SLAC-TN-11-002

Clicks versus Citations:

Click Count as a Metric in High Energy Physics Publishing

Ayelet Bitton

Office of Science, Science Undergraduate Laboratory Internship (SULI)
University of California, San Diego
SLAC National Accelerator Laboratory
Stanford, CA

August 20, 2010

Prepared in partial fulfillment of the requirements of the Office of Science, Department of Energy's Science Undergraduate Laboratory Internship under the direction of Travis Brooks at the SPIRES/INSPIRES Group, SLAC National Accelerator Laboratory.

Participant: ____________________________________

Signature

Research Advisor: ____________________________________

Signature

Work supported in part by US Department of Energy contract DE-AC02-76SF00515.

TABLE OF CONTENTS

ABSTRACT

Clicks versus Citations: Click Count as a Metric in High Energy Physics Publishing. AYELET BITTON (University of California, San Diego, La Jolla, CA 92093) TRAVIS BROOKS (SPIRES/INSPIRES Group, SLAC National Accelerator Laboratory, Menlo Park, CA 94025)

High-energy physicists worldwide rely on online resources such as SPIRES and arXiv to perform gather research and share their own publications. SPIRES is a tool designed to search the literature within high-energy physics, while arXiv provides the actual full-text documents of this literature. In high-energy physics, papers are often ranked according to the number of citations they acquire — meaning the number of times a later paper references the original. This paper investigates the correlation between the number of times a paper is clicked in order to be downloaded and the number of citations it receives following the click. It explores how physicists truly read what they cite.

INTRODUCTION

Traditionally, costly journals have ruled the world of published research papers. However, in high-energy physics (HEP), free online paper repositories have led to a huge change in how physicists perform research and share their findings. With the invention of the internet, the transmission of information — including research results — has moved online. Within the world of science, and particularly HEP, scientists are now able to reach more information than ever before, at tremendously increased speeds.

HEP is unique in that its participants have access to a majority of the field's literature online, free of cost. Over two decades ago, the first website in the United States came online: SPIRES, a database designed to effectively search HEP papers [1]. This open access resource initiated a change in the way scientists normally interact with research publication. It also provides a unique means of studying how a community with such infrastructure operates. This method of sharing and accessing information has already led to numerous changes in the publishing habits of high-energy physicists.

Past research has already shown that in the two decades since the invention of the internet, the field norms have moved to the point that over 95 percent of the HEP community publishes its papers online [2]. These papers are submitted in the form of preprints to arXiv — a database that stores the actual copies of the papers SPIRES allows users to search for [3]. Preprints await journal review, but are often still of a high caliber due to "the invisible hand of peer review" [4]. Papers that are published to arXiv later tend to be cited over 14 times on average, while papers not submitted to arXiv receive less than four citations on average [2].

Currently, a citation count — the number of times a specific paper is referenced by a later paper — acts as the sole objective metric for measuring a paper's or author's success and influence. SPIRES simplifies this by offering citation statistics, displaying a specific author's or paper's citation count alongside the article's information. In this paper, we shall investigate whether the number of clicks a paper receives in any way correlates with the number of citations it later receives. For the purposes of this paper, the word "click" represents only clicks on the SPIRES website that led to an actual download of a paper, which was then theoretically read by the user. To simplify the question: are researchers clicking (and therefore reading) what they cite from SPIRES and arXiv? If so, a paper's click count may also serve as a metric for relevance, and may be worth considering alongside its citation count. Certain paper repositories have already implemented a display of a paper's click counts [5], but as of this point, no concrete research has commenced the search for a correlation between clicks and citations. Should such a correlation exist, it could be a further adjustment that SPIRES and other online repositories have had on the way high-energy physicists perform their research.

