# Learning to Predict Citation-Based Impact Measures

Luca Weihs
University of Washington, Seattle
Department of Statistics, Box 354322
Seattle, WA 98195-4322
lucaw@uw.edu

Oren Etzioni
Allen Institute for Artificial Intelligence
2157 N Northlake Way Suite 110
Seattle, WA 98103
orene@allenai.org

## ABSTRACT

Citations implicitly encode a community's judgment of a paper's importance and thus provide a unique signal by which to study scientific impact. Efforts in understanding and refining this signal are reflected in the probabilistic modeling of citation networks and the proliferation of citation-based impact measures such as Hirsch's h-index. While these efforts focus on understanding the past and present, they leave open the question of whether scientific impact can be predicted into the future. Recent work addressing this deficiency has employed linear and simple probabilistic models; we show that these results can be handily outperformed by leveraging non-linear techniques. In particular, we find that these AI methods can predict measures of scientific impact for papers and authors, namely citation rates and h-indices, with surprising accuracy, even 10 years into the future. Moreover, we demonstrate how existing probabilistic models for paper citations can be extended to better incorporate refined prior knowledge. While predictions of "scientific impact" should be approached with healthy skepticism, our results improve upon prior efforts and form a baseline against which future progress can be easily judged.

## KEYWORDS

Citation prediction, citation network, scientific impact, h-index, reinforced poisson process

## 1 INTRODUCTION

This paper investigates the problem of predicting scientific impact for individual authors and papers up to 10 years into the future. As there is no consensus as to which measure of scientific impact is best, we follow prior work [1, 20], and quantify author impact using the h-index, and paper impact with citation counts. We test the efficacy of our methods on a data set of close to four million computer science papers written by approximately 800,000 authors and published in the years spanning 1975 to 2016. This data set, which we have made publicly available,[1] is unique in both size, more than an order of magnitude larger than others, and in breadth, covering papers published in over 7000 conferences and journals.

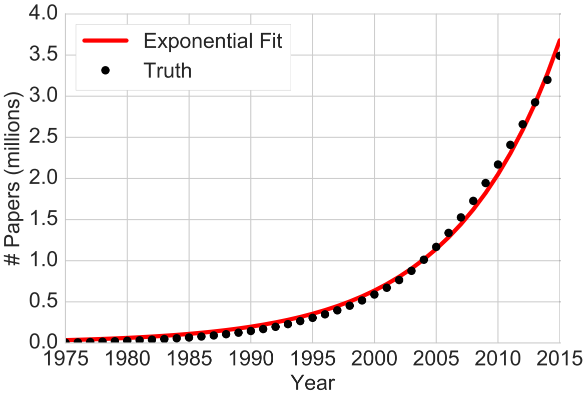


**Figure 1: The cumulative number of published papers in our data set over time. The (plotted) exponential fit suggests that the number of papers can be expected to double approximately every six years.**

For our prediction task, we use information available in 2005 to predict impact in the subsequent 10 year period, the years 2006 to 2015.

Because of their simplicity and ubiquity, citation counts have become perhaps the most popular measure of scientific impact. While citation counts are indeed a simple and useful proxy for impact, they exhibit a number of flaws, especially when being used to characterize authors. One common criticism of citation counts is that they fail to capture any notion of how citations are distributed across a researcher's publications. For instance, we might expect that an author with 60 citations is more impactful if those citations are distributed equally across 6 papers rather than across 30. Equally problematic, if an author has a short career with a single high citation paper, perhaps a survey or interdisciplinary publication, they may seem more impactful than a researcher with a long history of moderate impact publications. A growing realization, pioneered by Hirsch [12], that these flaws must be addressed has led to a recent explosion in surrogate measures of impact. A small subset of these measures include the h-index, g-index, c-index, eigenfactor, and hip-index [4, 10, 12, 25, 27]. The h-index, the author impact measure we predict, equals the largest value $N$ for which an author has $N$ publications each with greater than or equal to $N$ citations. Notice that the h-index is unaffected by outliers, a single publication can only ever increase the h-index by 1, and also penalizes having many papers with few citations, a publication only contributes to the h-index if it has sufficiently many citations.

[1] The data set, along with code to reproduce our results, can be found at https://github.com/Lucaweihs/impact-prediction.

*JCDL'17, Toronto, Canada*
© 2017 ACM. 978-x-xxxx-xxxx-x/YY/MM...$15.00
DOI: 10.1145/nnnnnnn.nnnnnnn

While quantifying current scientific impact is of substantial interest, there are a wide variety of problems for which the future is far more important than the present (e.g. the question of granting tenure). Moreover, given the exponential explosion in published papers, see Figure 1, it is critical to design automated systems which detect impactful work as early as possible. To better address these questions, there has been a surge of research in scientific impact prediction for authors and individual papers.

There are two primary strategies for impact prediction. The first is the spiritual successor of the work of Price [18] in which citation counts are statistically modeled using the intuitions provided by the *preferential attachment* model of network growth and empirical studies [20, 22]. The second predicts impact with a machine learning approach, that is, extensive feature engineering followed by supervised learning with a regression model [1, 6–9, 26]. We compare these two approaches on our data set and propose a method, for paper citation prediction, to bridge the gap between them. We note that there have been some recent efforts which do not fall precisely into either of the above categories. For instance, Nezhadbiglari et al. [15] exploit K-Spectral Clustering to discover cluster centroids among author citation histories invariant to both scaling and shifts; they then combine the information from these centroids with simple author-level features to predict author cluster membership and future citation counts.

The remainder of this paper is organized as follows. We begin with a summary of the features we use to characterize individual papers and authors. We then demonstrate how these features can be used to predict author h-index and paper citation counts. For papers, we compare the machine learning approaches to those inspired by probabilistic modeling; we are not aware of any predictive probabilistic models for author h-index and thus cannot make such a comparison for authors. We end with an analysis of which features are most strongly related to our prediction targets and a discussion of our results and future work. We also include an appendix detailing our modifications to the reinforced poisson model of Shen et al. [20] used to predict paper citations.

## 2 FEATURE ENGINEERING

In order to apply supervised learning techniques to our problem, we must first develop a collection of features summarizing individual papers and authors. In this work we focus our attention on features that can be extracted from the citation graph, coauthor graph, and paper metadata (e.g. authors and venue); we leave the creation of content-based features extracted from papers' text as future work. A specification of all 44 author features, many inspired by prior work [1, 7, 8, 19, 26], can be found in Table 1. The 63 features for papers are similar and remain unlisted to spare space;[2] we give a summary of all features below. Before continuing we should note that there are several one-to-many relationships between papers, authors, and venues; namely an author publishes many papers in different venues and a single paper can have many authors. Because of this, we often must summarize a variable length vector of information. For instance, suppose a paper had five authors who have, respectively, obtained an average of 5, 3, 2, 1, and 1 citations per paper. We will capture this variable length information with three summary statistics, namely the mean, minimum, and maximum. Hence the above vector (5, 3, 2, 1, 1) would be reduced to (2.4, 1, 5). As we always make such a reduction, we will not explicitly note so below.

### 2.1 Metadata

Some features can be extracted from paper metadata with little to no processing. For instance, the author count, whether or not the paper is a survey, and the number of years since the paper was published.

