A Stochastic Prediction for the h-index Utilizing Price’s Preferential Attachment Model

Tyler Schuessler
Tyler.Schuessler@Colorado.EDU

Follow this and additional works at: https://scholar.colorado.edu/honr_theses

Part of the Other Applied Mathematics Commons, Other Statistics and Probability Commons, and the Probability Commons

Recommended Citation
https://scholar.colorado.edu/honr_theses/1906
HONORS THESIS

A Stochastic Prediction for the h-index Utilizing Price’s Preferential Attachment Model

Tyler Schuessler
Department of Applied Mathematics
University of Colorado - Boulder

Defended on Friday, April 5 2019

Committee Members
Thesis Advisor: Manuel E. Lladser, Department of Applied Mathematics (APPM)
Anne Dougherty, College of Arts and Sciences Honors Program & APPM
Orit Peleg, Department of Computer Science (CSCI)
A Stochastic Prediction for the h-index Utilizing Price’s Preferential Attachment Model

Tyler Schuessler

Abstract

The h-index, or Hirsch index, is a bibliometric index intended to measure the impact of a researcher’s scholarly output. It is defined as the largest $k$ such that an author has $k$ papers with $k$ or more citations each. Motivated by our desire to understand anomalous researchers, we aim to characterize the asymptotic probability distribution of the h-index for a researcher with $m$ papers in a random citation network of $n$ papers described by Price’s preferential attachment model. Furthermore, we analyze our results implementing both synthetic and real citation networks.

Table of Contents

1 Introduction 2
2 The h-index Model 3
   2.1 Price’s Preferential Attachment Citation Model ........................................... 3
   2.2 Expected Citation Count ................................................................. 3
   2.3 Scaling Limit for Expected Citations .................................................... 4
   2.4 Stochastic Prediction for the h-index ..................................................... 5
3 Synthetic Data 7
   3.1 Simulation Methodology ................................................................. 7
   3.2 Synthetic Researchers ................................................................. 7
   3.3 Prediction Analysis for Synthetic Citation Networks ................................. 8
4 arXiv Data Analysis 9
   4.1 Description and Methodology ............................................................ 9
   4.2 Fitting Model Parameters ............................................................... 10
   4.3 Researchers Publication Frequency .................................................... 12
   4.4 Researchers with a Given Number of Publications ............................... 13
   4.5 Synthetic Researchers with a Given Number of Publications .................. 14
   4.6 The h-index for an Arbitrary Number of Publications ........................... 15
   4.7 Expected h-indices for Given Number of Publications ............................ 17
5 Conclusion 18
A Appendix 21
   A.1 Additional Figures for Section 4.3 .......................................................... 21
   A.2 Additional Figures for Section 4.4 .......................................................... 23
Introduction

The h-index was first proposed in 2005 by J. E. Hirsch to provide quantification for a scholar’s research impact [1]. He argued the harmony of a researcher’s per publication citation count and total number of publications mitigated biases that previous metrics had not. Advantageously, the h-index is a one-dimensional quantity, making comparisons between researchers straightforward.

The h-index is defined as the largest index, $k$, such that a researcher has $k$ publications all with a minimum of $k$ citations apiece. For example, a researcher with an h-index of 4 has 4 papers all with at least 4 citations apiece; however, all their remaining papers have strictly fewer than 5 citations each.

Algorithmically, computing the h-index involves ordering an author’s papers by decreasing citation count and then finding the largest index such that the citation count of that paper is greater than or equal to its associated index. Specifically, for an author to have an h-index of $k$, necessarily $k$ of their papers must have at least $k$ citations each, while all remaining papers have less than $(k + 1)$ citations each.

While criticisms of the h-index do exist and modifications or novel metrics have been considered, the use of the h-index has become somewhat ubiquitous. Various studies have been conducted regarding the outright applicability and the comparative advantage of the h-index [2, 3, 4, 5, 6, 7]. The consensus among this research suggests that the h-index is a valid tool for evaluating a researcher’s impact; however, these studies also commonly conclude the h-index has some inherent shortcomings. Much of the criticism arises from the h-index’s inability to highlight the performance of those who do not necessarily conform to the prototypical or ideal researcher Hirsch had in mind. Bornman and Daniel, for example, argue methodical researchers who publish fewer papers but obtain high per paper citation counts have h-indices capped by their comparatively small number of papers [3]. For this reason, they conclude that the h-index can mistakenly equate truly impactful researchers with the less qualified, simply because the former lack an abundance of publications. Costas and Bordons have similar conclusions; the h-index might promote quantity of publication rather than quality as to not limit one’s own h-index [4]. Some other criticisms derive from a researcher’s ability to artificially inflate their own h-index, either through frequent co-authorship or citing oneself [5, 6].

Some prevalent solutions include utilizing alternative indices [6] and using the h-index in conjunction with another metric [1, 2, 4, 5]. Many of these recently proposed bibliometrics are variants or extrapolations of the h-index itself. Comparisons with these by-product metrics are used to inform some of the aforementioned conclusions in [2, 4, 5, 6]. Other comparisons, using less derivative metrics, seem to have some hotly contested results [8]. For instance, Hirsch in 2007 defended his h-index against the mean citation count of an author. He argued that the h-index outperforms it and other simple metrics like total citation count and cumulative number of publications [1, 8].

The pervasiveness of the h-index in today’s research climate, coupled with the convoluted nature of its merits and demerits, motivate us to describe a baseline behavior for h-indices within a field of research. Therefore, we frame an h-index probability distribution for an author with $m$ publications in a random citation network informed by Price’s preferential attachment model. It is our hope that stochastic analysis will aid in more robust classification of a researcher’s performance as it relates to their h-index.
We first endeavor to interpret the citation behavior for a singular paper, independent of its associated researcher. Thus, in Sections 2.1-2.3, we study the stochastic nature of citations based on paper aggregations which constitute networks. In Section 2.4, we complete our prediction for the h-index distribution, turning to researchers’ behavior informed by our previous analysis regarding citations. Then, we analyze our results by comparison. This is done with synthetic citation networks (Section 3) and is then supplemented with real data driven citation networks (Section 4).

