Measuring the Early-Mover Advantage in High-Energy Physics Publication

Advait Sarkar
Social and Technological Network Analysis (L109) Project
February 25, 2013

ABSTRACT

The "first-mover" advantage is a well-documented effect in several fields whereby the first entrants in a competitive market benefit by virtue of their early timing, rather than the quality of their entry, an advantage which then prospers from a positive feedback cycle. We study a variant of this effect, which we call the "early-mover" advantage, in a large citation graph of high-energy physics publications spanning 10 years of citation activity. We investigate the relationship of publication time with three formalised notions of competitive advantage. For each we find weak but significant correlations, suggesting the presence of a slight early-mover advantage in high-energy physics publication. Additionally, this advantage is found to diminish with time.

1. INTRODUCTION

In *Networks of Scientific Papers* [1], Derek John de Solla Price demonstrated that the indegree distributions of citation graphs followed a power-law structure, where the vast majority of papers received few or no citations, but a "long tail" of a few papers received several citations. In other words, the citation graph is a *scale-free* network.

Price later postulated that this structure might be attributed to what he dubbed a "cumulative advantage" [2] process, wherein the rate at which a paper is cited is proportional to the number of citations it has. He mathematically formalised this process and was able to demonstrate that it generated a scale-free citation network. Several years later a similar process called "preferential attachment" was independently proposed by Barabási and Albert [3]. There is much empirical evidence from real-world networks that supports this [4, 5, 6, 7], but some networks have also been found to deviate from its predictions [8, 9, 10].

In this study we investigate the "first-mover" advantage enjoyed by early entrants into the competitive market of scientific publication. However, our treatment of the problem differs from previous work in the domain such as Newman's [11], the primary aim of which has been to provide empirical substantiation (or refutation) of the correctness of a formalised preferential attachment model by comparing its predictions to an existing network. Instead of starting with an observable advantage and proposing a citation growth model to explain it, we directly measured the first-mover advantage in a network. We define three distinct kinds of "advantage" and investigate their correlation with time of publication:

1. *Captured opportunity*, defined as the number of citations received within a certain time window, expressed as a fraction of the number of papers published in that window.
2. *Influence*, as measured by left eigenvector centrality over a threshold-trimmed graph.
3. *Value*, defined as the sum of citation values, where citations are weighted by their temporal proximity to the cited publication.

Another important departure from previous work is that we measured the advantage gained by one publication over another purely by virtue of how much earlier it was published, and *not* by virtue of how much longer it has existed. This distinction is made clearer in §2.2. In order to disambiguate between our approach and the model-evaluation approach of previous work, we use the term *early-mover advantage* to denote the effect from here onwards.

We used the citation graph from the high energy physics phenomenology ("HEP-PH") category of the e-print archive "arXiv" [12]. This dataset contains 34,546 papers and 421,578 edges. If a paper p_1 cites paper p_2, the graph contains a directed edge from p_1 to p_2. Therefore a node's outdegree is the number of papers it cites, and its indegree is the number of citations it has received. There is no information about citations to and from papers outside the dataset.

The data covers citation activity in the period from February 1992 to March 2002. It begins within a few months of the inception of arXiv, and thus represents essentially the complete history of its HEP-PH section up to 2002. The data was originally released as a part of the knowledge discovery and data mining competition "KDD Cup 2003" [13].

The date of publication for each paper is also known. Beyond this no further data or metadata is available about the authorship or the content of the paper. This makes any analysis related to authorship networks or semantic relevance impossible. Our analysis focused solely on the structure of the graph and its relationship to publication date.

2. STUDY

2.1 Preliminary Analysis

We began by removing duplicate entries in the file containing the dates of publication, attributable to cross-published papers, by only retaining the earliest date of publication recorded for each paper. This reduced the known publication dates by 6.1% from 38,556 to 36,201. It was found that 3,985 papers with known dates were not present in the citation graph, but this did not impact further analysis.