METHODS AND MATERIALS

SPIRES is a database holding the metadata for over 750,000 scientific papers, including papers published in journals and online repositories such as arXiv[1]. It features a search system for finding papers, and provides links to the journals and databases where these papers are available. SPIRES works hand in hand with arXiv, which provides the actual preprints of papers, offering a free alternative to costly journal versions of papers. High energy physicists worldwide rely on SPIRES and arXiv as a resource to perform research and access their colleagues' findings. The data in this paper is drawn from log files created over a six month period in 2009 consisting of all clicks made on the SPIRES website during that time, as well as the same data drawn from arXiv over the same six month period.

The SPIRES log files allow us to create different data sets to demonstrate various relationships between clicks and citations. About 87 percent of the links within the log file could be associated with a paper record identification number within the SPIRES system, while the remaining 13 percent were left aside, unable to be disambiguated. From the SPIRES record identification numbers, we are able to fetch various pieces of information about each paper clicked by a SPIRES user, such as its citation count, its topic, its year published, etc. Accordingly, we are able to plot different sets of the clicked papers. By plotting the number of clicks against the number of citations the paper received after the initial click, we can determine if any correlation between clicks and citations exist. To narrow the data in search of the specific click versus citation correlation, only papers that were published within three months of the recorded click were plotted; moreover, only their citations that occurred within the following year were plotted as well, to ensure that the citation followed a user accessing the paper and theoretically led to a citation (Figure 1). We generated the same plots for the arXiv data set. In order to reduce error, we compiled a list of the IP addresses drawn from each line of the log file, and removed computerized users such as Google Bot which had logged thousands of faux clicks.

In addition to plotting the clicks versus citations data set, we calculated the expected value of both clicks and citations for each point. This was calculated by determining the number of clicks versus its frequency (Figure 3) and the number of citations versus its frequency (Figure 4). From these two data sets, we were able to determine the expected value at each point on the original plot, and then found the difference between the actual and expected value, plotting the result (Figure 2). This allowed us to examine where the data deviates from the calculated expectation or null hypothesis — what we would expect to find should no correlation exist.

Finally, to find a statistical representation of the data, we calculated Pearson's rho a correlation coefficient that fluctuates between -1 and 1. Values near 0 indicate a lack of a correlation, while positive results correspond to a positive correlation and negative results correspond with negative values. In addition to calculating Pearson's rho for the data, we took a hundred sets of randomized data reassigning all x- and y-values to a new pair, recalculated the correlation coefficient, and averaged the resulting values. This would allow us to determine whether the resultant coefficient from the actual data is in fact remarkable in any way, or whether it is irrelevant — similar to the resultant coefficient from the randomized data.

RESULTS

Upon examination, the SPIRES and arXiv data revealed a consistent correlation between clicks and citations. Both data sets revealed remarkably similar plots. While the initial plot of the six months of SPIRES data (Figure 1) plotting only papers published within three months of their click and their citations within the next year looks mundane, the secondary plot displaying the difference between the actual and expected values clearly shows the correlation (Figure 2). It eliminates papers alongside both axes, dividing the data into three distinct areas the blue area along the x-axis, the blue area along the y-axis, and the central colored region. The area along the x-axis (or click-axis) represents an area of papers that are highly clicked but rarely cited, the area along the y-axis (or cite-axis) represents papers that are heavily cited but not often clicked, and the center region represents everything in between.

Pearson's correlation coefficient supports this finding. While the calculated coefficient for the six months of data was 0.34, which initially seemed too near to zero to indicate a correlation, the averaged, randomized data set consistently returned a correlation coefficient of 0 ± 0.002. This indicates that the method is not generating the correlation of 0.34, strengthening the idea of a correlation. To quantify this reading, we calculated a naive confidence interval of rho, which confirmed that zero was not within the confidence interval of the limits of rho, supporting a true, significant correlation.

CONCLUSION

While there are numerous outlier papers, the SPIRES and arXiv data still indicate a correlation between downloads and citations. Rather than simply clicking around, physicists utilize the papers they click and cite them in their own published research. This shows the unique role that SPIRES plays in the realm of HEP; SPIRES is an important tool for high-energy physicists to find relevant background for their own projects.