2 The h-index Model

2.1 Price’s Preferential Attachment Citation Model

We aim to understand how an individual paper accumulates citations in a research environment. To do so, we consider a random citation network in which the nodes are papers and (directed) edges are citations. We assume that papers in a field accumulate citations following Derek J. de Solla Price’s model [9, 10].

Price argued that cumulative advantage—or by its contemporary name, preferential attachment—captures the stochastic citation dynamics for a network of scientific papers. Preferential attachment is a stochastic model describing the probability that existing papers in a citation network obtain further citations: each citation a new paper makes is to an existing paper with a probability proportional to the number of citations that paper already possesses.

Price’s preferential attachment model has two parameters: \( c \) and \( a \). \( c \) is the average number of citations a paper makes upon publication, and \( a \) is the automatic number of citations (also called pseudo-citations) a paper inherits independently of other papers in the network. For simplicity, in what follows we assume that each paper makes exactly \( c \) citations.

As the network model is dynamic, it evolves through the publication of new papers. A paper is immediately considered part of the network upon publication, at which time it obtains its “automatic” \( a \) citations, and becomes referenceable itself. Since papers are added successively, a paper is uniquely defined by its temporal index, i.e. a number \( j \in \{1, \ldots, n\} \), where \( n \) is the generic number of papers in the network. Throughout the paper, we often refer to this index as the paper number or paper index. In mathematical terms, the probability the \( j \)-th paper obtains the \( i \)-th citation \((i \in \{1, \ldots, c\})\) made by the \( n \)-th paper is

\[
C_{j,n-1} + a\frac{1}{(c + a)(n - 1)},
\]

where \( C_{j,n-1} \) is number of citations the \( j \)-th paper has obtained after the \((n - 1)\)-st publication.

2.2 Expected Citation Count

Utilizing Price’s model, we next resolve to determine the expected number of citations the \( j \)-th paper acquires in a network comprised of \( n \) publications.

Again, let \( C_{j,n-1} \) be the random number of citations the \( j \)-th paper collects from the first \((n - 1)\) publications in the network. Note that

\[
\frac{C_{j,n-1} + a}{(c + a)(n - 1)} = p \frac{C_{j,n-1}}{c(n - 1)} + (1 - p) \frac{1}{(n - 1)},
\]

(1)
where \( p := \frac{c}{(c+a)} \) and hence \((1 - p) = \frac{a}{c+a} \). Equation (1) is a strategic dissection of Price’s governing principle, and has been employed to simulate this model efficiently [11]. For the purpose of simulation, this equation allows citations to materialize in two distinct ways; a paper added to the network makes a citation preferentially with probability \( p \), and non-preferentially with probability \((1 - p)\). Following this principle, the additional citations the \( j \)-th paper receives upon the publication of the \( n \)-th paper—conditional on \( C_{j,n-1} \)—has the distribution

\[
(C_{j,n} - C_{j,n-1}) \mid C_{j,n-1} = \text{Binomial} \left( c, p \frac{C_{j,n-1}}{c(n-1)} + (1 - p) \frac{1}{n-1} \right).
\]

Let us define \( E_{j,n} := E(C_{j,n}) \) and \( b_n := \frac{p}{n} \). From the rules of conditional expectation, it follows from equation (2) that

\[
E_{j,n} = a \cdot b_{n-1} + (1 + b_{n-1}) \cdot E_{j,n-1}.
\]

And recursively expanding the terms of (3), we find

\[
E_{j,n} = ab_{n-1} + ab_{n-2}(1 + b_{n-1}) + (1 + b_{n-1})(1 + b_{n-2})E_{j,n-2}
\]

\[
= ab_{n-1} + ab_{n-2}(1 + b_{n-1}) + ab_{n-3}(1 + b_{n-1})(1 + b_{n-2}) + (1 + b_{n-1})(1 + b_{n-2})(1 + b_{n-3})E_{j,n-3}
\]

\[
= ab_{n-1} + \cdots + ab_{j}(1 + b_{n-1}) \cdots (1 + b_{j+1}) + (1 + b_{n-1}) \cdots (1 + b_{j+1})(1 + b_{j})E_{j,j}.
\]

Since the \( j \)-th paper cannot possibly collect any citations from other papers until the \((j + 1)\)-st publication, \( E_{j,j} = 0 \). Therefore, compactly stated

\[
E_{j,n} = a \sum_{i=0}^{n-(j+1)} b_{n-(i+1)} \left( \prod_{k=1}^{i} (1 + b_{n-k}) \right),
\]

where \( \prod_{k=1}^{0} (1 + b_{n-k}) := 1 \).

Equation (4) provides an explicit, yet problematic, expression for the number of citations the \( j \)-th paper is expected to obtain in a citation network consisting of \( n \) papers. In the following section, we endeavor to produce a more tractable version of this result.

### 2.3 Scaling Limit for Expected Citations

For a sufficiently large number \( n \) of papers and \( 1 \leq j \leq n \), define \( t := \frac{j}{n} \) and the function \( g(t) := E_{j,n} \). The auxiliary variable \( t \) represents an index rescaling of all papers in the network, such that the new index range is between 0 and 1. Explicitly, paper index increments are now of length \( \frac{1}{n} \), which tend to 0 as \( n \) increases. Thus, we can expect

\[
g'(t) = \lim_{n \to \infty} \frac{E_{j,n} - E_{j+1,n}}{1/n}.
\]

Before the following analysis, we quickly recall two useful properties of the Gamma function:

\[
\text{(a) } x = \frac{\Gamma(x+1)}{\Gamma(x)} \text{ for } x > 0; \text{ and}
\]

\footnote{We remark that, in principle, Price’s model allows a new publication to cite an extant paper multiple times. As the size of the network increases this possibility becomes progressively less likely.}
distribution of the process is then given by the uncommon to model the random arrival times of a stochastic process in this manner. Conditioning on the smallest (expected) citation count. Naturally then, to have the greatest (expected) citation count, and correspondingly, the paper with the largest index to have

\[ E_{j,n} - E_{j+1,n} = a \sum_{i=0}^{n-(j+1)} b_{n-(i+1)} \left( \prod_{k=1}^{i}(1 + b_{n-k}) \right) - a \sum_{i=0}^{n-(j+2)} b_{n-(i+1)} \left( \prod_{k=1}^{i}(1 + b_{n-k}) \right) \]

Consequently, using property (b) in equation (5) we obtain

\[ g'(t) \approx \lim_{n \to \infty} \frac{ap}{t} \frac{\Gamma(n+p)}{\Gamma(n)} \frac{\Gamma(tn+1)}{\Gamma(tn+1+p)} \]

\[ = \lim_{n \to \infty} \frac{ap}{t} \frac{1}{(tn+1)^p} \]

\[ = \frac{ap}{tp+1}. \]

\[ E_{j,j} = 0 \] implies \( g(1) = 0 \); and thus, in accordance with the Fundamental Theorem of Calculus, we find

\[ g(t) = -\int_t^1 \frac{ap}{x^{p+1}} dx = ap(t^{-p} - 1), \]

for \( t \in (0, 1] \). Equation (6) is an clear and manageable representation of \( E_{j,n} \), where \( j = tn \), which permits us to estimate any particular paper’s expected citation count within a sufficiently large network.

