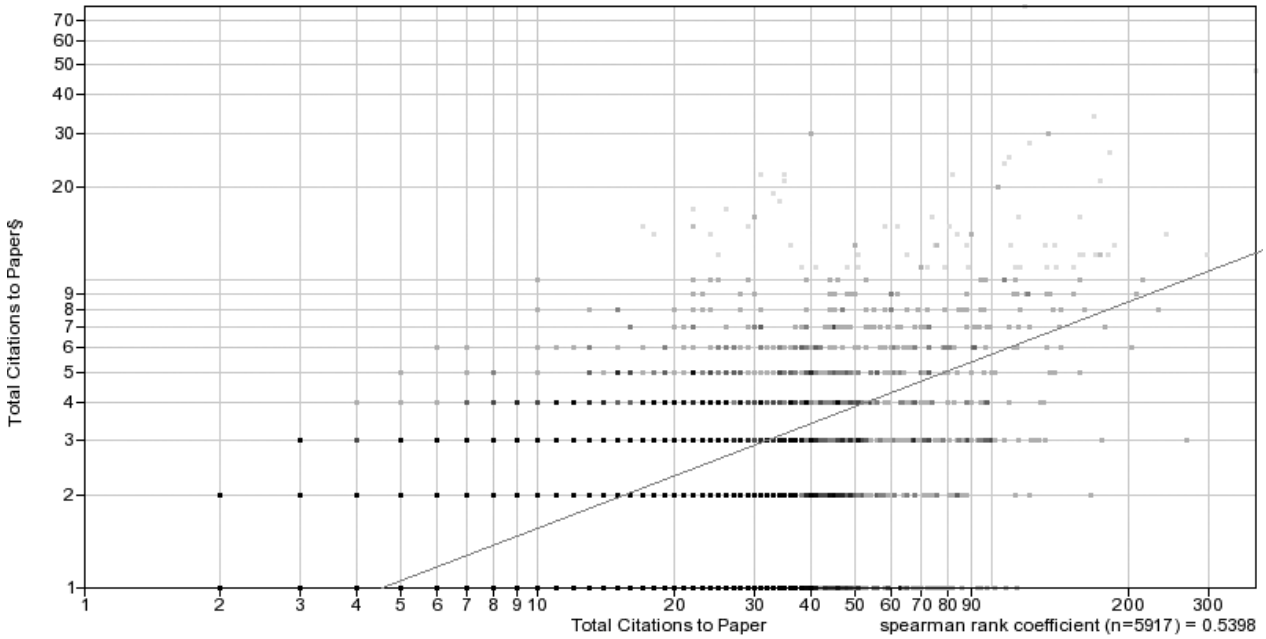
Comparing one month of citation data to the citation impact at 2 years results in a correlation of 0.565 (Figure 12). After 6 months this correlation rises to 0.834 (Figure 13). When compared against the ability of download/citation correlation to predict future citation impact (Figure 11) it is apparent that "early-days'" citations provide a stronger baseline correlation, but take longer (hence more citations) to reach the asymptotic point.

Conclusion

Whereas the use of citation counts as a measure of research impact is well established, Web-based access to the research literature offers a new potential measure of impact—download counts. Counting downloads is useful for at least two reasons: (a) The portion of later citation variance that is correlated with earlier download counts provides an early-days' estimate of probable citation impact that can already begin to be tracked the instant an article is made Open Access and that already attains its maximum predictive power after 6 months. (b) The portion of download variance that is uncorrelated with citation counts provides a second, partly independent estimate of the *usage impact* of an article, sensitive to another research performance indicator that is not reflected in citations (Kurtz, 2004).