The citation graph has 61 weakly connected components, but the majority of nodes (99.6%) are contained in the giant component. Inspection of the indegree (Fig. 1) and outdegree (Fig. 2) distributions of the citation graph confirmed that it is indeed a scale-free graph. The early-mover advantage in this graph is the extent to which this degree distribution can be explained by the date of publication.

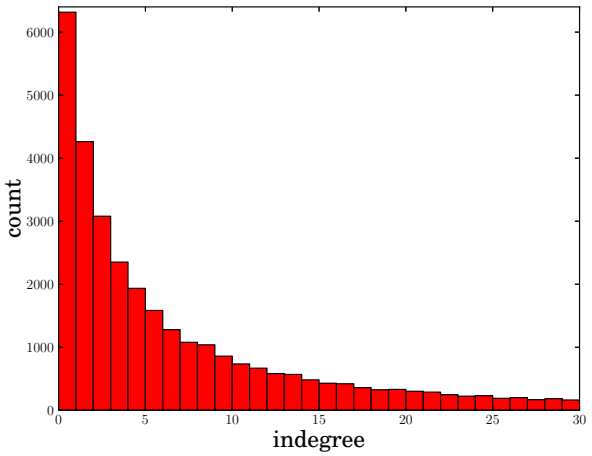

Figure 1: Histogram depicting the power law structure of indegree for the 30 most frequent counts.

We first normalised the publication dates. The earliest publication was 1992-02-11, and the latest publication was 2002-03-12, a span of 3682 days, or just over 10 years. We normalised dates on this scale to be a floating-point value between 0.0 and 1.0 to facilitate easier manipulation of the time variable. It is useful to note that on this time scale, a difference in publication dates of 0.1 corresponds approximately to one year.

2.2 Early-Mover Advantage for Captured Opportunity

Citation count is undoubtedly the *de facto* standard for measuring the worth of a scientific publication, routinely used to gauge the quality of research at universities [14, 15, 16].

Since the papers in this dataset have been published over the course of several years, it is not sound to directly measure the correlation between publication date and citation count, simply because older papers have had a much longer time in which to be cited. This conflates the advantage gained

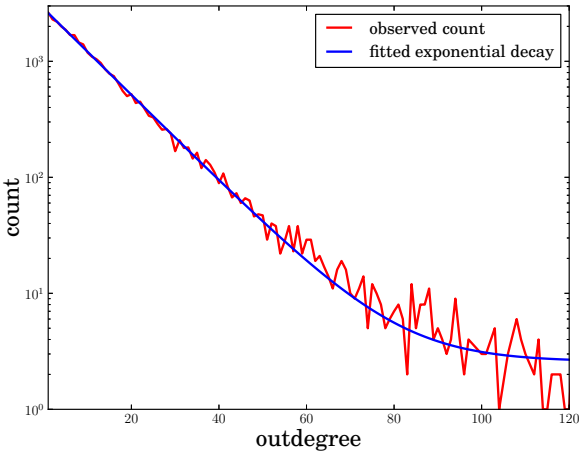

Figure 2: Lin-log plot depicting the power law structure of outdegree for the 120 most highly occurring counts. The curve has exponent $\alpha = -0.086$, and levels out towards the end due to a large constant.

by being published earlier with the advantage gained by existing for longer. We remove this bias by only counting the citations received within a fixed time window from the date of publication. To elaborate, let us say we set a threshold of t. Recall that an edge (p_1, p_2) means that p_1 cites p_2. We eliminate all edges (p_1, p_2) from the citation graph where $date(p_2) - date(p_1) > t$, thus eliminating any advantage gained purely by existing for longer. Therefore if our original citation graph was $\mathbf{G}$, we obtain our thresholded graph $\mathbf{G}_t$ for threshold t as follows:

$$edges\,(\mathbf{G}_t) = \{(p_1, p_2) \in edges(\mathbf{G}) \mid date(p_2) - date(p_1) \leq t\}$$

$$nodes\,(\mathbf{G}_t) = \{p \mid \exists (p_1, p_2) \in edges(\mathbf{G}).\ p = p_1 \vee p = p_2\}$$