2.4 Stochastic Prediction for the h-index

We now address an individual researcher’s h-index within a citation network. We denote \( N \) as the total number of publications authored by the researcher.

Imagine the researcher has authored \( N = m \geq 1 \) publications, it becomes reasonable to model the indices of these papers as i.i.d. \( U'_1, \ldots, U'_m \sim Uniform\{1, \ldots, n\} \). Applying the scaling discussed in the previous section, each \( U_i := \frac{U'_i}{n} \) should be approximately \( Uniform(0, 1) \). \(^2\) In what follows, we assume that \( U_1, \ldots, U_m \) are i.i.d. \( Uniform(0, 1) \).

Since \( g(t) \) is strictly decreasing on its domain, our model predicts an author’s paper with the smallest index to have the greatest (expected) citation count, and correspondingly, the paper with the largest index to have the smallest (expected) citation count. Naturally then, \( g(U_{(k)}) \) can be used as a proxy for the number of

\(^2\)Another motivation for assuming uniformly distributed paper indices is its relation to the Poisson point process. It is not uncommon to model the random arrival times of a stochastic process in this manner. Conditioning on \( m \) arrivals, the arrival time distribution of the process is then given by the \( m \) uniform order statistics.
Given that the probability that not one of an author’s citations the $k$-th paper obtains. Hence, this researcher has an h-index of $k$, if and only if, $g(U_{(k)}) \geq k$ and $g(U_{(k+1)}) < (k + 1)$. Following this logic, we obtain, for $1 \leq k < m$,

$$\mathbb{P}(\text{h-index} = k | N = m) = \mathbb{P}\left(g(U_{(k)}) \geq k, g(U_{(k+1)}) < k + 1\right)$$

$$= \mathbb{P}(U_{(k)} \leq g^{-1}(k), U_{(k+1)} > g^{-1}(k + 1))$$

$$= 1 - \mathbb{P}\left(U_{(k)} > g^{-1}(k) \text{ or } U_{(k+1)} \leq g^{-1}(k + 1)\right)$$

$$= 1 - \mathbb{P}\left(U_{(k)} > g^{-1}(k)\right) - \mathbb{P}\left(U_{(k+1)} \leq g^{-1}(k + 1)\right)$$

$$= \mathbb{P}\left(U_{(k)} \leq g^{-1}(k)\right) - \mathbb{P}\left(U_{(k+1)} \leq g^{-1}(k + 1)\right).$$

Similarly, for $k = m$

$$\mathbb{P}(\text{h-index} = m | N = m) = \mathbb{P}\left(g(U_{(m)}) \geq m\right) = \mathbb{P}\left(U_{(m)} \leq g^{-1}(m)\right).$$

Note that

$$\sum_{k=1}^{m} \mathbb{P}(\text{h-index} = k | N = m) = \mathbb{P}\left(U_{(1)} \leq g^{-1}(1)\right),$$

which implies

$$\mathbb{P}(\text{h-index} = 0 | N = m) := 1 - \mathbb{P}\left(U_{(1)} \leq g^{-1}(1)\right).$$

Intuitively, this last formulation not only accounts for the missing probability mass, but simply describes the probability that not one of an author’s $m$ papers has a single citation, i.e. $\mathbb{P}\left(g(U_{(1)}) < 1\right)$.

Given that $U_{(k)} \sim \text{Beta}(k, m - k + 1)$, we can define

$$p_k := \mathbb{P}\left(U_{(k)} \leq g^{-1}(k)\right) = \int_{0}^{g^{-1}(k)} \frac{m! u^{k-1} (1 - u)^{m-k}}{(k-1)! (m-k)!} \, du.$$ 

Summarily,

$$\mathbb{P}(\text{h-index} = k | N = m) = \begin{cases} 
(1 - p_1) & k = 0 \\
(p_k - p_{k+1}) & 1 \leq k < m \\
p_m & k = m 
\end{cases} \quad (7)$$

The favorable structure of (7) elicits a very straightforward expected h-index formulation for researchers with $m$ publications:

$$\mathbb{E}(\text{h-index}|N = m) = \sum_{k=1}^{m-1} k \cdot (p_k - p_{k+1}) + m \cdot p_m$$

$$= \sum_{k=1}^{m} k \cdot p_k - \sum_{k=1}^{m-1} k \cdot p_{k+1}$$

$$= \sum_{k=1}^{m} k \cdot p_k - \sum_{k=1}^{m-1} (k + 1) \cdot p_{k+1} + \sum_{k=1}^{m-1} p_{k+1}$$

$$= \sum_{k=1}^{m} p_k.$$ 

Throughout the remainder of this paper, we evaluate the above h-index predictions, analyzing applicability for both synthetic and real citation networks.
3 Synthetic Data

3.1 Simulation Methodology

In order to accede our model we simulate random citation networks utilizing Price’s preferential attachment model. Through this, it is possible to compare our theoretical results with relevant distributions arising from synthetic citation networks. As described in Section 2.1, a citation network is constructed by incrementally adding new papers to the existing network. Each additional paper generates \( c \) citations. Every paper which is already within the network has a distinct probability of obtaining any of these \( c \) citations. To reiterate, this particular probability, as given in equation (1), is proportional to the number of citations which the aforementioned paper has already collected.