Proprietors of databases such as SPIRES must consider whether it is worthwhile to introduce a new form of competition for physicists. While many physicists place great weight on their citation counts, it can not be a true form of evaluating the worth, value, or influence of a paper or author. The same can be said of downloads. The addition of a new statistic to the competitive number game may in actuality produce extraneous rivalry within the field — an effect that may not be beneficial to physicists.

Future Work Additional work is needed to determine individual characteristics that define which papers lie within each of the three sections revealed in the difference plot. One strategy may be to plot papers according to topic, review, an author's overall cite count, or a number of other categories. Another statistical form of analyzing the data may be to calculate Spearman's rank, to pair with Pearson's rho and the confidence interval. A final means of quantifying an author's work, as proposed by John Beacom, would be to investigate plotting an author's accumulated citations within the past five years versus the author's overall citations for each year throughout the author's career.

FIGURES

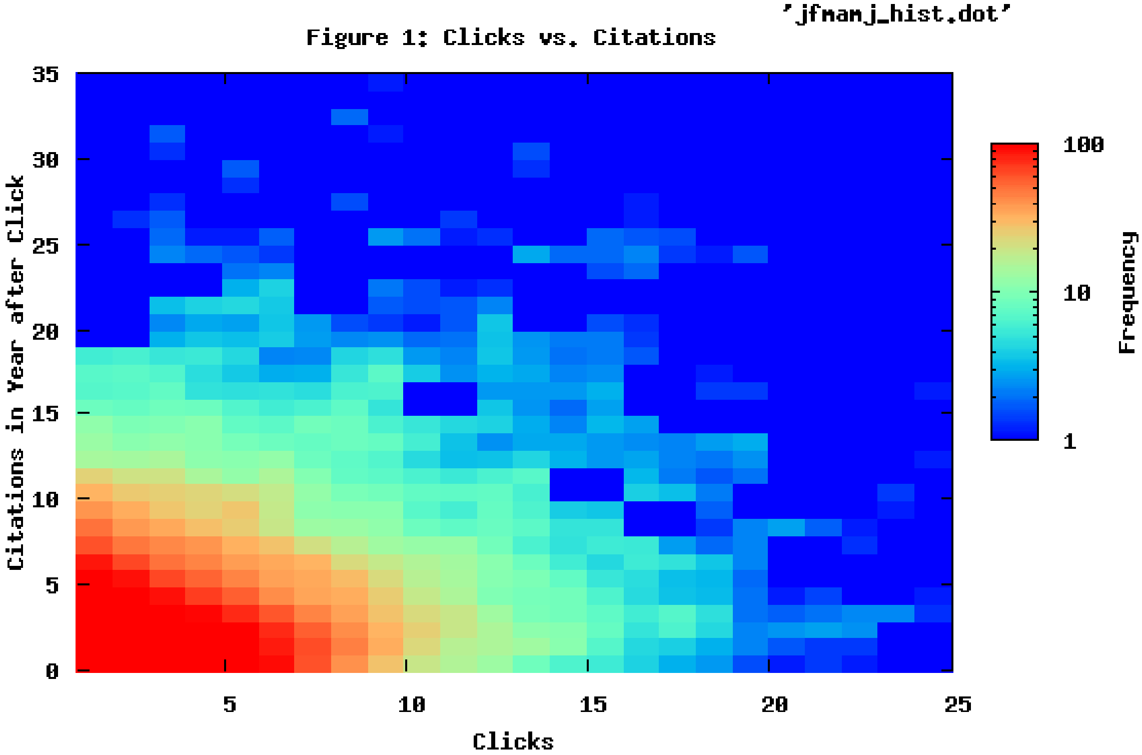

Figure 1

Clicks vs. Citations. Six months of 2009 SPIRES data narrowed to display only papers published within 3 months of the initial click and citations occurring in the year following the click.