### 2.2 Impact History

Some of the strongest predictors of future impact can be extracted directly from the time series of citation counts and h-indices for individual papers and authors. Such features include the total citation count, year-over-year change in citation rate, and long-term average citation rate. We also consider the impact history of venues, extracting essentially the same features for venues as for authors, and summarize this information for individual papers and authors.

### 2.3 Citation and Coauthor Graphs

The topology of the citation and coauthor graphs offer compelling information about the centrality and influence of papers and their authors. One might expect, for example, that coauthors will tend to cite one another; thus, having high degree in the coauthor graph suggests future h-index growth. Indeed Sarigölet et al. [19] demonstrate that measures of centrality in the coauthorship network alone provide strong signals of future success. For computational efficiency, our primary measures of centrality are in/out-degree and the PageRank [16].

## 3 AUTHOR IMPACT

We first consider author h-index prediction using a machine learning approach. We generate a collection of 44 features for each author, and use these features within several regression models. As we predict up to 10 years into the future, we generate features for authors whose first publication was in or before 2005 and *only* use data that would have been available in 2005. The observed author h-indices in 2006-2015 are then used as targets for prediction. Note that we train the same models several times, once for each of the 10 target years 2006-2015. To filter out inactive authors, we follow Acuna et al. [1] and only include authors having an h-index of at least 4 and whose first article was published between 5 and 12 years prior to 2005. We train the following regression models; these models are ordered by increasing complexity, beginning with simple baselines and ending with state-of-the-art machine learning algorithms.

(1) Plus-k (PK) - A baseline model that adds a fixed constant to all author's h-indices every year; this constant, equaling 0.402, is chosen by linear regression using the Huber loss which better handles outliers than the usual squared error loss [13].
(2) Simple Markov (SM) - A linear regression model that is only given features describing the author's h-index in 2005 and the change in the author's h-index from 2003 to 2005.

[2] A listing of the paper features can be found, along with the code and data, at https://github.com/Lucaweihs/impact-prediction.

**Table 1: All features used for author h-index prediction.**

| Feature Name | Description |
|---|---|
| author_hindex | • H-index |
| author_hindex_delta | • Change in h-index over the last two years |
| author_citation_count | • Cumulative citation count |
| author_key_citation_count | • Cumulative key citation count [27] |
| author_citations_delta_{0,1} | • Citations this year and one year ago |
| author_key_citations_delta_{0,1} | • Key citations this year and one year ago |
| author_mean_citations_per_paper | • Mean number of citations per paper |
| author_mean_citation_per_paper_delta | • Change in mean cites per paper over last two years |
| author_mean_citations_per_year | • Mean number of citations per year |
| author_papers | • Number of papers published |
| author_papers_delta | • Number of papers published in last two years |
| author_mean_citation_rank | • Rank of author (between 0 and 1) among all other authors in terms of mean citations per year |
| author_unweighted_pagerank | • PageRank of author in unweighted coauthorship network |
| author_weighted_pagerank | • PageRank of author in weighted coauthorship network |
| author_age | • Career length (years since first paper published) |
| author_recent_num_coauthors | • Total number of coauthors in last two years |
| author_max_single_paper_citations | • Max number of citations for any of author's papers |
| venue_hindex_{mean, min,max} | • H-indices of venues author has published in |
| venue_hindex_delta_{mean, min,max} | • 2-year h-index change for venues author has published in |
| venue_citations_{mean, min,max} | • Mean citations per paper of venues author has published in |
| venue_citations_delta_{mean, min,max} | • Change in mean citations per paper over last two years for venues author has published in |
| venue_papers_{mean, min, max} | • Number of papers in venues in which the author has published |
| venue_papers_delta_{mean, min, max} | • Change in number of papers in venues in which the author has published over the last two years |
| venue_rank_{mean, min, max} | • Ranks of venues (between 0-1) in which the author has published determined by mean number of citations per paper |
| venue_max_single_paper_citations_{mean, min, max} | • Maximum number of citations any paper published in a venue has received for each venue the author has published in |
| total_num_venues | • Total number of venues published in |

(3) Lasso (LAS) - A regularized linear regression model using all features with the regularization parameter chosen by 10-fold cross validation [21].
(4) Random forest (RF) - An ensemble of regression trees using randomization techniques to improve performance [5].
(5) Gradient boosted regression trees (GBRT) - a collection of simple regression trees that are trained iteratively by a type of functional gradient descent [11]. Gradient boosted trees can perform very well but, unlike random forests, can require extensive parameter tuning.

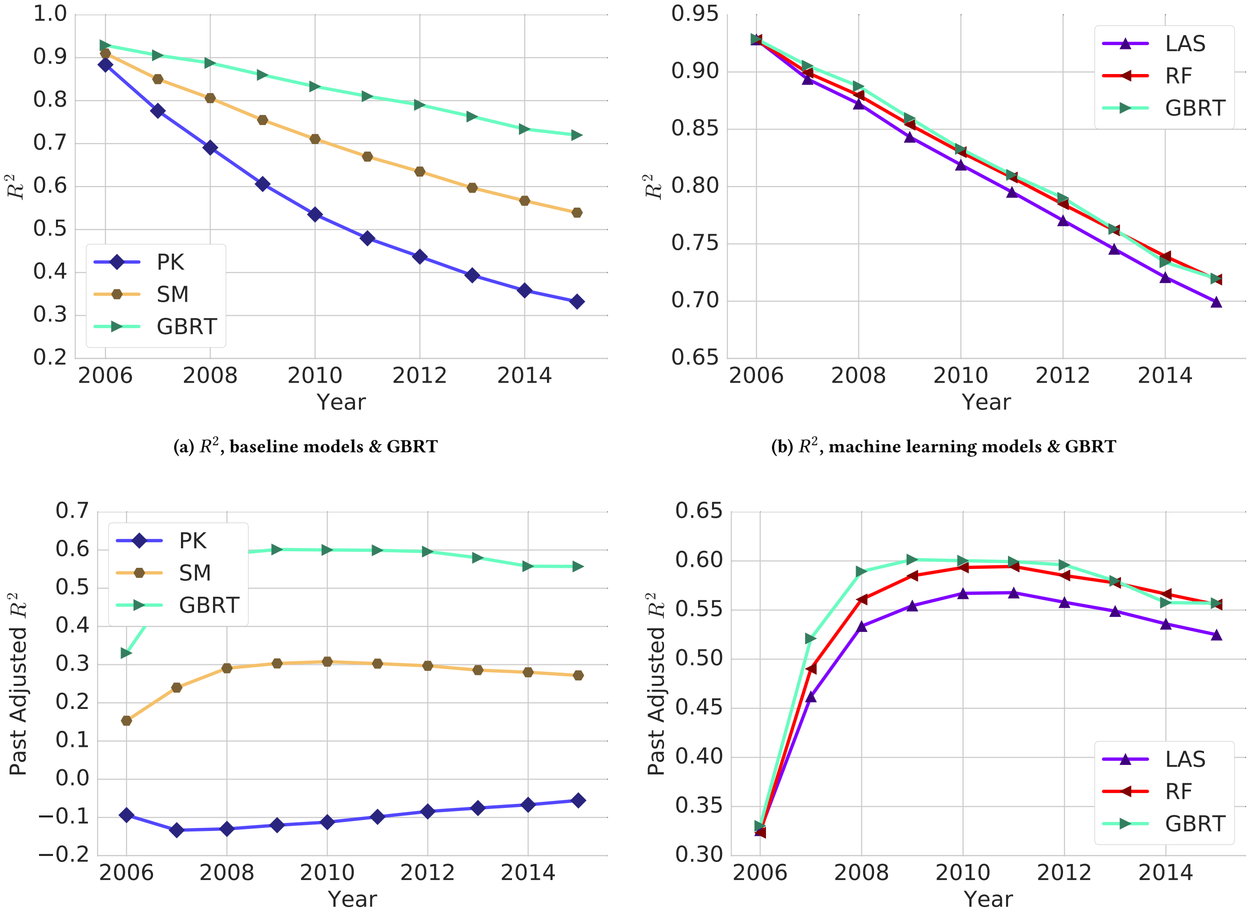
To evaluate the performance of our predictions we consider three performance metrics described below. The first measure we consider is the well-known $R^2$ metric. $R^2$ compares the relative performance of a model against a predictor that simply returns the mean of the labels. In our setting, $R^2$ equals

