CSE 481DS Final Report

Lirui Wang, Henry Hung, Eric Chan
Department of Computer Science
University of Washington

Abstract

When the CS community cites and reviews papers, the reputation of authors is implicitly considered. Researchers having a history of high citation counts, good conference publication tracks, in-field reputations, and well-known industry partners tend to get more citations for their works. This work hypothesizes and studies the disentanglement of confounding variables in this observation as well as contrasting actual research potential with fame to help the CS community consider factors beyond author reputation when citing works. We filtered CS papers from the large-scale Microsoft Academic Graph (MAG) dataset and investigated the hypothesis from multiple perspectives. We mainly validated the correlations between previous reputation and future academic performance across subfields and time. We also validated our results to ensure construct, internal, and external validities. Overall, we found that high reputation often leads to high future citation growth but does not necessarily reflect high research potentials. Despite several limitations, our work encourages CS field to consider factors beyond author reputation when citing works.

1 Introduction

Recent years, we observed a trend where CS authors and papers in the "hot" cluster tend to get more citations. Research community appears to pay extremely high attention to works published and advertised by famous authors and institutions. This phenomenon is intuitive since researchers who have published good works are expected to continue their research work. Metrics such as H-Index and its variants [9; 2; 11; 5] are designed exactly for the purpose of measuring research performance and potential. Moreover, the entire field of scientometrics [13] concerns measurement and evaluation of research performance. However, the recent development of the Computer Science (CS) and its subfield make us wonder if the reputation alone plays a role in citation growth for an author. Answering this question has some important implications on disentangling the confounding variables in the citation trend as well as contrasting research potential with author fame. It can also encourage the CS field to consider factors beyond author reputation when citing works.

In this work, we leverage large-scale cloud computing for the Microsoft Academic Graph (MAG) dataset [16; 7; 10; 8] that consists of components such as citations, field of study, institutions, etc. We filtered the dataset and created custom data tables to efficiently organize and generate information required in our analysis.

We propose two hypotheses at the core of our analysis. The first hypothesis (H1) is that authors that have high previous reputation get cited more for their future works. Despite appearing trivial, it questions the relationship between reputation and future citation growth, if we isolate the citations of the previous work and reduce the effects of confounding variables. Our results suggest that for the CS field, author previous h-index has a positive correlation with the future paper citations.

Our second hypothesis (H2) is that the relative research performance for an author's career often differs from its citation performance. Specifically we construct T-Split H-Index (Section 3.2) to measure relative research potential, which compares future work's citations compared to "core" early

work. We applies this analysis to both the entire CS field and also specific case study. We observe the quality work published in an CS author's career is not necessarily reflected by his reputation.

To validate our conclusion, we discovered that the result is also consistent across reputation proxy (construct validity), across time and subfield (external validity) and we analyze the confounding factors through causal inference (internal validity). Specifically we used index variants such as A/M/R index and cross check the correlations in different subfields and times. We use propensity score matching to eliminate the effects of confounding variables and provide visualizations for our results.

Overall, our contributions are three-fold: 1) We propose two hypothesis regarding authors' reputations with future citation growth and reseach potentials. 2) We use large-scale MAG dataset to confirm our hypothesis and present interpretations. 3) We discuss the validity of our conclusion as well as limitations and future works.

2 Dataset

For citation analysis in our project, we used Microsoft Academic Graph (MAG) [16; 7; 10; 8], backboned by Microsoft Academic Service (MAS). At the core of MAS are six types of entities that model the scholarly activities with a heterogeneous graph: field of study, author, institution, paper, venue, and event. To the best of our knowledge, it is the largest publicly available dataset of citation data and therefore suitable for us to study the relationships between reputation and scientometrics. Previous work [8] has shown that MAG has a good correlation with external datasets and good coverage across different domains, despite certain limitations to completeness. Moreover, the dataset provided in the Microsoft Academic Research provides large database dumps every week (overall about 350 GB text file) and has integrated functions such as forums and existing database access. Specifically, the dataset contains over 200 million papers, 1 billion references, and 100 million authors, etc. A detailed analysis of the dataset can be found in [8; 7] and the official website. We access the dataset with pyspark SQL through Databricks on Azure servers. Note in addition to quantities such as citations and H-Index, we have also looked into the semantic information of the dataset such as paper semantics, confidence for paper subfield, and institution ranks computed from the underlying graph.