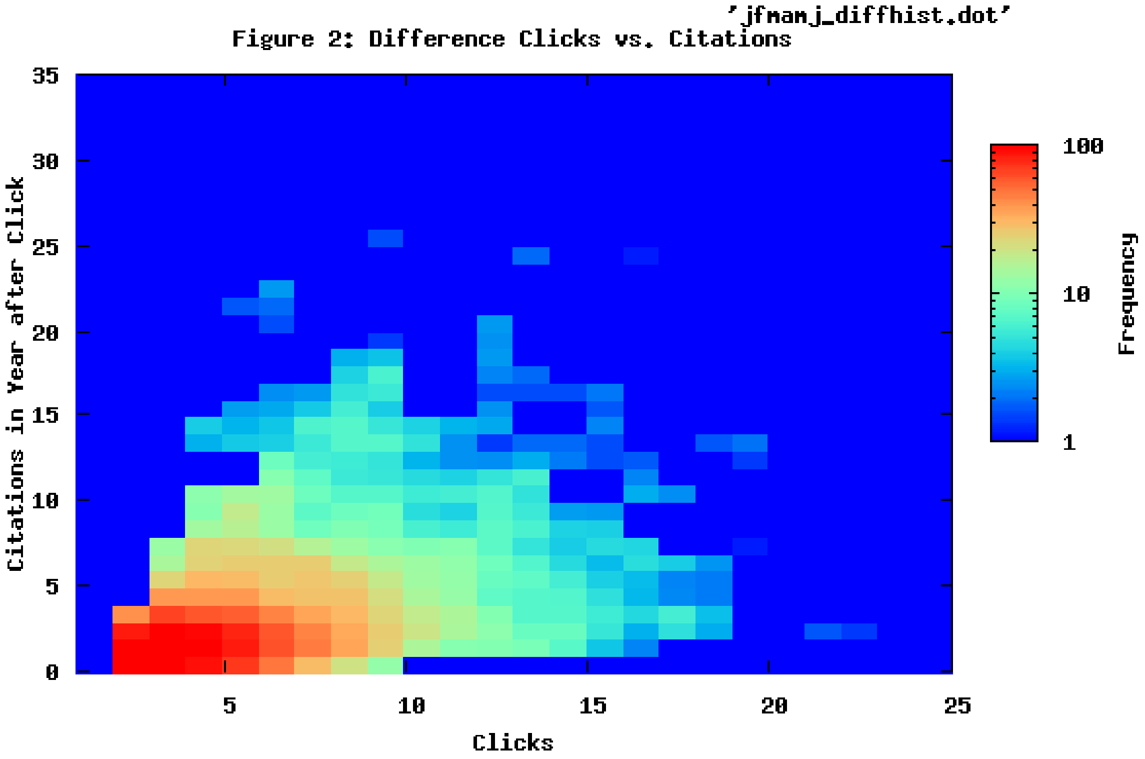

Figure 2

Difference of Actual Value and Expected Value Clicks vs. Citations. The difference of the six months of SPIRES data (shown in Figure 1) and the calculated expected value from the six month data set.