$$1 - \frac{\sum_{i=1}^{N}(y_{i,j} - \hat{y}_{i,j})^2}{\sum_{i=1}^{N}(y_{i,j} - \overline{y}_j)^2}$$

where $N$ is the total number of authors, $y_{i,j}$ is the h-index of the $i$th author in year $j$, $\hat{y}_{i,j}$ is the predicted h-index for the author in that year, and $\overline{y}_j = \frac{1}{N}\sum_{i=1}^{N} y_{i,j}$ is the average h-index over all authors. While $R^2$ is a popular measure of regression performance, it tends to overstate predictive power in the impact prediction setting [17]. This inflation occurs because citation counts and h-indices cannot decrease and are highly auto-correlated, that is, dependent on their value in the prior year. To remove some of this auto-correlation, we modify the $R^2$ metric by subtracting the known number of citations in 2005 from the prediction targets in 2006-2015. We define a new metric, which we call the *Past Adjusted* $R^2$ (PA-$R^2$), as

$$1 - \frac{\sum_{i=1}^{N}(y_{i,j} - \hat{y}_{i,j})^2}{\sum_{i=1}^{N}(z_{i,j} - \overline{z}_j)^2}$$

where $z_{i,j} = y_{i,j} - y_{i,2005}$ and $\overline{z}_j = \frac{1}{N}\sum_{i=1}^{N} z_{i,j}$. By subtracting the known quantity $y_{i,2005}$ from $y_{i,j}$ in the denominator of PA-$R^2$, we make the denominator strictly smaller and thus remove some misleading inflation in the statistic. While it is often stated that the $R^2$ metric lies between 0 and 1, this need only be the true in some special cases, for instance, when computing training error with linear regression. Both $R^2$ and PA-$R^2$ will always be less than or equal to 1 but may be negative.



(a) $R^2$, baseline models & GBRT

(b) $R^2$, machine learning models & GBRT

(c) PA-$R^2$, baseline models & GBRT

(d) PA-$R^2$, machine learning models & GBRT

**Figure 2: $R^2$ and PA-$R^2$ of author h-index predictions. The GBRT outperform the baseline models significantly and also provide notable improvements over the other machine learning models.**

For a test set of 2,566 authors, we compute the $R^2$ and PA-$R^2$ measures for all of the above models and display the results in Figure 2. In terms of $R^2$, we see in (Fig. 2a) that the gradient boosted regression trees substantially outperform the simple baseline models. Between the three machine learning models, the differences in performance are less stark: the GBRT and RF models are essentially tied, while the lasso model performs only slightly worse (Fig. 2b). However, when using the PA-$R^2$ metric, the differences become more pronounced, and interesting trends appear (Fig. 2c, 2d). The PA-$R^2$ metric illuminates the surprising difficulty in predicting the h-index in short time periods of 1-2 years, a reversal from the $R^2$ plots where short time periods appear very predictable. This is intuitive as an author's h-index can only increase in integral jumps and tends to grow slowly; while the long-term cumulative effects of these changes are predictable the short term is much less so. Note also that the PA-$R^2$ metric suggests that the GBRT offer a notable improvement over the other machine learning models (Fig. 2d).

Acuna et al. (2012) consider the task of predicting author h-index using elastic-net regularized linear regression on a crowd-sourced data set of neuroscientists; Table 2 displays their results alongside ours. Our model substantially outperforms theirs with a 50% relative improvement in $R^2$ when predicting 10 years into the future. Note, however, that the data sets are distinct. Their data set was pieced together from multiple evolving sources and is not publicly available as a cohesive whole; as such, we could not run our models on their data set.

While $R^2$-type measures are popular, we prefer Mean Absolute Percentage Error (MAPE), which averages the percentage error of each prediction. MAPE is defined as

$$\frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_{i,j}-\hat{y}_{i,j}}{y_{i,j}}\right|.$$

Here, the $i$th summand $|(y_{i,j}-\hat{y}_{i,j})/y_{i,j}|$ is the absolute percentage error of the $i$th prediction, and the MAPE is the average of these errors. Note that a smaller MAPE is better, the opposite of $R^2$.

**Table 2: Unadjusted $R^2$ values for 1, 5, and 10 year author h-index predictions. Our GBRT model attains substantially higher $R^2$ scores than those reported by Acuna et al. (2012), especially so when predicting 10 years into the future.**

| Years out | 1 | 5 | 10 |
|---|---|---|---|
| Acuna et al. (2012) | 0.92 | 0.67 | 0.48 |
| GBRT | 0.93 | 0.84 | 0.72 |
| Relative improvement | 3% | 24.8% | 50.6% |

Examining Figure 3b we see that, in terms of MAPE, the strengths of GBRT over the other models become even more apparent.

Beyond its intuitive description, MAPE has two distinct advantages over the $R^2$ measures. First it normalizes, on a per paper basis, the error in the prediction. Without this normalization, being off by three is equally poor for an author with an h-index of one as for an author with an h-index of fifty, a result that makes little intuitive sense. Moreover, unlike MAPE, $R^2$ measures are highly sensitive to outliers and can change dramatically when experiencing a small number of poor predictions. The MAPE produces easily interpretable results; indeed, the h-index is seen to be surprisingly predictable, even when predicting 10 years into the future, the GBRT are within ±19% of the truth on average (Fig. 3a).

## 4 PAPER IMPACT

The early history of study into the citation graph of papers was concerned with the describing the observed power-law distribution of citation counts. The first major success in modeling this phenomenon came from Price [18] who modeled the probability that a newly published paper, $p_{new}$, would cite some other paper, $p_{old}$, as being proportional to the number of citations $P_{old}$ had at the time of $p_{new}$'s publication. Price showed that a network growing with this "rich get richer" mechanism resulted in node degrees following a power law distribution closely mimicking that observed in real citation networks. This model was later rediscovered and popularized by Barabási and Albert [2] who coined the mechanism *preferential attachment.* This preferential attachment model has recently inspired probabilistic models for predicting citation counts of individual papers [20, 22]. One such predictive model describes citation trajectories using a *Reinforced Poisson Process* (RPP). In the RPP model, obtaining a citation increases the probability of receiving a citation in the future, a type of self-reinforcement analogous to the notion of preferential attachment [20]. In particular, the RPP models $C_p(t)$, the number of citations a paper $p$ has attained by time $t > 0$ after its publication, as a Poisson process with rate function

$$r_p(t) = \lambda_p \cdot f_p(t \mid \theta_p) \cdot (C_p(t) + m)$$

where $\lambda_p$ is a fitness parameter, $f_p$ is a non-negative temporal decay function with parameters $\theta_p$, and $m$ is a positive integer representing initial visibility. The parameters of the above model can then be inferred by maximum likelihood estimation. As maximum likelihood estimation is prone to overfitting, this model can be naturally augmented by using the Bayesian framework where $\lambda_p$ is assumed to be generated from some prior distribution. In our case we assume that $\lambda_p$ is generated by a Gamma$(\alpha, \beta)$ distribution. This prior substantially reduces the number of parameters that have to be estimated and helps mitigate the effect of overfitting [20]. Unfortunately, even this updated model does not perform quite as well as one might expect [23]. In order to improve the accuracy of this elegant model, we consider the following three modifications.