Due to its large size and our focus on computer science papers, we have filtered out a subset of the dataset based on papers that have "Computer Science" as one of their subfields. The filtered dataset $\mathcal{D}$ contains 30 million authors, 20 million papers, 20 thousand institutions, etc (around 30GB). The citation counts of each papers at each year were then re-calculated solely from the paper publishing years and references among these papers. We present detailed analysis of the dataset in Appendix.8.1. We plotted out an overview of a few attributes paper counts for authors and subfields on Fig. 10. We observe a trend of skewed author and subfield publication distribution. To compare and understand subfields over time, we compare Machine Learning with Computer Hardware Fields in Fig. 10 and observe an upward trend of ML research in recent years.

In order to efficiently process such dataset and validate our research, we generate several custom data tables and save them locally. More details can be found at Appendix 8.1. The most common operations in this processing step is to join paper and author information such as citations with other tables such as conference and institutions, and then aggregate over time. We used distributed computing framework [3] to process these large joined dataset. We present our dependency graph demonstrating our pre-processing work on Fig. 11.

3 Analytical Approach

As a motivation, it is commonly observed and believed that researchers having a history of high citation counts, good conference publication tracks, and well-known industry partners affiliations tends to get more citations for their works. Therefore, we make two hypotheses in our work to present the phenomenon and strive to validate them with data:

- Hypothesis 1 (H1): Authors that have high previous reputation get cited more for their future works;
- Hypothesis 2 (H2): The relative research performance for an author's career often differs from its citation performance.

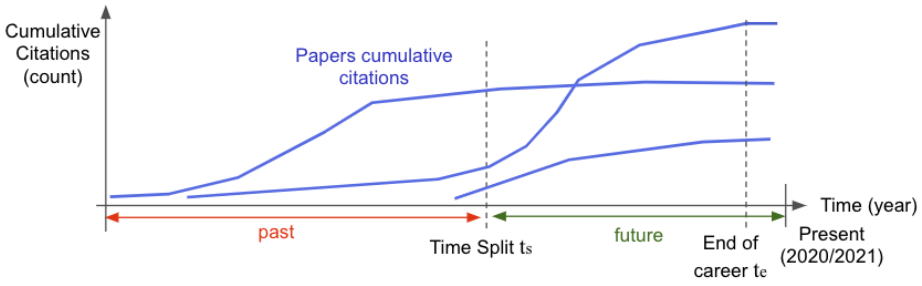


Figure 1: A timeline view of an author's career and T-Split point.

Specifically, we use the MAG dataset to model the correlations between reputation and future citations, as well as reputation and relative research potential. We also use an author's H-Index [9] and its variants [2; 11; 5] to represent his/her reputations. The H-Index [9] is defined as " A scientist has index h if h of his or her N papers have at least h citations each, and the other $(N - h)$ papers have $<= h$ citations each."

We then validate the findings from multiple perspectives, such as different indexes to represent reputations as construct validity, causal inference for internal validity, and subfield conditioning for external validity. We also discuss the assumptions and limitations of our work, and propose a few interesting future directions.

At the first look, using H-Index for construction of reputation make our hypotheses seems trivial, since H-Index is simply a function of citations, and previous performance surely has some correlations with future performance. However, It is not directly clear that this relationship holds if we isolate the citations of the previous work, reduce the effects of confounding variables, and normalize across factors such as subfield and time. For instance, a researcher could have a successful career start and becomes famous with a high H-Index. Then, external factors such as industry opportunities might influence his focus on the new quality works, resulting in a low future citation growth. Our research studies the correlation between reputation established from previous works and the citations of future work, as well as author-based research performance, which is a highly nontrivial problem given that the research ability and paper quality are hard to quantify.