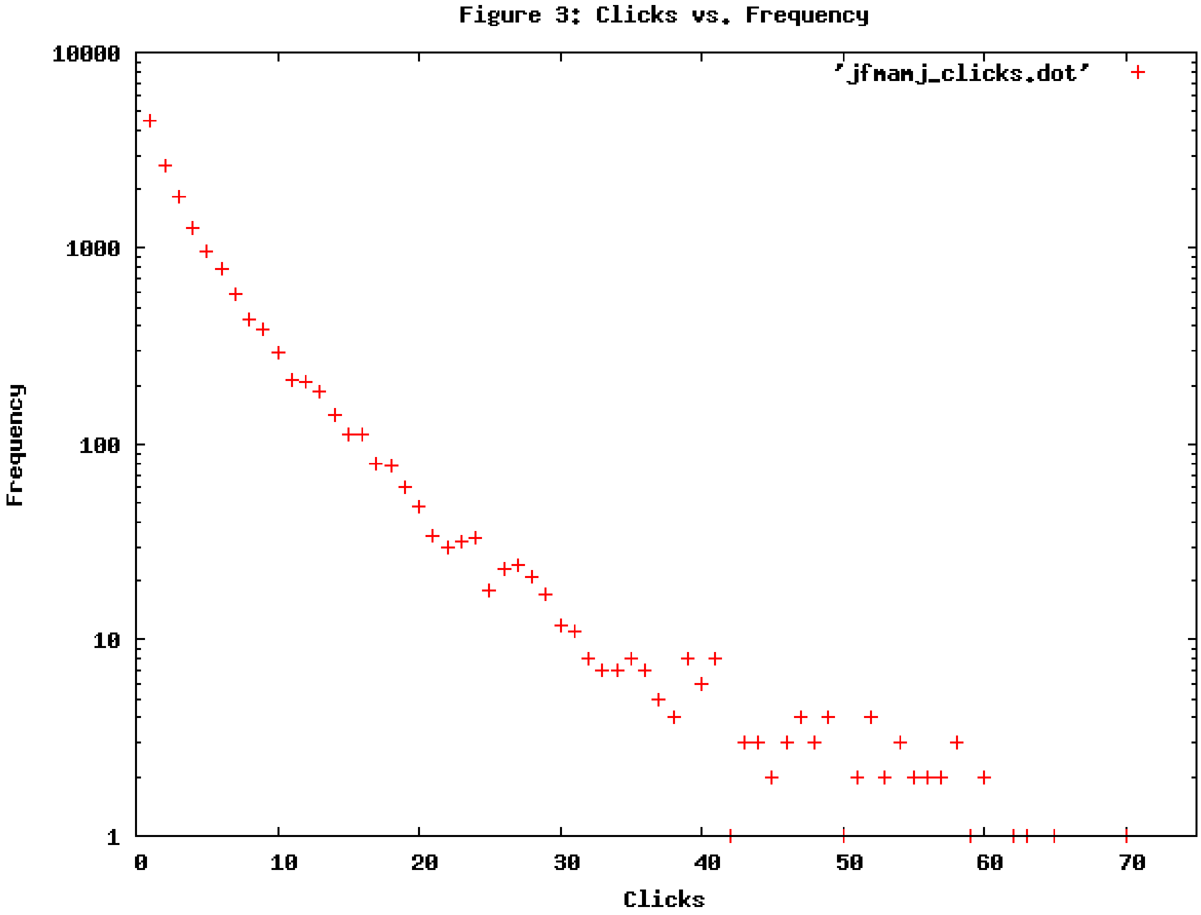

Figure 3

Clicks vs. Frequency. Used to calculate expected value.