However, there is still a bias in the indegrees of $\mathbf{G}_t$. Since the rate of papers published varies with time, newer papers often have a greater opportunity to be cited within the same window t than older papers. To eliminate this bias, we express the number of citations as a fraction of the number of papers published in the window t. We name this variable "captured opportunity". We use $captured\_opportunity_t(p)$ to denote the captured opportunity for a paper p at threshold t, and formalise it as as follows, using the abbreviation $d(p)$ for $date(p)$:

$$captured\_opportunity_t(p) = \frac{indegree(p)}{\left|\left\{p' \in nodes(\mathbf{G}_t) \mid \big(d(p') > d(p)\big) \wedge \big(d(p') \leq d(p) + t\big)\right\}\right|}$$

So, for example, if t is set to 1 year, and a paper p receives 3 citations after one year of publication, and in that year there were a total of 100 papers published, the captured opportunity for that paper is 0.03.

We correlated captured opportunity with time for the thresholds $t = 0.1$, $t = 0.25$, $t = 0.5$, and $t = 0.75$, corresponding

approximately to 1, 2.5, 5, and 7.5 years after publication respectively. We found small but highly significant negative correlations (Table 1). For $t = 0.1$ and $t = 0.25$, a stronger anticorrelation is obtained by taking the logarithm of the captured opportunity, indicating an exponential early-mover advantage, but for $t = 0.50$ and $t = 0.75$ the difference in correlation obtained by taking the logarithm is negligible. We performed least-squares regression and in each case discovered a negative slope.

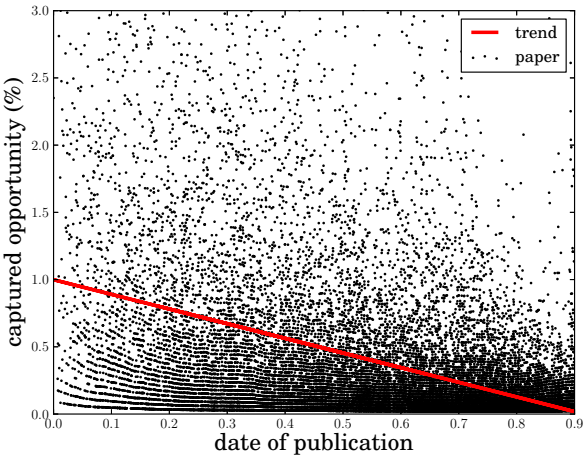

Figure 3: Temporal trend in captured opportunity for $t = 0.1$. Each dot represents a single paper.

Our analysis indicates that early publications do indeed have a small advantage over newer publications in terms of captured opportunity. Interestingly, the negative correlation and slope decreased as t increased, suggesting that the early-mover advantage gets weaker over time. This could be a result of the fact that as t gets larger, the range of dates available for measuring the correlation gets smaller. However, it may also indicate that good papers are eventually recognised, even if they are published later.

Threshold t	Pearson's r	Spearman's ρ
0.10	-0.353	-0.335
0.25	-0.244	-0.230
0.50	-0.167	-0.164
0.75	-0.096	-0.105

Table 1: Correlation of *log*(captured opportunity) with date of publication at different values of t. The 2-tailed p-values are negligible ($p < 10^{-8}$).

Threshold t	Regression slope
0.10	-1.09
0.25	-0.38
0.50	-0.22
0.75	-0.19

Table 2: Slope of least-squares regression line for *log*(captured opportunity) with date of publication at different values of threshold t.

2.3 Early-Mover Advantage for Influence

In §2.2 we expressed citation count as a fraction and used that as a measure for the impact of a paper. However, this is based on the assumption that all citations carry equal weight. This is not necessarily always the case. We might want to assign a greater weight to a well-cited paper than we give to a citation from a poorly-cited paper. We can take this a step further and recursively weight the citing papers themselves according to the same criterion.