This study found a significant and sizeable correlation between the citation and download impact of articles in physics (0.462), as well as in other arXiv fields: mathematics (0.347), astrophysics (0.477), and condensed matter (0.330). This was based on Web downloads from the UK arXiv.org mirror only, and on those citations that could be automatically found and linked by Citebase. The true correlation may in fact prove somewhat higher once more download sites are monitored and automatic linking becomes more accurate. The correlation will no doubt vary from field to field, and may also change as the proportion of Open

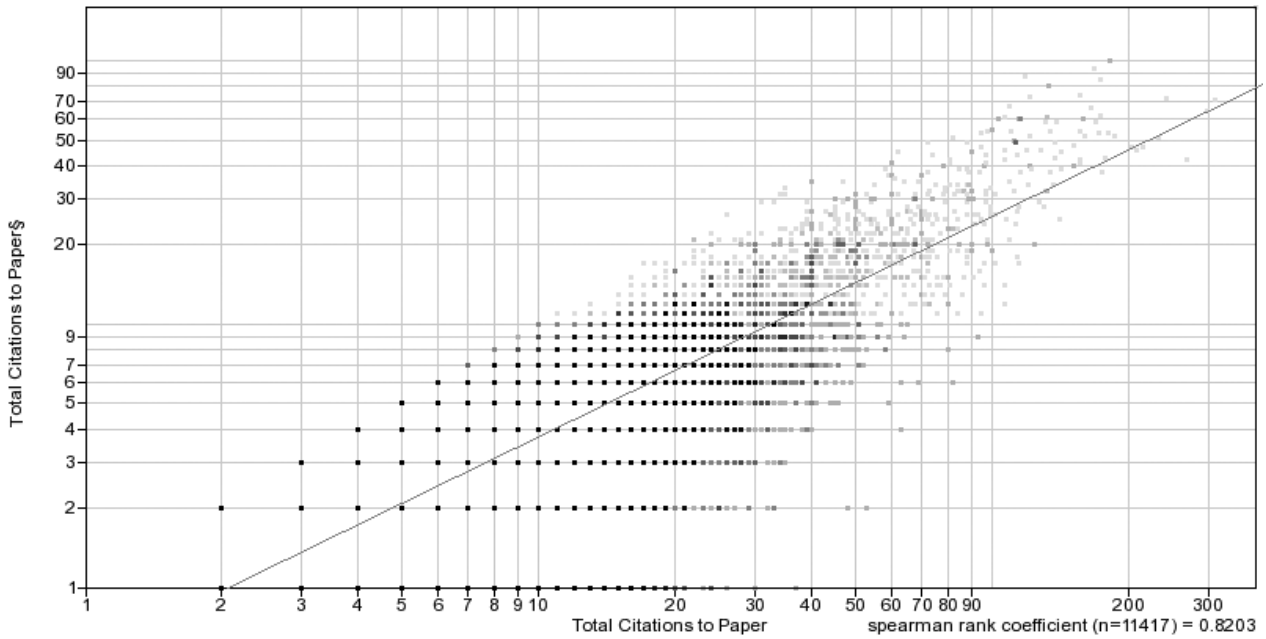
DOI: 10.1002/asi



	Min Filter	Max Filter	Sum (5.25)	Mean	Standard Dev
Cites§	1	10000	2565.87	0.43	0.61
Cites	1	10000	13478.87	2.28	1.10

r (squared)	0.5647 (0.319)
n (t)	5917 (52.629)
Non-Dir. Sig.	<0.0001

FIG. 12. Correlation between the citation impact for articles after 1 month (30 days) and 2 years (730 days)—high energy physics papers deposited between 2000 and end 2002.



	Min Filter	Max Filter	Sum (1.92)	Mean	Standard Dev
Cites§	1	10000	11402.03	1.00	0.91
Cites	1	10000	21894.14	1.92	1.08

r (squared)	0.8342 (0.696)
n (t)	11417 (161.633)
Non-Dir. Sig.	<0.0001

FIG. 13. Correlation between the citation impact for articles after 6 months (180 days) and 2 years (730 days)—high energy physics papers deposited between 2000 and end 2002.

DOI: 10.1002/asi

Access refereed research content (now only 10–20%) approaches 100% (and includes not only refereed journal and conference paper citations and consultations, but books and research data too).