(i) The RPP model requires knowledge of the exact date when papers are published, we extend it to the more realistic setting where only the publication year is known.
(ii) We employ regularization to help mitigate the model's propensity to overfit.
(iii) Instead of requiring that the Gamma prior parameters, $\alpha$ and $\beta$, be shared across all papers, we allow them to be the output of a fully connected single layer neural network taking as input the same features we use in the below machine learning models. This allows more refined information about the paper to inform our prior knowledge.

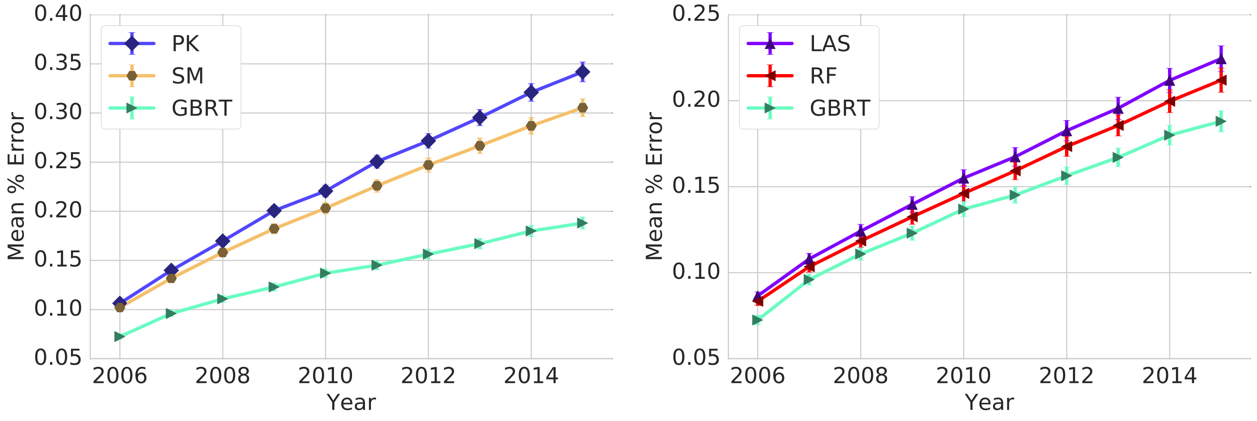
For details on how the above three items are accomplished, see Appendix A. These changes leave us with two candidate models; the first, which we call an RPPNet, applies all three modifications leveraging the features we extract, while the second, which we call an RPP, only implements the first two modifications.

We also consider a machine learning approach and extract 63 features for each paper exclusively using information available in 2005. We then use these features to train the same collection of regression models described for author h-index prediction to predict citation counts in the years 2006-2015. We filter our data to only include papers having received at least five citations before the end of 2005, a minimum threshold of impact. By simply replacing author h-indices with paper citations we adapt the MAPE, PA-$R^2$, and $R^2$ measures to the paper impact setting.

For a test set of size 10,000 papers, we plot the MAPE for all of the above models in Figure 4. As for author impact prediction, GBRT substantially outperform the baseline models, a difference of almost 15 percentage points after 10 years (Fig. 4a). Among all models, GBRT are consistently the best with the RPPNet a close second (Fig. 4b). Given that the RPPNet is more easily interpretable than GBRT and naturally produces error estimates along with its predictions, its slight loss in performance when compared to the GBRT may be acceptable. Up, we do not display plot of $R^2$ and PA-$R^2$; instead we note that the GBRT out perform all other models but, surprisingly, the simple SM model performs almost as well as the GBRT. This performance is, however, largely a reflection of these measures' sensitivity to outliers. When we remove 50 of the best and worst predictions of each model from their evaluations (1% of the test data) the GBRT outperform the SM models in PA-$R^2$ by $\approx 0.1$ (0.76 v.s. 0.66) when predicting 10 years into the future.

## 5 FACTORS CONTRIBUTING TO PREDICTION

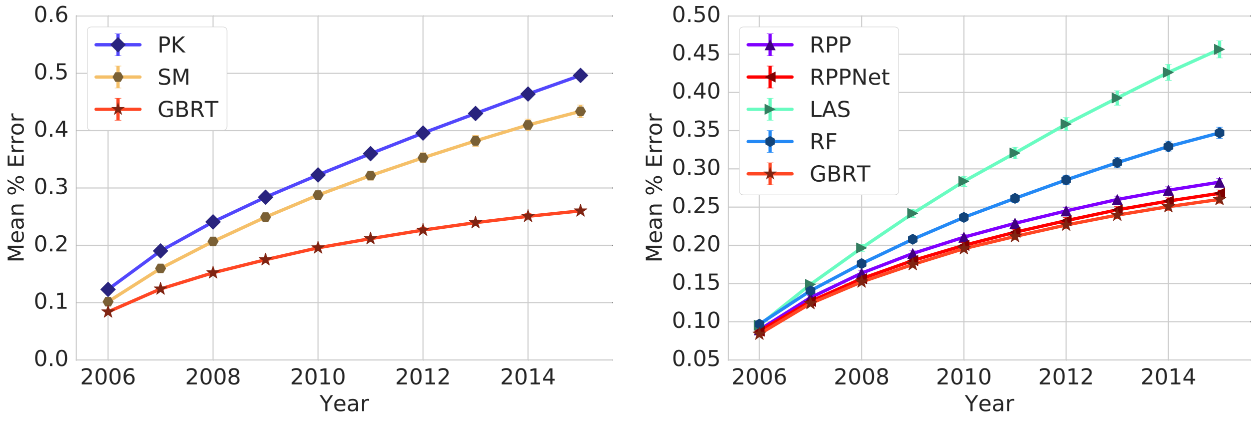
One notable omission from our analyses up to this point has been a discussion of the effect of author career and paper age on predictive performance. In particular, one may expect that citation rate of a paper stabilizes with age and thus, citation counts of older papers may become easier to predict. Similarly, an author with a well-established career should have a more stable h-index than that of a relative newcomer. To address this question, Figure 5 plots the percentage error of our GBRT predictions after 10 years for each



**(a) Baseline models & GBRT**

**(b) ML models**

**Figure 3: MAPE for author h-index prediction with 95% confidence intervals, these intervals are very narrow. Averaging over years, the GBRT obtain an error of 0.138 in comparison to 0.211 for the SMs and 0.151 for the RFs.**



**(a) MAPE, baseline models & GBRT**

**(b) MAPE, machine learning models & GBRT**

**Figure 4: MAPE for paper citation prediction with 95% confidence intervals, these intervals are very narrow. As with author h-index prediction, the GBRT outperform all other models; averaging over years, the GBRT obtain an error of 0.192 in comparison to 0.29 for SMs and 0.204 for RPPNets.**

author and paper in our test sets when stratifying by age.[3] Notice that while the errors in our predictions seem to be concentrated near zero, suggesting a lack of bias, we observe a substantially larger variability in the accuracy of our predictions for the younger authors and papers. This exactly fits our above intuition that older authors and papers are more predictable. Figure 5 also displays the MAPE for authors and papers for each of the different age groups. It's interesting to note that, for both authors and papers, the predictivity of citations and h-indices, as measured by the MAPE, increases rapidly in the first few years and then begins to plateau. Notice that, for instance, authors with a 25 year old career by 2005 have a MAPE of 0.13 which is only four percentage points less than those authors with a 10 year old career. This suggests that there is an inherent variability in citation counts and h-indices that cannot be explained by our model even in the best case. This motivates a need for future work developing features beyond those present in the citation history.