This notion of influence, or "importance", is well-known to have been formalised as the *eigenvector centrality* measure. Specifically, for a directed graph, the *left* eigenvector centrality measures the importance or influence of a node i by considering those nodes which have a directed edge to i [17]. It is denoted $\varphi_1^L(i)$ and is given by the i^{th} entry of the principal left eigenvector of the adjacency matrix:

$$\varphi_1^L(i) = \left(\frac{1}{\lambda_1}\mathbf{A}^T\varphi_1^L\right)_i$$

where $\mathbf{A}$ is the adjacency matrix of the graph and λ_1 is the principal eigenvector of the adjacency matrix.

We correlated time with left eigenvector centrality as a measure of influence. As in the previous section, we thresholded the citations to ensure that older papers do not gain an artificial advantage. We performed correlation and regression at various thresholds.

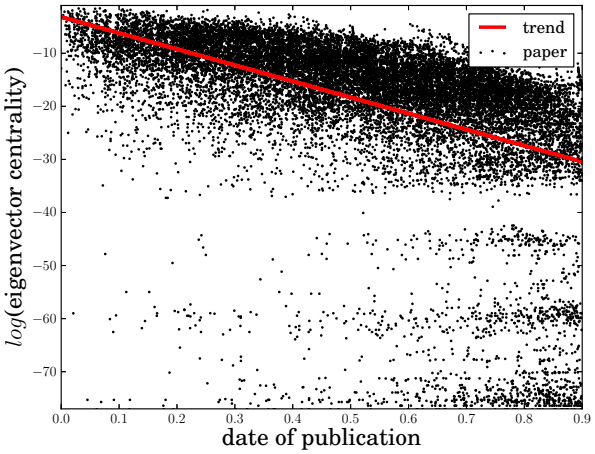

Figure 4: Influence trend for $t = 0.1$.

With influence, as with captured opportunity, a stronger anticorrelation is obtained by taking its logarithm. Our analysis shows a strong negative correlation at low values of threshold t. We again found that this anticorrelation becomes weaker as the threshold is increased, possibly as a result of the smaller sample size, but also possibly because good papers tend to be eventually recognised, suggesting that the influence of a paper that has been recently been published is not necessarily a good predictor of how influential it will eventually become.

Threshold t	Pearson's r	Spearman's ρ	Kendall's τ
0.10	-0.533	-0.526	-0.390
0.25	-0.379	-0.414	-0.293
0.50	-0.209	-0.279	-0.184
0.75	-0.142	-0.159	-0.107

Table 3: Correlation of log(influence) with date of publication. The p-values are negligible ($p < 10^{-11}$).

Threshold t	Regression slope
0.10	-20.33
0.25	-17.29
0.50	-14.85
0.75	-19.73

Table 4: Slope of least-squares regression line for log(influence) with date of publication at different values of threshold t.

2.4 Early-Mover Advantage for Citation Value

In §2.3 we established that all citations are not equal. Using eigenvector centrality to weight the contribution of citations is one way of addressing this inequality. However, indirect citations are usually not valuable to the author of the original paper. Another simple way of weighting the value of a citation to a researcher is by how soon the citation appears after the publication. All else being equal, getting cited earlier is more valuable than getting cited later.

We use a basic definition of the value of a citation:

$$value_c(p_1, p_2) = 1 - [date(p_2) - date(p_1)]$$

This notion of value simply inverts the date difference between the citing paper and the cited paper. Recall that date is expressed on a scale of 0 to 1. As a consequence, citations which occur very soon after a paper is published have a value close to 1, and citations which occur 10 or more years (the period under investigation) after the paper is published have values close to 0. In reality, very late citations still have value to the researcher. In future work, it would be interesting to explore the use of more sophisticated penalty curves, such as a heavy-tailed exponential. However, this is beyond the scope of the current investigation.