It is likely that download impact is just the first of many new and powerful indicators of research impact and direction that will emerge from an Open Access corpus (Hitchcock, 2005, has compiled a list of many studies using and looking at OA material)—indicators that will include co-citation analysis (to and from jointly cited or citing articles and authors), co-download analysis (Bollen, van Sompel, Smith, & Luce, 2005), co-text analysis (from Boolean word conjunctions to latent semantic indexing and other measures of text similarity patterns and lineage, e.g., see Deerwester et al., 1990), citation-based analogues of Google's recursive PageRank (Brin & Page, 1998) algorithm weighting cited papers (or authors') citation ranks with the citation weights of the citing papers (authors), hub/authority analysis (papers cited by many papers vs. papers citing many papers, see Kleinberg, 1999), and time-series chronometric analyses. Citation and download counts are just the first two terms in what will be a rich and diverse multiple regression equation predicting and tracking research impact.

Acknowledgments

The authors thank Michael Kurtz and Simeon Warner for their suggestions and comments on drafts of this paper.

References

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems, 30, 1–7, 107–117.

Bollen, J., Van de Sompel, H., Smith, J., & Luce, R. (2005). Toward alternative metrics of journal impact: A comparison of download and citation data (Preprint). Retrieved February 9, 2006, from http://arXiv.org/cs/0503007

Bollacker, K., Lawrence, S., & Giles, L.C. (1998). CiteSeer: An autonomous Web agent for automatic retrieval and identification of interesting publications. In K.P. Sycara & M. Wooldridge, Proceedings of the Second International Conference on Autonomous Agents. Retrieved from http://citeseer.ist.psu.edu/bollacker98citeseer.html

Borgman, C., & Furner, J. (2002). Scholarly communication and bibliometrics. Annual Review of Information Science and Technology, 36, 3–72 (Preprint). Retreived February 9, 2006, from http://polaris.gseis.ucla.edu/jfurner/arist02.pdf

Brody, T. (2004). Citebase correlation generator. Retrieved May 17, 2004, from http://citebase.eprints.org/analysis/correlation.php

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(1), 391–407.

Garfield, E. (1994). The impact factor. Current Contents, 25, 3–7. Retrieved June 20, 1994, from http://www.isinet.com/isi/hot/essays/journalcitationreports/7.html

Ginsparg, P. (2003). Can peer review be better focused? Retrieved November 27, 2004, from http://arxiv.org/blurb/pg02pr.html

Harnad, S. (2001). The self-archiving initiative. Nature, 410, 1024–1025. Retrieved February 9, 2006, from http://cogprints.ecs.soton.ac.uk/archive/00001642/

Harnad, S., & Brody, T. (2004). Prior evidence that downloads predict citations. British Medical Journal. Retrieved February 9, 2006, from http://bmj.bmjjournals. com/cgi/eletters/329/7465/546#73000

Hitchcock, S. (2005). The effect of open access and downloads ('hits') on citation impact: A bibliography of studies. Retrieved March 3, 2004, from http://opcit.eprints.org/oacitation-biblio.html

Kleinberg, J.M. (1999). Hubs, authorities, and communities. ACM Computing Surveys. Retrieved February 9, 2006, from http://doi.acm.org/10.1145/345966.345982

Kurtz, M.J. (2004). Restrictive access policies cut readership of electronic research journal articles by a factor of two. Cambridge, MA; Harvard-Smithsonian Centre for Astrophysics. Retrieved February 9, 2006, from http://opcit.eprints.org/feb19oa/kurtz.pdf

Kurtz, M.J., Eichhorn, G., Accomazzi, A., Grant, C.S., Demleitner, M. & Murray, S.S. (2004). The effect of use and access on citation. Information Processing and Management, 41, 1395–1402. Retrieved February 9, 2006, from http://cfa-www.harvard.edu/~kurtz/IPM-abstract.html

Perneger, T.V. (2004). Relation between online "hit counts" and subsequent citations: Prospective study of research papers in the BMJ. British Medical Journal, 329, 546–547. Retrieved February 9, 2006, from http://bmj.bmjjournals.com/cgi/content/full/329/7465/546

Thomson ISI. (2003). ISI journal citation reports. Philadelphia, PA: Author.

Warner, S. (2001, March). arXiv, the OAI and peer review. Workshop on OAI and peer review journals in Europe, Geneva. Retrieved February 9, 2006, from http://eprints.rclis.org/archive/00000890/

DOI: 10.1002/asi