Throughout the remainder of this section, the synthetic networks we reference follow Price’s preferential attachment model. The parameters of any particular network are \( c = 15 \), \( a = 1 \), and are contained at 1000 papers. Additionally, in this Section there are instances in which we refer to an average network. An average network is realized by synthesizing a number of citation networks. Then, for each particular index \( j \), the citation count is averaged across all of these synthetic networks.

As seen in Figure 1a, our formula for the expected citation count, \( E_{j,n} = g \left( \frac{j}{n} \right) \), follows the simulated citation trend. But, there is considerable uncertainty in regards to exclusively one network. In contrast, as seen in Figure 1b, this uncertainty is significantly reduced in respect to an average citation network comprised of 100 synthetic networks. Hence, it is reasonable to let \( g(t) \) serve as a surrogate for the citation count of papers.

![Figure 1: (a) Citation counts and expected number of citations for one synthetic network. (b) Average citation counts and expected number of citations for an average network.](image)

3.2 Synthetic Researchers

The next logical step is to consider researchers’ behavior within our synthetic networks. In adherence to the assumptions made in Section 2.4, an author with \( m \) publications’ paper indices are reasonably modeled by i.i.d. \( \text{Uniform}\{1, \ldots, n\} \). Functionally, for simulation, all researchers are therefore realized simply as subsets of \( m \) papers drawn with replacement from the \( n \) papers in the network.

Figure 2, depicts the citation accrual for 30 papers or one simulated author. As expected, in general the earlier a paper was published (i.e. the smaller the paper number), the larger its citation count is.
3.3 Prediction Analysis for Synthetic Citation Networks

To analyze our theoretical predictions for $\Pr(h\text{-index} = k|N = m)$ we conceive 1000 synthetic researchers, for each relevant $m$, in the manner described in Section 3.2. Tabulating the h-indices for every synthetic author, we obtain histograms representing the true h-index distribution. The comparisons are shown in Figure 3 for $m \in \{11, 22, 33\}$, and our two distinct types of synthetic citation networks: a single realized synthetic network and an average citation network.

The h-index distributions for these synthetic researchers in a single simulated citation network are given in Figures 3a, 3c, and 3e. From this, it is evident that our theoretical predictions capture the true h-index distribution structure reasonably well, even for a single synthetic citation network. Furthermore, the h-index distribution for researchers derivative of an average network is essentially identical to our predicted h-index distribution (see Figures 3b, 3d, and 3f).

Hence, for a citation network exhibiting preferential attachment and assuming authors’ paper indices are uniformly distributed, we expect our model to forecast the h-index distribution for a researcher with $m$ publications. Consequentially, in the following section, we venture to evaluate our model on real-life citation networks.
(a) h-index for $m = 11$: one synthetic network
(b) h-index for $m = 11$: average synthetic network
(c) h-index for $m = 22$: one synthetic network
(d) h-index for $m = 22$: average synthetic network
(e) h-index for $m = 33$: one synthetic network
(f) h-index for $m = 33$: average synthetic network

Figure 3: Histograms of h-indices for 1000 simulated researchers having published 11, 22, and 33 papers with stochastic predictions for the same $m$. Results for one synthetic network are shown on the left, and for an average network on the right.

4 arXiv Data Analysis

4.1 Description and Methodology

We implement citation networks built from data obtained by D. P. George and R. Knegjens from arXiv.org [12]. The data consists of papers published between 1991 and 2018 in research fields with a minimum of 1000 papers. We realize our citation networks by connecting papers to the references they have made to other papers within the arXiv database. Specifically, a paper’s citation count increases only if referenced by another paper from the database with the same field designation. This methodology provides a unique citation network for 21 different fields of research (see Table 1). However, not all references a paper makes are found in the arXiv database, or are made to papers in the same field. Therefore, the citation networks
from different fields have varying proportions of references which are actually realized as citations to the total references made (see last column in Table 1).  

For the following analysis we display graphics from two fields: High Energy Physics - Theoretical (hep-th), and High Energy Physics - Experimental (hep-ex). This choice is not arbitrary. hep-th has the highest ratio of found citations to total references (51.3%); in particular, this is the field with the most complete realized citation network. Contrastingly, hep-ex is displayed in an attempt to limit bias since it is not in the upper echelon of found citation percent (22.0%), but is still a physics field.  

<table>
<thead>
<tr>
<th>Field Name</th>
<th>Abbrev.</th>
<th># Papers</th>
<th># Researchers</th>
<th>% Citations Found</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Energy Physics - Theoretical</td>
<td>hep-th</td>
<td>80364</td>
<td>518321</td>
<td>51.3%</td>
</tr>
<tr>
<td>High Energy Physics - Lattice</td>
<td>hep-lat</td>
<td>14432</td>
<td>366740</td>
<td>45.1%</td>
</tr>
<tr>
<td>Astrophysics</td>
<td>astro-ph</td>
<td>205651</td>
<td>365146</td>
<td>29.8%</td>
</tr>
<tr>
<td>General Relativity and Quantum Cosmology</td>
<td>gr-qc</td>
<td>41159</td>
<td>291784</td>
<td>27.5%</td>
</tr>
<tr>
<td>Quantum Physics</td>
<td>quant-ph</td>
<td>63112</td>
<td>460476</td>
<td>25.3%</td>
</tr>
<tr>
<td>Nuclear Theory</td>
<td>nucl-th</td>
<td>25524</td>
<td>524250</td>
<td>23.4%</td>
</tr>
<tr>
<td>High Energy Physics - Experimental</td>
<td>hep-ex</td>
<td>17432</td>
<td>34248</td>
<td>22.0%</td>
</tr>
<tr>
<td>Condensed Matter</td>
<td>cond-mat</td>
<td>210793</td>
<td>253199</td>
<td>21.0%</td>
</tr>
<tr>
<td>Nuclear Experimental</td>
<td>nucl-ex</td>
<td>8498</td>
<td>136347</td>
<td>15.2%</td>
</tr>
<tr>
<td>Math</td>
<td>math</td>
<td>272702</td>
<td>122705</td>
<td>6.3%</td>
</tr>
<tr>
<td>Mathematical Physics</td>
<td>math-ph</td>
<td>23212</td>
<td>143522</td>
<td>5.3%</td>
</tr>
<tr>
<td>High Energy Physics - Phenomenology</td>
<td>hep-ph</td>
<td>100698</td>
<td>269530</td>
<td>4.5%</td>
</tr>
<tr>
<td>Physics</td>
<td>physics</td>
<td>88300</td>
<td>512664</td>
<td>3.8%</td>
</tr>
<tr>
<td>Nonlinear Sciences</td>
<td>nlin</td>
<td>13917</td>
<td>521263</td>
<td>3.7%</td>
</tr>
<tr>
<td>Computer Science</td>
<td>cs</td>
<td>131611</td>
<td>446820</td>
<td>3.6%</td>
</tr>
<tr>
<td>Algebraic Geometry</td>
<td>alg-geom</td>
<td>1209</td>
<td>269636</td>
<td>2.5%</td>
</tr>
<tr>
<td>Chaotic Dynamics</td>
<td>chao-dyn</td>
<td>1770</td>
<td>123951</td>
<td>2.3%</td>
</tr>
<tr>
<td>Quantitative Finance</td>
<td>q-fin</td>
<td>5621</td>
<td>5670</td>
<td>2.0%</td>
</tr>
<tr>
<td>Statistics</td>
<td>stat</td>
<td>17340</td>
<td>281481</td>
<td>1.7%</td>
</tr>
<tr>
<td>Quantum Algebra</td>
<td>q-alg</td>
<td>1177</td>
<td>136572</td>
<td>1.1%</td>
</tr>
<tr>
<td>Quantitative Biology</td>
<td>q-bio</td>
<td>14384</td>
<td>535863</td>
<td>1.1%</td>
</tr>
</tbody>
</table>