Beyond understanding the variability in our predictions, we are also interested in which features contributed most to our predictions. In order to quantify this feature importance, we measure the dependence between our features and the observed scientific impact measure using the $t^*$ statistic, a non-parametric measure of correlation [3, 24]. We use $t^*$ as it captures *any* dependence

[3] Recall that we previously restricted our data set to only those authors with an h-index of $\geq$ 4 and with a career length between 5-12 years by 2005. To form these new predictions we fit a GBRT model where we allowed authors of any age (but still required an h-index of $\geq$ 4).

structure between two variables, unlike, for example, Pearson correlation which only captures linear trends. We will focus only on the features contributing to the h-index predictions (see Fig. 6), since the results are very similar for paper citation prediction. Perhaps surprisingly, the relative importance of several features change when considering predictions at different forecasting periods. For instance, when predicting author h-index in 2006 the best feature is clearly the author's h-index in 2005. But when predicting the h-index in 2015, the h-index in 2005 is less important than the number of papers published in the years 2004-2005 (Fig. 6). One might expect that the papers an author publishes today will take several years before they accumulate enough citations to influence the author's h-index; but once sufficiently many years have passed, an authors h-index is strongly determined by those papers. These results mirror the trends seen by Acuna et al. [1].

## 6 DISCUSSION AND FUTURE WORK

While the above results suggest that scientific impact prediction is possible, even for a ten-year horizon, we should stress that our metrics assess average trends and do not apply to each author and paper uniformly. Indeed, as we discussed in Section 5, factors such as author and paper age play an important role in determining the accuracy of our predictions; we observed the intuitive result that the variability in our prediction decreases with the age of papers and length of researchers' careers. Somewhat surprisingly, the variability in our predictions does not appear to tend to zero for very old papers and authors. This suggests that there is still room for new features and modeling techniques to improve our upon our predictive performance in future work. We suspect, for instance, that features describing the topic of a paper may result in substantial predictive gains, especially for those relatively young authors and papers. The need for such features is highlighted by the work of Newman [14], who showed that the preferential attachment model predicts a substantial first-mover advantage for those publishing in a new area. Beyond developing new features, we also expect that gains can be made by more directly modeling citation events; while the RPP and RPPNet models bring us part of the way it is clear that they do not capture all of the underlying dynamics, moreover, we still lack such a predictive probabilistic model for author h-indices and citations. Finally, as it may soon be the case that decisions are being made on the basis of impact predictions, it is clear that it is no longer enough to produce a single-number prediction of impact. Instead, future techniques must be able to reliably assess the confidence in their predictions, indeed this is one reason why combining machine learning techniques with probabilistic modeling approaches, such as the RPPNet, is so appealing.

Of course, any attempt to summarize scientific impact is limited. Existing measures, such as citation counts and h-indices, do not provide a comprehensive assessment of a paper or author. Instead, they provide a signal that helps to inform our understanding of a paper's or author's impact. In this way, our results suggest that scientific impact predictions may be a useful tool, among many, in guiding our focus to where it will be most fruitful.

## A MODIFICATIONS TO THE REINFORCED POISSON PROCESS MODEL

Recall that in the reinforced poisson process models $C_p(t)$, the number of citations a paper $p$ has at time $t > 0$ after it's publication, as a poisson process with rate function

$$r_p(t) = \lambda_p \; f(t \mid \theta_p) \; (C_p(t) + m)$$

where $\lambda_p$, $\theta_p$ and $m$ are parameters and $f$ is a temporal decay function [20]. While the $\lambda_p$ and $\theta_p$ parameters are inferred by maximum likelihood, [20] found that performance is only weakly dependent on the value of $m$ and thus we follow this prior work and simply let $m = 10$ be fixed. To help reign in the problem of overfitting one can place a Gamma$(\alpha, \beta)$ prior upon the $\lambda_p$ parameters. Recall from Section 4 that we make three modifications to the above model, these modifications are described below.

### A.1 Discrete Time

As a clear extension of the continuous time RPP model we model $C_p(n)$, the number of citations a paper $p$ has $n \geq 1$ years after it's publication, as a discrete time poisson process with rate function

$$r_p(n) = \lambda_p \; (C_p(n-1) + m) \int_{n-1}^{n} f(t \mid \theta_p) \; \mathrm{d}t.$$

Note that we have integrated above as $r_p(n)$ represents the mean for the entire time period between $n-1$ and $n$. Following prior work we let $f(t \mid \theta) = \frac{1}{\sqrt{2\pi}\sigma t} \exp(-\frac{1}{2\sigma^2}(\ln t - \mu)^2)$ so that $f$ a log-normal probability density function for $\theta = (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_{>0}$.

To avoid degenerate cases we will always assume that $\mu \geq -1$ and $\sigma \geq 0.5$. Let $\Phi_\theta$ be the log-normal cumulative distribution function corresponding to $f(t \mid \theta)$ and for all $i \geq 1$ let $\Delta\Phi_\theta(i) = \Phi_\theta(i) - \Phi_\theta(i-1)$. Using this notation we may rewrite the rate function as $r_p(n) = \lambda_p \; (C_p(n-1) + m)\Delta\Phi_\theta(n)$.

The above definition gives us, for all $n \geq 1$, the following, self-reinforcing, relationship,

$$C_p(n) - C_p(n-1) \mid C_p(n-1) \sim \text{Poisson}(\lambda_p \; (C_p(n-1) + m)\Delta\Phi_\theta(i))$$

where $C_p(0) = 0$. For the sake of simplifying notation we will, for the moment, drop the $p$ subscripts.