Before we discuss our hypothesis and results, we briefly cover a few terms. Our time granularity for analysis is year, denoted by t. We use superscript to denote the measurement time of metrics (c for citation and h for H-Index) and subscript to indicate the time according to which a certain set of work are considered. For instance, c_{-}^{t-} denote citations before t of work published before t, c_{+}^{t+} denote total citations from year t to 2020/2021 of work published after t. For simplicity we overload the notation for paper citation list and total citations. We also denote $\bar{c}_{+}^{t+}$ to be the average citation growth over years. Despite the direct relationship with citations, H-Index has the advantage of jointly measuring productivity and work influence [13].

3.1 Hypothesis 1

Our first hypothesis aims to verify the intuition that higher reputation is correlated with higher future performance. To begin with, an author's career can be described as in Fig. 1. The plot records the cumulative citations of his/her papers over the years. Each blue line describes a paper's cumulative citation, and when all lines stop increasing, we consider the corresponding year to be the author's end of career t_e. To provide a context for reasoning about pasts and futures, we split each author's career into the "past" and "future" by finding the year t_s where this author achieves half of the total citations at t_e. The author's "current" reputation is then defined as his/her H-Index at t_s. On the other hand, future citations are evaluated by counting citations of future works and divide it by the time span until the author's end of career, namely $t_e - t_s$. Notice the citations received by the past works in the future are excluded to eliminate the effect that papers with high citation usually continue to received many citations. By doing so, we restrict our focus to citing behavior toward future works only, given the reputation gained from works in the past. Finally, the relationship between reputation and future citations is fitted against authors' data by local linear regression.

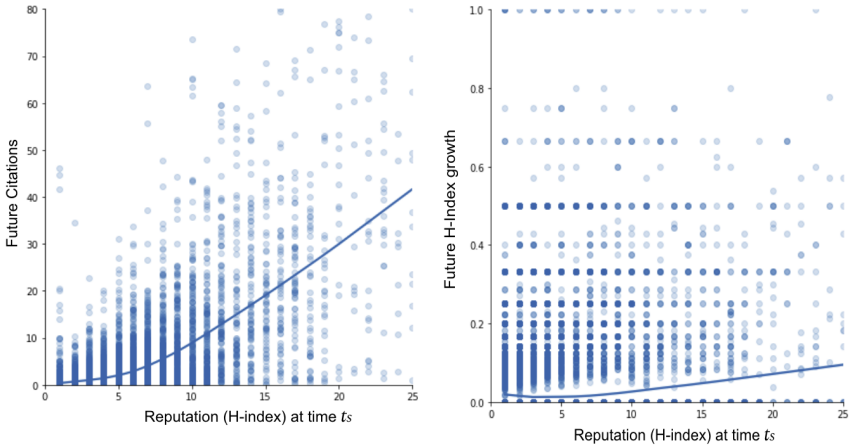


Figure 2: Left) Reputation vs future citation. Right) Reputation vs future H-Index growth. Our main result on Hypothesis 1: The correlation between reputation and future citation over the entire CS field. A similar pattern also holds for H-Index growth but with a lower extent.

Besides future citations, we also explore the future growth in H-Index, which we define as the increase of H-Index from t_s to the author's end of career t_e, divided by this time span. Similarly, future citations of papers in the past do not count toward the calculation of the H-Index at t_e. We call such calculation T-split H-Index h^t_{split}, which is introduced more formally in hypothesis 2. Here, $t = t_s$ and $h^{t_s}_{split}$ is intended to represent author's H-Index at time t_e, while keeping citations of past works fixed after t_s. Hence future H-Index growth gives us insight into the the degree to which future works are able to potentially improve the H-Index.

3.2 Hypothesis 2

We introduce T-Split H-Index h^t_{split} which is the H-Index from citation $c^{t+}_{+} \cup c^{t-}_{-}$. This simple definition is constructed to determine the relative research potential of an author at time t, and there are two ways to see it. Note that by the definition of H-Index, citations of future works c^{t+}_{+} needs to reach h^{t-}_{-} to increase H-Index, so it compares future work's citations compared to "core" early work. Alternatively, since this is the actual H-Index subtracting the contribution of the future citations of early work h^{t+}_{-}, it excludes the H-Index contribution of previous work before t. First note that it is expected to see the average future H-Index growth for h^{t+}_{+} of an author can decrease over years by definition of H-Index, verified in the left plot of Fig. 3. We study this effect across time, and normalize all author's career years between 0 and 1. This is in contrast to the monotocity of future citation growth $\bar{c}^{t+}_{+}$ with H-Index h^{t-}_{-} in H1. Moreover, we plot out the T-Split H-Index h^t_{split} with previous H-Index h^{t-}_{-}. We observe the quality work published in an CS author's career is not necessarily reflected by his reputation. The interesting implication for this phenomenon is that, given that his potential of publishing giant works (that meet his standard of good works) is not high, should the community still pays extremely high attention to his future paper? We give a specific author example in our finding section.