Table 1: Fields in arXiv data set with more than 1000 papers identified.

4.2 Fitting Model Parameters

To evaluate our model we need to fit the parameters $c$ and $a$ from the realized citation networks.

The parameter $c$ is simply found by calculating the average number of citations all papers in a field have obtained. This serves as an adequate stand-in for the average number of citations all papers make because every citation made is self-contained within that network. So, the average number of citations received by any paper is equivalent to the average number of citations made.

Identifying $a$ is more involved and requires some insight into the structure of citation networks themselves. The in-degree distribution of large networks exhibiting preferential attachment properties adhere to a power shape:

$$P(k) \propto k^{-\gamma}$$

where $\gamma$ is a parameter that characterizes the network. In this case, $\gamma = 3$ indicates a very strong preferential attachment, meaning that newer papers are cited less frequently than older papers.

3In general, physics fields have the highest proportions of found to total references, likely because arXiv was originally physics focused.

4Relevant graphics associated with the remaining 19 fields can be found in the Appendix.
law such that \( d(k) \propto k^{-\alpha} \), where \( d(k) \) is the fraction of nodes with in-degree \( k \) \[13\]. Moreover, \( \sum_{i \geq k} d(i) \propto k^{1-\alpha} \), this is often referred to as the (in-degree) tail distribution. Specifically, for Price’s model: \( \alpha = 2 + \frac{a}{c} \) which, given an estimate of \( c \), allows for the fitting of \( a \).

![In-degree distributions for the derived hep-th and hep-ex networks.](image1)

Figure 4: In-degree distributions for the derived hep-th and hep-ex networks.

![Tail distributions and corresponding linear fit for the derived hep-th and hep-ex networks.](image2)

Figure 5: Tail distributions and corresponding linear fit for the derived hep-th and hep-ex networks.

As expected \[13\], and seen in Figures 4a and 4b, fitting \( a \) from the in-degree distribution is problematic due to the non-linearity observed for large and small \( k \). Following the usual methodology \[13\], we instead fit the parameter \( a \) using the tail distribution. As demonstrated by Figures 5a and 5b, the data is still not quite linear. Therefore, we fit only the data within 1.5 standard deviations of the corresponding mean.

The fitted parameters for each realized citation network carry through the rest of our analysis (see Table 2).
<table>
<thead>
<tr>
<th>Abbrev.</th>
<th>Fitted $c$</th>
<th>Fitted $a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>hep-th</td>
<td>19.101</td>
<td>17.115</td>
</tr>
<tr>
<td>hep-lat</td>
<td>12.113</td>
<td>9.924</td>
</tr>
<tr>
<td>astro-ph</td>
<td>14.077</td>
<td>11.344</td>
</tr>
<tr>
<td>gr-qc</td>
<td>9.781</td>
<td>9.599</td>
</tr>
<tr>
<td>quant-ph</td>
<td>7.958</td>
<td>5.896</td>
</tr>
<tr>
<td>nucl-th</td>
<td>8.813</td>
<td>8.766</td>
</tr>
<tr>
<td>hep-ex</td>
<td>6.737</td>
<td>6.774</td>
</tr>
<tr>
<td>cond-mat</td>
<td>6.795</td>
<td>5.588</td>
</tr>
<tr>
<td>nucl-ex</td>
<td>5.065</td>
<td>3.255</td>
</tr>
<tr>
<td>math</td>
<td>1.341</td>
<td>1.564</td>
</tr>
<tr>
<td>math-ph</td>
<td>1.450</td>
<td>1.982</td>
</tr>
<tr>
<td>hep-ph</td>
<td>19.689</td>
<td>20.545</td>
</tr>
<tr>
<td>physics</td>
<td>0.974</td>
<td>1.212</td>
</tr>
<tr>
<td>nlin</td>
<td>1.068</td>
<td>1.263</td>
</tr>
<tr>
<td>cs</td>
<td>0.773</td>
<td>0.276</td>
</tr>
<tr>
<td>alg-geom</td>
<td>0.261</td>
<td>0.507</td>
</tr>
<tr>
<td>chao-dyn</td>
<td>0.499</td>
<td>0.602</td>
</tr>
<tr>
<td>q-fin</td>
<td>0.515</td>
<td>0.834</td>
</tr>
<tr>
<td>stat</td>
<td>0.469</td>
<td>0.594</td>
</tr>
<tr>
<td>q-alg</td>
<td>0.180</td>
<td>0.080</td>
</tr>
<tr>
<td>q-bio</td>
<td>0.315</td>
<td>0.306</td>
</tr>
</tbody>
</table>

Table 2: Price’s model’s fitted parameters for each of the 21 realized citation networks.

4.3 Researchers Publication Frequency

In order to substantiate one of the fundamental assumptions of our h-index prediction, we explore the paper publication frequency for arXiv researchers. Every paper published in this database has an associated date of publication. This allows us to determine the number of papers published in a realized network by other researchers during each author’s publication interim. We call this an inter-publication index.