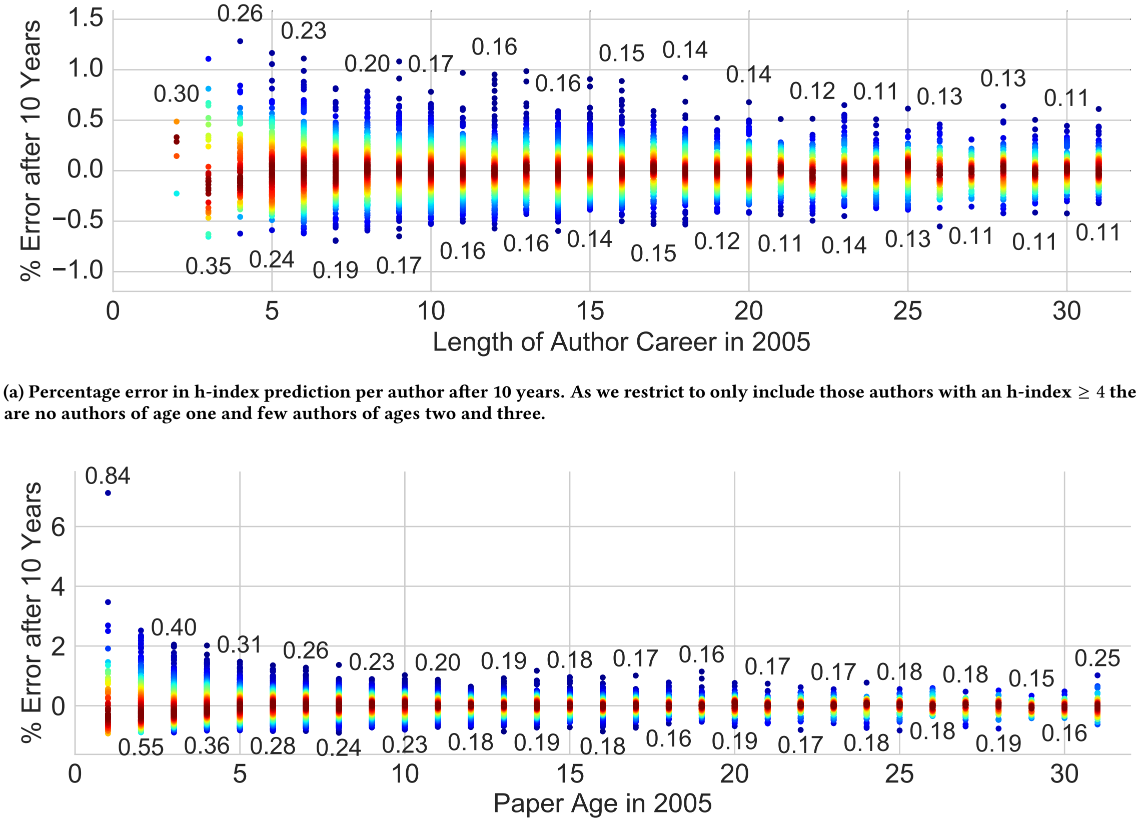
Suppose we observe per-year citation counts $C(1) - C(0) = d_1, \; \ldots, \; C(n) - C(n-1) = d_n$ and wish perform maximum likelihood estimation of $\lambda, \theta$. To do this we will first need an explicit form of the likelihood function. By definition,

$$P(C(i) - C(i-1) = d_i \mid d_1, \ldots, d_{i-1}) = e^{-\lambda C_{i-1} \Delta\Phi_\theta(i)} \lambda^{d_i} \Delta\Phi_\theta(i)^{d_i} \frac{C_{i-1}^{d_i}}{d_i!}.$$

From the above it is easy to see that the likelihood of the observations is simply

$$\begin{aligned} L(\lambda, \theta \mid d_1, \ldots, d_n) &= \prod_{i=1}^{n} e^{-\lambda C_{i-1} \Delta\Phi_\theta(i)} \lambda^{d_i} \Delta\Phi_\theta(i)^{d_i} \frac{C_{i-1}^{d_i}}{d_i!} \\ &= \exp\left(-\lambda \sum_{i=1}^{n} C_{i-1} \Delta\Phi_\theta(i)\right) \; \lambda^{\sum_{i=1}^{n} d_i} (\prod_{i=1}^{n} \Delta\Phi_\theta(i)^{d_i} \frac{C_{i-1}^{d_i}}{d_i!}) \end{aligned}$$

Note that $\sum_{i=1}^{n} d_i$ is equal to the total number of citations by time $n$, call this quantity $N$. Given this explicit form of the likelihood,



**(a) Percentage error in h-index prediction per author after 10 years. As we restrict to only include those authors with an h-index $\geq 4$ there are no authors of age one and few authors of ages two and three.**

**(b) Percentage error in citation count prediction per paper after 10 years.**

**Figure 5: For every author and paper in our test sets, we plot the percentage error of our GBRT predictions after 10 years. We have separated authors and papers into groups based on their respective ages at the end of 2005, that is, papers published in 2004 and 2005 would, respectively, be considered to have ages 2 and 1 at the end of 2005. Author career ages are based off of the date of their first publications. As many points overlap in these plots we have, within each age group, colored the points using a kernel density estimate, dark red corresponds many overlapping points while dark blue corresponds to few overlapping points. Above and below each group we have also included the MAPE when restricting the papers and authors within that group, for example, the MAPE among authors with a career length of 5 is 24%.**

we may estimate $\lambda, \theta$ directly by maximum likelihood estimation. In particular, note that the log-likelihood has the form

$$\mathcal{L}(\lambda, \theta) = \log L(\lambda, \theta \mid d_1, ..., d_n)$$
$$= -\lambda \sum_{i=1}^{n} C_{i-1} \Delta\Phi_\theta(i) + N \log \lambda + \sum_{i=1}^{n} d_i \log(\Delta\Phi_\theta(i)) + \sum_{i=1}^{n} \log(\frac{C_{i-1}^{d_i}}{d_i!}),$$

differentiating we find that

$$0 = \frac{\partial}{\partial \lambda} \mathcal{L}(\lambda, \theta) \iff \lambda = \frac{N}{\sum_{i=1}^{n} C_{i-1} \Delta\Phi_\theta(i)}.$$

Plugging this optimum value, $\lambda^* = \frac{N}{\sum_{i=1}^{n} C_{i-1} \Delta\Phi_\theta(i)}$, into $\mathcal{L}$ gives

$$\mathcal{L}(\lambda^*, \theta) = \sum_{i=1}^{n} d_i \log(\Delta\Phi_\theta(i)) - N \log(\sum_{i=1}^{n} C_{i-1} \Delta\Phi_\theta(i)) + const.$$

We now differentiate the above in $\mu, \sigma$. Let $\phi_\theta$ be the density corresponding to a Normal$(\mu, \sigma^2)$ distribution. We then have that

$$\frac{\partial}{\partial \mu} \mathcal{L}(\lambda^*, \theta) = \sum_{i=1}^{n} (\lambda^* C_{i-1} - \frac{d_i}{\Delta\Phi_\theta(i)})(\phi_\theta(\log(i)) - \phi_\theta(\log(i-1))),$$

$$\frac{\partial}{\partial \sigma} \mathcal{L}(\lambda^*, \theta) = \sum_{i=1}^{n} (\lambda^* C_{i-1} - \frac{d_i}{\Delta\Phi_\theta(i)})(\frac{\log(i) - \mu}{\sigma} \phi_\theta(\log(i))$$
$$- \frac{\log(i-1) - \mu}{\sigma} \phi_\theta(\log(i-1))).$$

Using the above we can find $\mu, \sigma$ using derivative based optimization approaches. Using the discovered $\theta = (\mu, \sigma)$ we may predict into the future using the mean of the poisson process given the prior observations. Let $c_n(t \mid \lambda, \theta) = E[C(n+t) \mid C(1), ..., C(n)]$,