3.3 Construct Validity

Since we use H-Index, proposed in 2005 [9], to measure reputations. It is important to use convergent validity to show that different measures of the same construct correlate. Specifically, given a h-core citation set (the set of papers with citations great than the H-Index for an author), there are several other proposals that measure statistics on the papers in the h-core.

- A-index [11] is the mean number of citations of the papers in the h-core.
- R-Index [11] is the square root of the sum of the citations of the h-core papers.

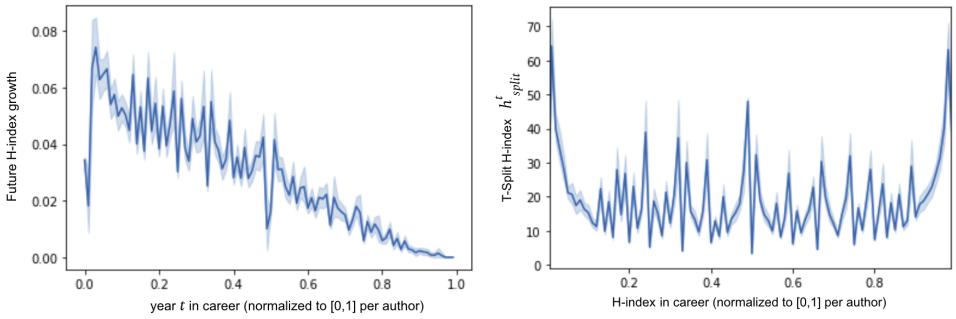


Figure 3: Left) The correlation between previous H-Index and future H-Index growth under normalized years for authors. Right) The correlation between previous H-Index and T-Split H-Index growth under normalized years for authors.

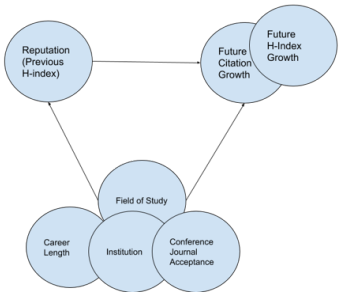
- M-Index [2] represents the median of the papers in the h-core to counteract highly skewed data.

There have been many comparisons of the H-Index with other indicators [2; 9; 4]. These comparisons often show that different indices are highly correlated but often focus on the productivity rather than the pure impact [13]. For instance, Dr. Durk Kingma has contributed Adam, VAE, and IAF as some current key concepts in Machine Learning. He publishes 25 papers in total, with an H-Index of 22, total citation over 80000, M-Index of 181, A-index of around 3900, and R index of 286. These numbers have different trade-offs and it is hard to claim there is a best candidate for measurement, which is also author-dependent. In this project, we ensure the construct validity of reputation by substituting A/R/M-indices for H-Index as the reputation, and making sure the result of Hypothesis 1 is reproduced. By trying a few variants, we expect some of them to capture key properties (e.g. research impact) that are in proximity to the notion of reputation. The correlations among these indices then signify the validity of H-Index as reputation.

3.4 Internal Validity

Internal validity concerns the potential selection effects, confounding variables, and robustness of the methods. It also considers the number of tested hypothesis and distributional or parametric assumptions. In our analysis for Hypothesis 1, it is possible that author's reputation and future citations/H-Index growth are inherently high or low in certain subfields of computer science, causing incorrect interpretation on the result. As a result, we tried to eliminate such effect of population bias with two approaches. Both involve defining a set of subfields, which we build from the "field of study level" information directly available in the MAG dataset. We have picked out 34 subfields labelled as Level 1 for our study. The first approach is then to condition the dataset on one subfield at a time, and check that the result of Hypothesis 1 were reproduced. We achieve this by associating each author with his/her fields of study, determined by whether he/she has published a paper in each subfield (at time t_s). Then we re-perform the analysis of Hypothesis 1 with only authors in a specific subfield. Note that directly filtering papers to each field and re-perform the analysis may be better than the method of assigning authors subfields that we use here. We choose to do so in the interest of time, and under the assumption that each author only works in few subfield in his/her career. In addition, the 34 Level-1 fields are not disjoint in terms of the papers each of them contains.