Figure 6 shows the inter-publication indices for all researchers with at least two papers in the hep-th (Figure 6a) and hep-ex (Figure 6b) realized networks. We can see that inter-publication indices are somewhat reasonably described by an exponential distribution, but it is still important to address the imperfect fit. As seen in Figure 6, there is significantly more mass near 0 than the maximum likelihood estimate can account for. This indicates that a consequential number of researchers have smaller inter-publication indices than expected. We believe this is likely due to co-authorship. The shared workload accompanying co-authorship could increase an individual’s publication rate, therefore, shrinking their inter-publication indices. Additionally, co-authors’ inter-publication distributions are correlated, adding confounding complexities.

Going forward, we assess it is safe to assume that inter-publication indices of researchers are exponentially distributed within each field. Consequently, if we are to condition on $m$ publications an author’s paper indices can instead be given by $m$ uniform order statistics, from the properties of a Poisson point process. This is the same distribution we assume for a researcher’s paper indices when ordered by (expected) citation count (see Section 2.4).

It is also important to note, however, that the exponential fit is field dependent. This is exemplified in Figure
6: hep-ex, despite having an apparent exponential trend, has a significantly worse fit than hep-th and, in fact, most fields (see Appendix A.1). Inherent differences between fields themselves, and disparity in the integrity of the data (see Table 1) are likely the primary factors behind fit diversity.

![Histograms of inter-publication indices and a fitted exponential probability density for hep-th and hep-ex.](image)

(a) Inter-publication indices: hep-th

(b) Inter-publication indices: hep-ex

Figure 6: Histograms of inter-publication indices and a fitted exponential probability density for hep-th and hep-ex.

### 4.4 Researchers with a Given Number of Publications

To continue evaluating our model, we now consider the h-index distributions within our newly fashioned arXiv citation networks. For each realized network, we compute the h-indices for all researchers having published \( m \) papers and compare this distribution to our theoretical prediction for \( P(h\text{-index} = k | N = m) \). Specifically, we again consider \( m \in \{11, 22, 33\} \). \(^5\) These values were chosen summarily, with the specification that, for a particular field there are at least 50 researchers with the appropriate \( m \). \(^6\)

---

\(^5\)The analogous figures, with \( m \in \{11, 22, 33\} \), for the remaining fields are found in Appendix A.2. Also, figures for hep-th and hep-ex with other values of \( m \) which meet the specified criteria are similarly in Appendix A.2.

\(^6\)Since the h-index of an author is at most \( m \), graphics for small \( m \) are less significant than for larger \( m \).
As seen in Figure 7, our predictions systematically fail to represent the true distribution of researchers’ h-indices in their entirety. Rather, the true h-index distributions are more widely spread around their means than our model anticipates. This is found to be a consistent issue for the majority of $m$ across all realized citation networks (see also Appendix A.2).

![Histograms of h-indices (blue bars) and model predictions (red dots) for researchers with 11, 22, and 33 publications within hep-th and hep-ex.](image)

Figure 7: Histograms of h-indices (blue bars) and model predictions (red dots) for researchers with 11, 22, and 33 publications within hep-th and hep-ex.

### 4.5 Synthetic Researchers with a Given Number of Publications

The comparison between our predicted and true h-indices in Section 4.4 are underwhelming. To pinpoint reasons for this, we simulate “synthetic” researchers utilizing the same method described in Section 3.2, i.e. we sample $m$ papers uniformly at random from an entire realized citation network and attribute those papers to a single “synthetic” author.

As seen in Figure 8 and Appendix A.3, the h-index distributions of “synthetic” researchers adhere significantly better to our predicted h-index model. The discrepancy between “synthetic” and real arXiv researchers
indicates there are factors that exist outside the scope of our model. We believe the two principle culprits are co-authorship and the assumption that an author’s range of publication indices is indiscriminate in the network. Firstly, co-authorships are presumably an important factor because “synthetic” researchers are less likely to “collaborate” than real researchers. For a large enough network, abundant co-authorship is improbable for the reason that “synthetic” researchers are uniformly assigned random publications within the network. Secondly, it is highly implausible that all real authors (e.g. younger researchers or inactive researchers) have published papers at indices uniformly distributed between 1991 and 2018.

Figure 8: Histograms of h-indices for our “synthetic” researchers with 11, 22, and 33 papers and associated predictions.

4.6 The h-index for an Arbitrary Number of Publications

Our model can be used to predict the h-index distribution for a researcher with an arbitrary number of publications using,

$$
P(\text{h-index} = k) = \sum_{i \geq 1} P(\text{h-index} = k | N = i) \cdot P(N = i).$$
\( \mathbb{P}(N = i) \) is the probability a researcher published \( i \) papers in the network and is estimated empirically.

To construct confidence intervals for \( \mathbb{P}(h\text{-index} = k) \) for each \( k \geq 0 \), we exploit that, under our model, this probability is a linear function of the vector \( p := (\mathbb{P}(N = i))_{i \geq 1} \). And, if we define the vector \( q_k := (\mathbb{P}(h = k | N = 1), \mathbb{P}(h = k | N = 2), \mathbb{P}(h = k | N = 3), \ldots) \), then \( \mathbb{P}(h\text{-index} = k) = \langle p, q_k \rangle \). Since the estimator \( \hat{p} \) of \( p \) is an average of say \( \ell \) vectors, the Central Limit Theorem implies for a large \( \ell \) that

\[
\langle \hat{p}, q_k \rangle \overset{d}{\sim} \text{Normal} \left( \mathbb{P}(h\text{-index} = k), \frac{q_k^T \Sigma q_k}{\ell} \right),
\]

where \( \Sigma \) is the variance-covariance matrix

\[
\Sigma := \begin{bmatrix}
  p_1 (1 - p_1) & -p_1 p_2 & -p_1 p_3 & \\
  -p_2 p_1 & p_2 (1 - p_2) & -p_2 p_3 & \\
  -p_3 p_1 & -p_3 p_2 & p_3 (1 - p_3) & \\
  \vdots & \vdots & \vdots & \\
\end{bmatrix}.
\]