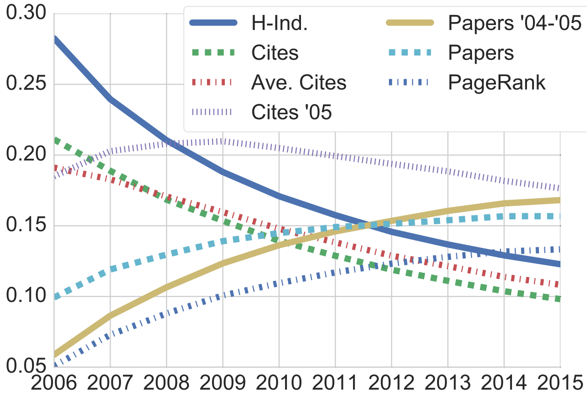


**Figure 6: $t^*$ between feature values and observed h-indices in the years 2006-2015, larger values mean more dependent. All features use data available in 2005. The features are H-ind. (h-index), Cites (total citations), Ave. Cites (average citations per year), Cites '05 (citations in 2005), Papers '04-'05 (papers published between 2004-2005), Papers (total number of papers published), PageRank (author PageRank in coauthor network). We see that the importance of authors' h-indices in 2005 decreases while the importance several other predictors, e.g. the number of papers an author published in 2004-2005, increases.**

then we may compute $c_n(t \mid \lambda, \theta)$ recursively to find that, for $t \geq 1$,

$$c_n(t \mid \lambda, \theta) = (C(n) + m) \prod_{i=1}^{t} (1 + \lambda \Delta\Phi_\theta(n+i)) - m.$$

## A.2 Prior Extensions and Regularization

We now place a Gamma$(\alpha, \beta)$ prior on $\lambda$ and compute the marginalized likelihood. To simplify notation we will let $A = (\prod_{i=1}^{n} \Delta\Phi_\theta(i)^{d_i} \frac{C_{i-1}^{d_i}}{d_i!})$. As in the previous section, we suppose we have observed $C(1) - C(0) = d_1, \ \ldots, \ C(n) - C(n-1) = d_n$ for which we may write the marginalized likelihood

$$\begin{aligned} L(\theta, \alpha, \beta \mid d_1, \ldots, d_n) &= \int_0^\infty L(\lambda, \theta \mid d_1, \ldots, d_n) \ \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \, \mathrm{d}\lambda \\ &= A \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+N)}{(\beta + \sum_{i=1}^{n} C_{i-1}\Delta\Phi_\theta(i))^{\alpha+N}}. \end{aligned}$$

From the above we also immediately observe that the posterior distribution of $\lambda$ given the observation is $\lambda \mid d_1, \ldots, d_n \sim \text{Gamma}(\alpha + N, \beta + \sum_{i=1}^{n} C_{i-1}\Delta\Phi_\theta(i))$ which allows us to easily compute the posterior mean,

$$\overline{\lambda} = E[\lambda \mid d_1, \ldots, d_n] = \frac{\alpha + N}{\beta + \sum_{i=1}^{n} C_{i-1}\Delta\Phi_\theta(i)}.$$

Now taking the log of the marginalized likelihood gives

$$\begin{aligned} \mathcal{L}(\theta, \alpha, \beta) = B &+ \sum_{i=1}^{n} d_i \log(\Delta\Phi_\theta(i)) + \alpha \log\beta - \log\Gamma(\alpha) \\ &+ \log\Gamma(\alpha+N) - (\alpha+N)\log(\beta + \sum_{i=1}^{n} C_{i-1}\Delta\Phi_\theta(i)) \end{aligned}$$

where $B = \sum_{i=1}^{n} \log(\frac{C_{i-1}^{d_i}}{d_i!})$ is constant with respect to the parameters $\theta, \alpha, \beta$. Now letting $\psi$ be the digamma function we have that

$$\begin{aligned} \frac{\partial}{\partial\alpha}\mathcal{L}(\theta, \alpha, \beta) &= \log\beta - \psi(\alpha) + \psi(\alpha+N) - \log(\beta + \sum_{i=1}^{n} C_{i-1}\Delta\Phi_\theta(i)), \\ \frac{\partial}{\partial\beta}\mathcal{L}(\theta, \alpha, \beta) &= \frac{\alpha}{\beta} - \frac{\alpha+N}{\beta + \sum_{i=1}^{n} C_{i-1}\Delta\Phi_\theta(i)} = \frac{\alpha}{\beta} - \overline{\lambda}, \\ \frac{\partial}{\partial\mu}\mathcal{L}(\theta, \alpha, \beta) &= \sum_{i=1}^{n} (\overline{\lambda} C_{i-1} - \frac{d_i}{\Delta\Phi_\theta(i)})(\phi_\theta(\log(i)) - \phi_\theta(\log(i-1))), \\ \frac{\partial}{\partial\sigma}\mathcal{L}(\theta, \alpha, \beta) &= \sum_{i=1}^{n} (\overline{\lambda} C_{i-1} - \frac{d_i}{\Delta\Phi_\theta(i)})(\frac{\log(i) - \mu}{\sigma}\phi_\theta(\log(i)) \\ &\qquad - \frac{\log(i-1) - \mu}{\sigma}\phi_\theta(\log(i-1))). \end{aligned}$$

It is interesting to note that, when comparing the derivatives in $\mu, \sigma$ to those computed in the previous section, we have simply replaced the optimal $\lambda^*$ in that setting with posterior mean $\overline{\lambda}$.

When performing maximum likelihood inference in the last section it was sufficient to consider each paper individually as no two papers shared any common parameters. In our current setting however, all papers share the same $\alpha, \beta$ parameters and hence we will need to perform maximum likelihood estimation for all papers simultaneously. To this end let $\mathcal{P}$ be a collection of papers. For each $p \in \mathcal{P}$ suppose we have observations $C_p(1) = d_1^p, \ C_p(2) - C_p(1) = d_2^p, \ \ldots, \ C_p(n_p) - C_p(n_p - 1) = d_{n_p}^p$, and let $\mathcal{L}_p(\theta_p, \alpha, \beta)$ be the log-likelihood for the individual paper $p$. Then the log-likelihood of all papers simultaneously is simply the sum

$$\mathcal{L}_{\mathcal{P}}(\alpha, \beta, \{\theta_p\}_{p\in\mathcal{P}}) = \sum_{p\in\mathcal{P}} \mathcal{L}_p(\theta_p, \alpha, \beta).$$

It is easy to compute gradients of $\mathcal{L}_{\mathcal{P}}$ using the derivates we have computed for the individual $\mathcal{L}_p$ and thus performing maximum likelihood inference for the parameters $\alpha, \ \beta$, and $\{\theta_p\}_{p\in\mathcal{P}}$ can be done using any gradient based optimizer.

Now suppose that $\alpha, \beta, \theta_p$ are fixed. Obtaining future predictions is straightforward by using posterior means. By iterated conditioning one may easily check that

$$c_{p,n_p}(t \mid \alpha, \beta, \theta_p) = E[c_n(n_p + t \mid \lambda_p, \theta_p) \mid C_p(1), \ldots, C_p(n_p)].$$

Now for fixed $\lambda_p$ we can compute $c_n(n_p + t \mid \lambda_p, \theta_p)$ using the results in the previous section. Hence we can approximate the above expectation to arbitrary precision using a Monte-Carlo strategy. Namely we draw samples from the posterior distribution of $\lambda_p$, Gamma$(\alpha + N, \beta + \sum_{i=1}^{n} C_{i-1}\Delta\Phi_\theta(i))$, compute $c_n(n_p + t \mid \lambda_p, \theta_p)$ for each of these samples and then averaging the results.