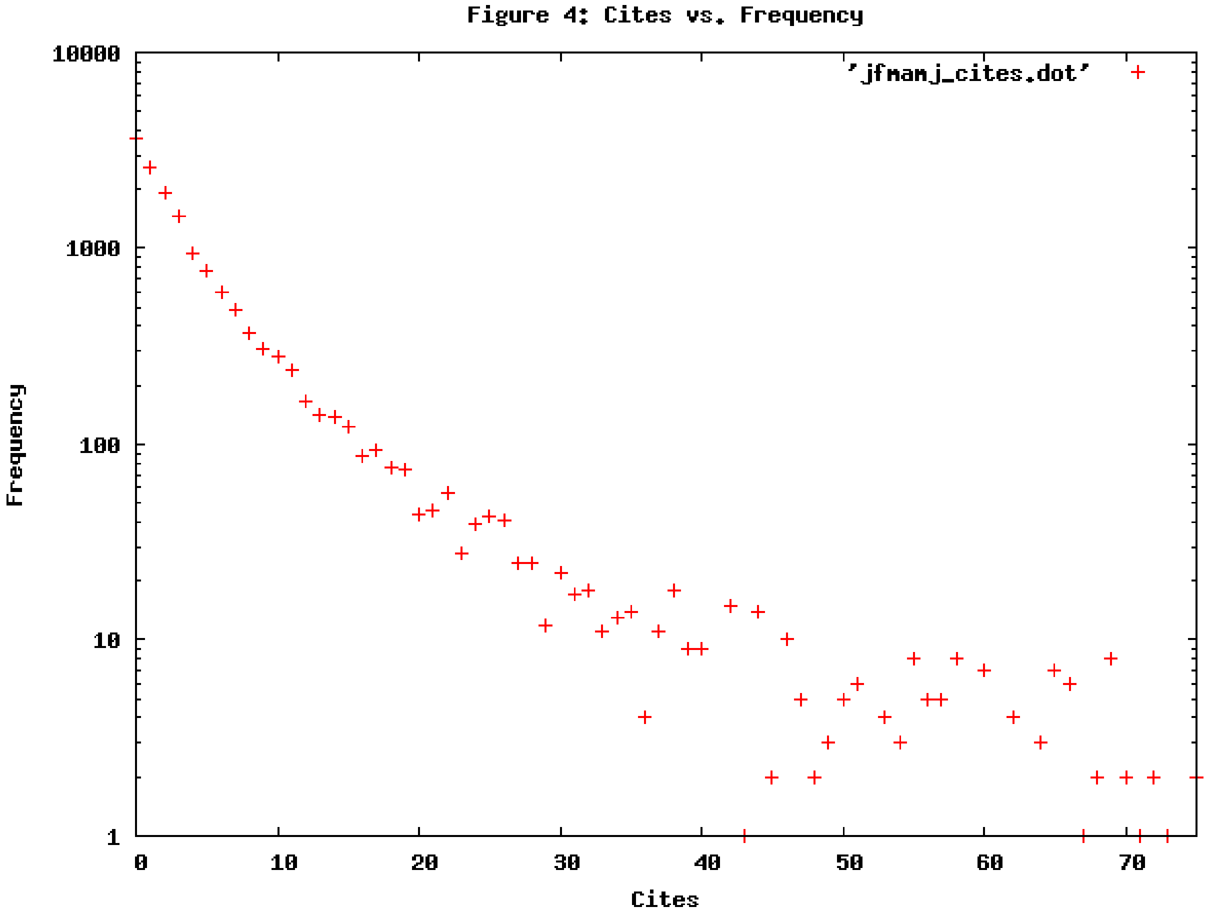

Figure 4

Cites vs. Frequency. Used to calculate expected value.

ACKNOWLEDGEMENTS

This research was conducted at SLAC National Accelerator Laboratory. I would like to thank my mentor, Travis Brooks, who has shown me the value of computer science and the sort of research I did, and has instilled in me an excitement for the field. I would also like to especially thank Joseph Blaylock for his consistent support, assistance, and teaching. Further thanks to the SPIRES staff both at SLAC and abroad. Finally, thank you to SLAC and the Department of Energy for supporting the SULI Program and offering me this invaluable

opportunity.

REFERENCES

[1] SPIRES Website http://www.slac.stanford.edu/spires [Last visited August 20, 2010]

[2] A. Gentil-Beccot, S. Mele, T. C. Brooks, “Citing and Reading Behaviours in High-Energy Physics. How a Community Stopped Worrying about Journals and Learned to Love Repositories,” Scientometrics **84**, 345 (2010). [arXiv:0906.5418 [cs.DL]]

[3] arXiv Website http://arxiv.org/ [Last visited August 20, 2010]

[4] Harnad, S. (2000). The invisible hand of peer review. Exploit Interactive. http://www.exploit-lib.org/issue5/peer-review/.

[5] SSRN Website http://www.ssrn.com/ [Last visited August 20, 2010]