An asymptotic \( 100(1 - \alpha)\% \) confidence interval for the probability that h-index = \( k \) is therefore given by the formula:

\[
\mathbb{P}(h\text{-index} = k) = \langle \hat{p}, q_k \rangle \pm z_{\alpha/2} \sqrt{\frac{q_k^T \Sigma q_k}{\ell}}.
\]

In what follows, we use the above analysis to construct approximate 95\% confidence intervals for \( \mathbb{P}(h\text{-index} = k) \); in particular, \( \alpha = 0.05 \) and \( z_{0.025} = 1.644854 \). Note, however, that \( \Sigma \) is unknown because \( p \) is unknown as well. As a result, we estimate the variance-covariance matrix by

\[
\hat{\Sigma} := \begin{bmatrix}
  \hat{p}_1 (1 - \hat{p}_1) & -\hat{p}_1 \hat{p}_2 & -\hat{p}_1 \hat{p}_3 & \\
  -\hat{p}_2 \hat{p}_1 & \hat{p}_2 (1 - \hat{p}_2) & -\hat{p}_2 \hat{p}_3 & \\
  -\hat{p}_3 \hat{p}_1 & -\hat{p}_3 \hat{p}_2 & \hat{p}_3 (1 - \hat{p}_3) & \\
  \vdots & \vdots & \vdots & \\
\end{bmatrix}.
\]

As seen in Figure 9 and Appendix A.4, our model predicts within approximate 95\% confidence the distribution of non-zero h-indices. Nevertheless, underestimates of the probability that the h-index = 0 are consistently sizable; meaning, the number of papers which obtain no citations is significantly larger than our model predicts.

We believe the insubstantial estimation of \( \mathbb{P}(h\text{-index} = 0) \) is predominantly a product of the deficiencies in the arXiv data. Specifically, for all fields, the ratio of found to total citations in the associated networks is much less than one (see Table 1). So, it is reasonable to expect that a significant subset of the papers incorrectly have 0 citations.
4.7 Expected h-indices for Given Number of Publications

As a final macroscopic evaluation of our model, for each $m \geq 1$, we compare the average h-index of researchers with $m$ papers to our prediction for $\mathbb{E}(h\text{-index}|N = m)$. Recall from Section 2.4 that according to our model

$$\mathbb{E}(h\text{-index}|N = m) = \sum_{i=1}^{m} p_k.$$ 

As seen in Figure 10 and Appendix A.5, our model accurately predicts expected h-indices of researchers with a relatively small number $m$ publications. The orange lines give some indication as to what a “small” $m$ is: $m \leq 49$ in hep-th and $m \leq 46$ in hep-ex. Although most of the 95% confidence intervals still encompass our predictions, persistent divergence from the theoretical expected h-index appears as $m$ increases. The parity between growth of $m$ and confidence interval length is a result of the shrinking sample size of authors with comparatively large publication quantities. For instance, in hep-th (hep-ex), for 95% of $m > 73$ ($m > 159$), less than 10 researchers have that particular amount of $m$ publications. With such small sample sizes, there is too much uncertainty to adequately assess the quality of our predictions for larger quantities of $m$ publications. Nevertheless, it is apparent that our model tends to underestimate the expected h-index of researchers with a large number of publications. Various model imperfections could cause this underestimation. However, we believe co-authorship is likely one contributing phenomenon because it correlates and can inflate the h-indices of the applicable authors.
5 Conclusion

In recent times, the accessibility of research has grown immeasurably. The information researchers are in pursuit of has never been more attainable. As a result, the evaluation of a researcher regarding the impact
in their field has become considerably more substantial. This has led to bibliometrics receiving appreciable development and analysis. Since its inception in 2005, the h-index has expanded in popularity, arguably achieving some universality amid bibliometrics for research impact. In accordance, this paper has outlined a model regarding the stochastic behavior of the h-index for an author with $m$ publications within a random citation network obeying Price’s preferential attachment model.

We were first able to understand the expected citation count of any paper originating from a preferential attachment citation network, contingent only upon its relative index of publication. Precisely, our asymptotic formulation for $E_{j,n} \approx g \left( \frac{j}{n} \right)$ dependably characterizes the expected citation count of the $j$-th publication within a sufficiently large citation network of $n$ papers. Our result bears similarities to a formula in [14], despite their use of a different “agent-based” model.

In the future, to more fundamentally comprehend the limits of our model, we would also need to establish the model implications of citation count variability. Doing so would necessitate either discerning an explicit expression for the variance of $C_{j,n}$ or at least bounding said expression in terms of $j$, $n$, and Price’s model parameters.

Using $g \left( \frac{j}{n} \right)$ as a proxy for the actual number of citations the $j$-th paper in the network receives, and assuming their paper indices are uniformly distributed, we established an h-index probability distribution for a researcher with $m$ papers.

For synthetic networks generated via Price’s model, our model accurately predicts the h-index of (also synthetic) authors in the network. However, upon application to real citation networks with actual researchers, our h-index distribution underperforms. It is unable to capture a significant subset of the actual h-index distribution with regularity. Nonetheless, for researchers with a relatively small number $m$ publications, we reasonably characterize expected h-indices.

We assess the discrepancy between actual and predicted h-indices to likely be a result of two primary aspects of our model: assumptions about uniform publication indices, and correlation effects due to co-authorship. Since it is unlikely that a researcher’s body of work spans the entire duration of a particular field, we propose limiting the possible range of paper indices attributed to an author. Essentially, each researcher would have a window of indices (determined empirically based on the index of their first and last publication) during which their remaining publications would be assumed to occur uniformly at random. Co-authorship, however, is a more nuanced problem because, for a co-authored publication, both the index of publication and the citation count are correlated for all of the involved researchers. As a result, no straightforward solution is immediately apparent. With extended investigation there is potential for a more robust stochastic h-index prediction.

**Acknowledgments.** I am very thankful of Dr. Dougherty for partially funding this research through her NSF EXTREEMS training grant No. DMS 1407340.

I am also incredibly appreciative of my advisor, Professor Lladser, for his encouragement, insight, and guidance. As well as grateful of my father, my brothers and Miss Tara Khan for all of their advice and endless support. And, of course, I am perpetually inspired by my mother.
References


[3] L. Bornmann and H.-D. Daniel, “The state of h index research. is the h index the ideal way to measure research performance?,” *EMBO reports*, vol. 10, pp. 2–6, Jan 2009.