As we have found that simply using maximum likelihood inference results in overfitting we consider adding a regularization penalty of the form $-\gamma(\frac{\alpha}{\beta})^2$ to the optimization procedure where

$\gamma \geq 0$ is a hyper parameter. In particular, rather than attempting to maximize $\mathcal{L}_{\mathcal{P}}$ we maximize

$$\mathcal{L}^*_{\mathcal{P}}(\alpha, \beta, \{\theta_p\}_{p\in\mathcal{P}}) = \frac{1}{|\mathcal{P}|}\sum_{p\in\mathcal{P}} \mathcal{L}_p(\theta_p, \alpha, \beta) - \gamma(\frac{\alpha}{\beta})^2.$$

Here $\alpha/\beta$ is equal to the mean of the prior distribution Gamma$(\alpha, \beta)$ and so this penalty discourages the prior distribution from being too large on average. In practice, $\gamma$ is chosen by cross validation.

## A.3 Neural Network Structure

In order to allow more subtle information to influence our prior distribution we also consider letting $\alpha, \beta$ be learned functions of features extracted from each paper. For each paper $p$ let $x_p \in \mathbb{R}^k$ be a collection of features corresponding to the paper. Then we consider $\alpha, \beta$ as functions of the $x_p$, written as $\alpha(x_p), \beta(x_p)$, so that $\lambda_p \sim$ Gamma$(\alpha(x_p), \beta(x_p))$. We learn the functions $\alpha, \beta$ as the output of a single layer, fully connected, neural network with softplus non-linearities, by maximizing the penalized log-likelihood

$$\mathcal{L}^{**}_{\mathcal{P}}(\alpha, \beta, \{\theta_p\}_{p\in\mathcal{P}}) = \frac{1}{|\mathcal{P}|}\sum_{p\in\mathcal{P}} \left(\mathcal{L}_p(\theta_p, \alpha(x_p), \beta(x_p)) - \gamma\frac{\alpha(x_p)^2}{\beta(x_p)^2}\right).$$

As was shown empirically in the main paper, allowing $\alpha, \beta$ to depend on $x_p$ can improve performance.

## REFERENCES

[1] Daniel E Acuna, Stefano Allesina, and Konrad P Kording. 2012. Future impact: Predicting scientific success. *Nature* 489, 7415 (sep 2012), 201–202.

[2] Albert-László Barabási and Réka Albert. 1999. Emergence of Scaling in Random Networks. *Science* 286, 5439 (1999), 509–512.

[3] Wicher Bergsma and Angelos Dassios. 2014. A consistent test of independence based on a sign covariance related to Kendallfis tau. *Bernoulli* 20, 2 (05 2014), 1006–1028.

[4] Maria Bras-Amorós, Josep Domingo-Ferrer, and Vicenç Torra. 2011. A bibliometric index based on the collaboration distance between cited and citing authors. *Journal of Informetrics* 5, 2 (2011), 248 – 264.

[5] Leo Breiman. 2001. Random Forests. *Mach. Learn.* 45, 1 (Oct. 2001), 5–32.

[6] Carlos Castillo, Debora Donato, and Aristides Gionis. 2007. *String Processing and Information Retrieval: 14th International Symposium, SPIRE 2007 Santiago, Chile, October 29-31, 2007 Proceedings.* Springer Berlin Heidelberg, Berlin, Heidelberg, Chapter Estimating Number of Citations Using Author Reputation, 107–117.

[7] T. Chakraborty, S. Kumar, P. Goyal, N. Ganguly, and A. Mukherjee. 2014. Towards a stratified learning approach to predict future citation counts. In *Digital Libraries (JCDL), 2014 IEEE/ACM Joint Conference on.* 351–360.

[8] J. Chen and C. Zhang. 2015. Predicting citation counts of papers. In *Cognitive Informatics Cognitive Computing (ICCI*CC), 2015 IEEE 14th International Conference on.* 434–440.

[9] Yuxiao Dong, Reid A. Johnson, and Nitesh V. Chawla. 2015. Will This Paper Increase Your H-index?: Scientific Impact Prediction. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM '15).* ACM, New York, NY, USA, 149–158.

[10] Leo Egghe. 2013. Theory and practise of the g-index. *Scientometrics* 69, 1 (2013), 131–152.

[11] Jerome H. Friedman. 2001. Greedy function approximation: A gradient boosting machine. *Ann. Statist.* 29, 5 (10 2001), 1189–1232.

[12] J. E. Hirsch. 2005. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America* 102, 46 (2005), 16569–16572.

[13] Peter J. Huber. 1964. Robust Estimation of a Location Parameter. *Ann. Math. Statist.* 35, 1 (03 1964), 73–101.

[14] Newman, M. E. J. 2009. The first-mover advantage in scientific publication. *EPL* 86, 6 (2009), 68001.

[15] Masoumeh Nezhadbiglari, Marcos André Gonçalves, and Jussara M. Almeida. 2016. Early Prediction of Scholar Popularity. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries (JCDL '16).* ACM, New York, NY, USA, 181–190. DOI:http://dx.doi.org/10.1145/2910896.2910905

[16] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. *The PageRank Citation Ranking: Bringing Order to the Web.* Technical Report 1999-66. Stanford InfoLab. Previous number = SIDL-WP-1999-0120.

[17] Orion Penner, Raj K Pan, Alexander M Petersen, Kimmo Kaski, and Santo Fortunato. 2013. On the Predictability of Future Impact in Science. *Scientific Reports* 3 (oct 2013), 3052.

[18] Derek De Solla Price. 1976. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science* 27, 5 (1976), 292–306.

[19] Emre Sarigöl, René Pfitzner, Ingo Scholtes, Antonios Garas, and Frank Schweitzer. 2014. Predicting scientific success based on coauthorship networks. *EPJ Data Science* 3, 1 (2014), 1–16.

[20] Huawei Shen, Dashun Wang, Chaoming Song, and Albert-László Barabási. 2014. Modeling and Predicting Popularity Dynamics via the Poisson Processes. *AAAI Conference on Artificial Intelligence* (2014).

[21] Robert Tibshirani. 1994. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society, Series B* 58 (1994), 267–288.

[22] Dashun Wang, Chaoming Song, and Albert-László Barabási. 2013. Quantifying Long-Term Scientific Impact. *Science* 342, 6154 (2013), 127–132. DOI:http://dx.doi.org/10.1126/science.1237825

[23] Jian Wang, Yajun Mei, and Diana Hicks. 2014. Comment on "Quantifying long-term scientific impact". *Science* 345, 6193 (2014), 149–149.

[24] Luca Weihs, Mathias Drton, and Dennis Leung. 2016. Efficient computation of the Bergsma–Dassios sign covariance. *Computational Statistics* 31, 1 (2016), 315–328.

[25] Jevin D. West, Theodore C. Bergstrom, and Carl T. Bergstrom. 2010. The Eigenfactor Metrics$^{TM}$: A network approach to assessing scholarly journals. *College & Research Libraries* 71, 3 (2010), 236–244.

[26] Rui Yan, Congrui Huang, Jie Tang, Yan Zhang, and Xiaoming Li. 2012. To Better Stand on the Shoulder of Giants. In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '12).* ACM, New York, NY, USA, 51–60.

[27] Xiaodan Zhu, Peter Turney, Daniel Lemire, and André Vellino. 2015. Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science and Technology* 66, 2 (2015), 408–427. DOI:http://dx.doi.org/10.1002/asi.23179