Finally, we define the value of a paper to be the sum of values for all its citations in the thresholded graph:

$$value(p) = \sum_{(p',p) \in edges(\mathbf{G}_t)} value_c(p', p)$$

and study the relationship of this "value" to date of publication in a manner similar to the previous studies.

Unlike with influence and captured opportunity, taking the logarithm of value does not yield a much stronger anticorrelation. This shows that while being an early-mover is exponentially advantageous in terms of citation capture or influence, it is only linearly advantageous in terms of how quickly the paper is cited. Again, there is a negative correlation at low values of threshold t. Furthermore this anticorrelation also becomes weaker as the threshold is increased, possibly as a result of the smaller sample size. The pattern of striations towards the end of the time period in Figure 5

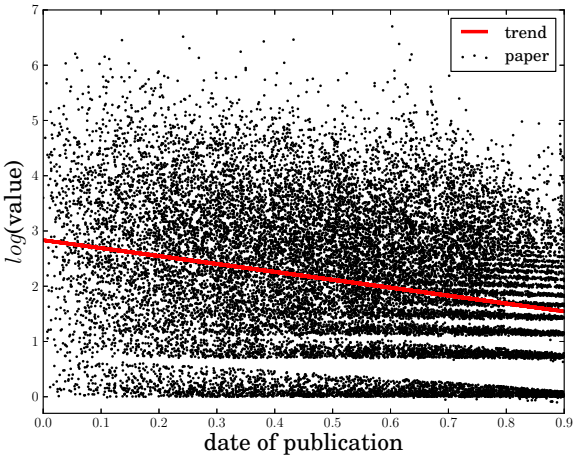

Figure 5: Value trend for $t = 0.1$.

is interesting; it indicates that in the initial 2-3 years after publication, papers fall into discrete categories according to their value to the researcher, but thereafter assume a more or less uniform distribution. Judging by the lower striations, it does not appear as though papers often 'escape' the category in which they begin. This is to be expected, as by our definition the highest-value citations are only available to a paper in its initial years after publication.

Threshold t	Pearson's r	Spearman's ρ	Kendall's τ
0.10	-0.254	-0.253	-0.172
0.25	-0.160	-0.154	-0.103
0.50	-0.115	-0.108	-0.073
0.75	-0.070	-0.066	-0.044

Table 5: Correlation of log(value) with date of publication. The p-values are negligible ($p < 10^{-5}$).

Threshold t	slope	slope (log)
0.10	-32.63	-1.43
0.25	-30.99	-1.09
0.50	-42.61	-1.23
0.75	-59.14	-1.54

Table 6: Slope of least-squares regression lines for value and log(value) with date of publication.

3. DISCUSSION

In each of the three notions of advantage studied, there is a weak but significant anticorrelation with publication date. This indicates the presence of a slight early-mover advantage. The strongest anticorrelations were found with influence as measured by left eigenvector centrality. This is expected, since indirect citations count towards our measure of influence, and older papers have had a greater opportunity to build long citation chains, especially if they are cited by heavily influential papers.

Furthermore, we found that the anticorrelation decreases as the threshold for citation inclusion is increased, illustrated in Figure 6. There are two plausible explanations for this:

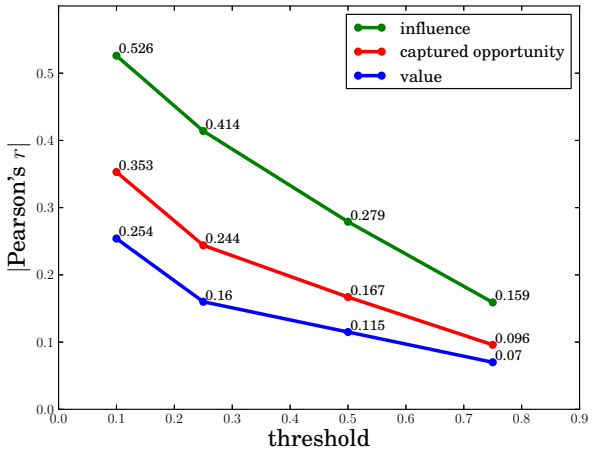

Figure 6: Decay of anticorrelation with threshold.