A Appendix

A.1 Additional Figures for Section 4.3

Figure 11: Histograms of inter-publication indices and a fitted exponential probability density for remaining 19 fields (see Table 1).
Figure 11: Histograms of inter-publication indices and a fitted exponential probability density for remaining 19 fields (see Table 1) (cont.).
Figure 11: Histograms of inter-publication indices and a fitted exponential probability density for remaining 19 fields (see Table 1) (cont.).

A.2 Additional Figures for Section 4.4

Figure 12: Histograms of h-indices (blue bars) and model predictions (red dots) for researchers with 11 publications for remaining fields with at least 50 total researchers.
Figure 12: Histograms of h-indices (blue bars) and model predictions (red dots) for researchers with 11 publications for remaining fields with at least 50 total researchers (cont.).
Figure 12: Histograms of h-indices (blue bars) and model predictions (red dots) for researchers with 11 publications for remaining fields with at least 50 total researchers (cont.).

Figure 13: Histograms of h-indices (blue bars) and model predictions (red dots) for researchers with 22 publications for remaining fields with at least 50 total researchers.
Figure 13: Histograms of h-indices (blue bars) and model predictions (red dots) for researchers with 22 publications for remaining fields with at least 50 total researchers (cont.).

Figure 14: Histograms of h-indices (blue bars) and model predictions (red dots) for researchers with 33 publications for remaining fields with at least 50 total researchers.
Figure 15: hep-th: h-index distribution for remaining $m$ with more than 50 researchers.
Figure 15: hep-th: h-index distribution for remaining $m$ with more than 50 researchers (cont.).

Figure 16: hep-ex: h-index distribution for remaining $m$ with more than 50 researchers.
Figure 16: hep-ex: h-index distribution for remaining $m$ with more than 50 researchers (cont.).
Figure 16: hep-ex: h-index distribution for remaining $m$ with more than 50 researchers (cont.).

A.3 Additional Figures for Section 4.5

Figure 17: Histograms of h-indices (blue bars) and model predictions (red dots) for “synthetic” researchers with 11 publications for remaining fields with at least 50 total researchers.
Figure 17: Histograms of h-indices (blue bars) and model predictions (red dots) for “synthetic” researchers with 11 publications for remaining fields with at least 50 total researchers.
Figure 17: Histograms of h-indices (blue bars) and model predictions (red dots) for “synthetic” researchers with 11 publications for remaining fields with at least 50 total researchers (cont.).

Figure 18: Histograms of h-indices (blue bars) and model predictions (red dots) for “synthetic” researchers with 22 publications for remaining fields with at least 50 total researchers.
Figure 18: Histograms of h-indices (blue bars) and model predictions (red dots) for “synthetic” researchers with 22 publications for remaining fields with at least 50 total researchers (cont.).
Figure 19: Histograms of h-indices (blue bars) and model predictions (red dots) for researchers with 33 publications for remaining fields with at least 50 total researchers.

Figure 20: hep-th: h-index distribution for remaining m with more than 50 researchers.
Figure 20: hep-th: h-index distribution for remaining $m$ with more than 50 researchers.
Figure 20: hep-th: h-index distribution for remaining $m$ with more than 50 researchers (cont.).

Figure 21: hep-ex: h-index distribution for remaining $m$ with more than 50 researchers.
Figure 21: hep-ex: h-index distribution for remaining $m$ with more than 50 researchers (cont.).
A.4 Additional Figures for Section 4.6

(a) h-index distribution for all $m$: alg-geom

(b) h-index distribution for all $m$: astro-ph

(c) h-index distribution for all $m$: chaodyn

(d) h-index distribution for all $m$: cond-mat

(e) h-index distribution for all $m$: cs

(f) h-index distribution for all $m$: gr-qc

Figure 22: h-index distributions for arbitrary $m$ for the remaining 19 fields (see Table 1).
Figure 22: h-index distributions for arbitrary $m$ for the remaining 19 fields (see Table 1) (cont.).
Figure 22: h-index distributions for arbitrary $m$ for the remaining 19 fields (see Table 1) (cont.).
Figure 22: h-index distributions for arbitrary \( m \) for the remaining 19 fields (see Table 1) (cont.).

A.5 Additional Figures for Section 4.7

(a) Theoretical expected h-index and average h-index for all \( m \): alg-geom
(b) Theoretical expected h-index and average h-index for all \( m \): astro-ph

(c) Theoretical expected h-index and average h-index for all \( m \): chao-dyn
(d) Theoretical expected h-index and average h-index for all \( m \): cond-mat

Figure 23: Expected h-index (red) versus average h-index (blue) in terms of the number \( m \) of publications. The blue segments represent 95% confidence intervals for the average h-index of authors with \( m \) papers. The orange line defines the regime to the left of which 95% of the \( m' \)’s have more than 25 researchers and the green line, the regime, such that to the right of the line 95% of the \( m' \)’s have less than 10 researchers. Figures for remaining 19 fields (see Table 1).
Figure 23: Expected h-index (red) versus average h-index (blue) in terms of the number \( m \) of publications. The blue segments represent 95\% confidence intervals for the average h-index of authors with \( m \) papers. The orange line defines the regime to the left of which 95\% of the \( m \)’s have more than 25 researchers. The green line defines the regime to the right of which 95\% of the \( m \)’s have less than 10 researchers. Figures for remaining 19 fields (see Table 1) (cont.).
Figure 23: Expected h-index (red) versus average h-index (blue) in terms of the number $m$ of publications. The blue segments represent 95% confidence intervals for the average h-index of authors with $m$ papers. The orange line defines the regime to the left of which 95% of the $m$'s have more than 25 researchers. The green line defines the regime to the right of which 95% of the $m$'s have less than 10 researchers. Figures for remaining 19 fields (see Table 1) (cont.).
Figure 23: Expected h-index (red) versus average h-index (blue) in terms of the number $m$ of publications. The blue segments represent 95% confidence intervals for the average h-index of authors with $m$ papers. The orange line defines the regime to the left of which 95% of the $m$’s have more than 25 researchers. The green line defines the regime to the right of which 95% of the $m$’s have less than 10 researchers. Figures for remaining 19 fields (see Table 1) (cont.).