The second approach is to normalize authors' reputation with respect to their subfields so that future citations/H-Index growth are comparable between authors working in different fields. To represent the normalized reputation, we re-define a normalized H-Index $\hat{h}$ by computing the average H-Index h_f for each field (over authors) in advance, then $\hat{h} = h / \left(\frac{1}{|F|} \sum_{f \in F} h_f \right)$, where h is author's H-Index (at time t_s) and F is the set of subfields associated with the author as described in previous approach. Again, there are better normalizing scheme compared to the variance-introducing steps involved here, but a quick-and-dirty one is chosen in the interest of time.



OLS Regression Results

Dep. Variable:	AvgFutureCitationGrowth	R-squared:	0.424
Model:	OLS	Adj. R-squared:	0.424
Method:	Least Squares	F-statistic:	1.431e+05
Date:	Mon, 07 Dec 2020	Prob (F-statistic):	0.00
Time:	23:34:45	Log-Likelihood:	-8.2677e+05
No. Observations:	194099	AIC:	1.654e+06
Df Residuals:	194097	BIC:	1.654e+06
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-9.2257	0.057	-162.432	0.000	-9.337	-9.114
PreviousHIndex	2.6102	0.007	378.254	0.000	2.597	2.624

Omnibus:	408164.327	Durbin-Watson:	1.995
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4950376737.154
Skew:	17.833	Prob(JB):	0.00
Kurtosis:	784.558	Cond. No.	12.1

Figure 4: Left) Causal Graph. We have binary treatment previous H-Index and outcome citation growth. The confounding variables include career length, conference accepted paper count, field of study, and institution rank. Right) The ordinary least square regression statistics result.

Finally, the previous two approaches is carried out as if subfield of authors is the only confounder to future citations/H-Index. For a more comprehensive validation, we applied causal inference [15] technique to disentangle confounding variables such as field of study, institution, career length, and author's peer review performance for H1. Shown on Fig. 4, our graph has a treatment T of previous H-Index at half-citation year and outcome Y of future citation growth. The confounding variables X are career length, field of study, conference paper count, and institution rank. Recall again that the half-citation year is the year where the author reaches half of its total citations. Conference paper count is defined as the number of accepted papers to either conference or journals known in the dataset. Institution rank is retrieved from the dataset and reordered for only CS institutions. We study the Average Treatment Effects (ATE) with confounders using propensity score matching.

3.5 External Validity

The target of our research was motivated specifically by the computer science community. Therefore, the set of all computer science papers is the largest data of our interest. Considering the coverage of Microsoft Academic Graph is already huge, we do not try incorporating other academic graph (e.g. AMiner) to demonstrate our results. However, we verify the trend in two different time for deep learning, a Level-2 subfield that has grown rapidly in recent years. We present our results in the following section.

4 Results and Findings

In this section, we conduct a few experiments and analyses on the research hypothesis.

4.1 Hypothesis 1

As seen from left plot of Fig. 2, we find that higher reputation is indeed correlated with higher future citations. Therefore, we have demonstrated the positive correlation claimed in Hypothesis 1. The right plot of Fig. 2 plots the same relationship with future citations replaced by future H-index growth. Similarly, with higher reputation, future growth in H-index seems to increase but correlates less with reputation. It might be due to the fact that properties and the discrete nature of H-index propagate through our definition of H-index growth and lead to less variation in the resulting values. We did not investigate further into the formulation of H-index growth since it is intended to be a side observation. We interpret this result as implying that with higher reputation, future works are likely to boost H-index, but still encounter the bottleneck introduced by increase number of papers, which is observed from the many 0-growth data in Fig. 2.