1. As t gets larger, a smaller range of dates is available for computing the correlation. For instance, $t = 0.75$ represents a window of approximately 7.5 years. Since our entire dataset spans 10 years, we only have 2.5 years of papers for which we can conduct analysis of a 7.5 year window. This amplifies the effect of variables other than publication date on the variable being studied (captured opportunity, influence, value, etc.) thus weakening the correlation, or

2. The early-mover advantage diminishes with time. This would mean that, fortunately, good papers are eventually recognised, even if they are published later. Conversely, papers which benefit from a strong early-mover advantage, despite the fact that their popularity is not justified by their content, do not retain this advantage. Consequently, it may be said that the performance of a paper in the initial 2-3 years after publication is not necessarily a good predictor of how it will perform in the long term.

4. CONCLUSION

We empirically investigated the early-mover advantage in high-energy physics publications by measuring the correlation between publication date and three distinct notions of advantage: opportunity capture, influence, and citations weighted by temporal proximity. In each case we found small but significant correlations, indicating a slight early-mover advantage. Furthermore, we found evidence to suggest that this advantage weakens over time.

Consequently, if a researcher is interested in maximising the short-term popularity of their publications, they might consider publishing extensively in an emerging field, instead of intensively in an established field. However, this strategy may not be as profitable in the long term.

Acknowledgments

Many thanks to Anastasious Noulas for clarifications regarding the dataset.

References

[1] Derek John de Solla Price. Networks of scientific papers. *Science*, 149:510–515, 1965.

[2] Derek John de Solla Price. A general theory of bibliometric and other cumulative advantage process. *Journal of the American Society of Information Science*, 27:292–306, 1976.

[3] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[4] Paul L Krapivsky and Sidney Redner. Organization of growing random networks. *Physical Review E*, 63(6):066123, 2001.

[5] Sergey N Dorogovtsev, AV Goltsev, and José Ferreira F Mendes. Pseudofractal scale-free web. *Physical Review E*, 65(6):066122, 2002.

[6] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47, 2002.

[7] MEJ Newman. The size of the giant component of a random graph with a given degree sequence. *SIAM Review*, 45(2):167–256, 2003.

[8] Katy Börner, Jeegar T Maru, and Robert L Goldstone. The simultaneous evolution of author and paper networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5266–5273, 2004.

[9] Sune Lehmann, AD Jackson, and B Lautrup. Life, death and preferential attachment. *EPL (Europhysics Letters)*, 69(2):298, 2007.

[10] Sidney Redner. Citation statistics from 110 years of physical review. *Physics Today*, 58(6):49, 2005.

[11] MEJ Newman. The first-mover advantage in scientific publication. *EPL (Europhysics Letters)*, 86(6):68001, 2009.

[12] *arXiv.org e-Print archive.* `http://arxiv.org` (as of March 6, 2013).

[13] *KDD Cup 2003.* `http://www.cs.cornell.edu/projects/kddcup/` (as of March 6, 2013).

[14] *Rating Universities on Research Quality.* `http://www.topuniversities.com/qs-stars/rating-universities-research-quality-qs-stars` (as of March 6, 2013).

[15] *Methodology of ARWU.* `http://www.shanghairanking.com/ARWU-FIELD-Methodology-2012.html` (as of March 6, 2013).

[16] *THE World University Rankings.* `http://www.timeshighereducation.co.uk/world-university-rankings/2012-13/world-ranking/methodology` (as of March 6, 2013).

[17] Ernesto Estrada. *The structure of complex networks: theory and applications.* Oxford University Press, 2011.