4.2 Hypothesis 2

We also take a look at the career path of a single author Richard Sutton at Fig. 6. We observe that the T-Split H-Index is high near the end points since most of his major work are counted in those two cases, (for instance "Introduction to reinforcement learning"[17] was published in the middle of

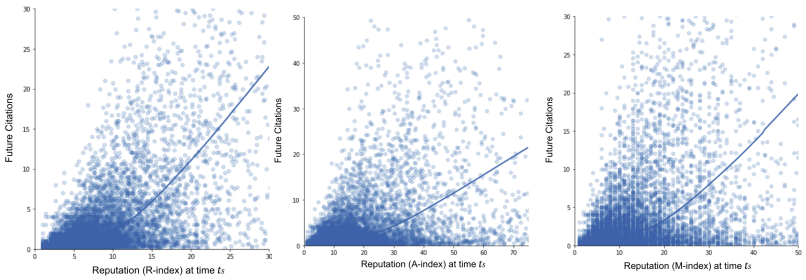


Figure 5: Different Indexes (R/A/M index) to illustrate the construct validity of H1. Left plot is R-Index, middle plot is A-Index, and right plot is M-Index. The x-axes the indexes at time T and the y-axis is the citation growth $\bar{c}_{+}^{t+}$.

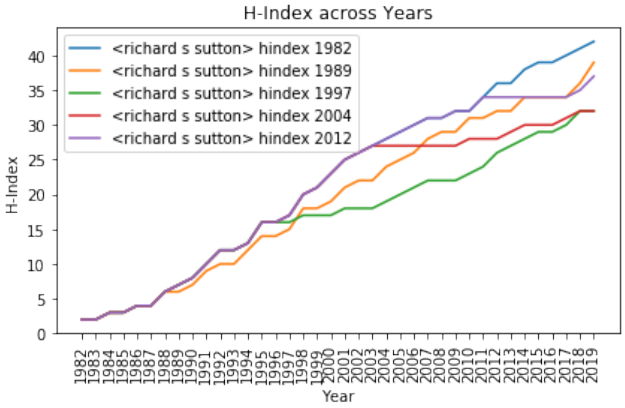


Figure 6: Different color lines represent different T-Split H-Index for the author Richard Sutton.

career, around 1998). We also observe that the slope of lines in his late career grows slowly, likely due to the difficulty of increasing H-Index at that point of career.

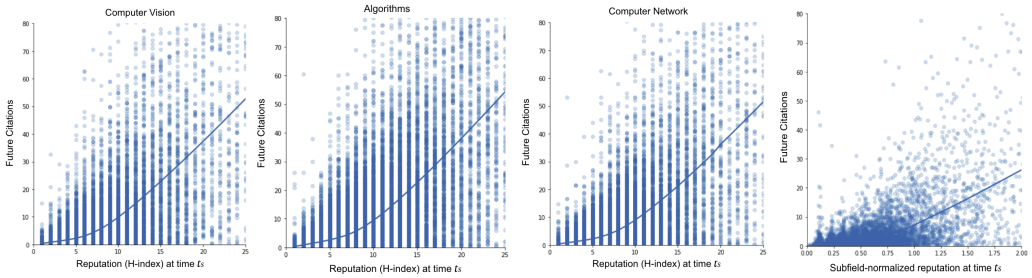
4.3 Construct Validity

We construct the R,A,M indexes to show the convergent validity of our construct. Note that for each of the indexes presented here, all the plots corroborate our findings where there is a positive correlation between A/R/M indexes with future citation growth.

4.4 Internal Validity

Our results of conditioning on the selected 34 Level-1 subfields have reproduced the positive correlations in Hypothesis 1, despite the internal difference among these subfields. Fig. 7a-7c shows the reputation-vs-future-citations relationship in three of the subfields containing similar number of papers. Similar pattern are observed in all plots. For approach with normalized reputation, Fig. 7d suggests that positive correlation still holds if we normalize the effects of subfield when measuring reputation.

We encode the 34 subfields as an multi-hot encoding. We empirically observe that one-hot encoding performs better than other hand-designed features such as the mean citations or H-Indexes for that field, which might come from more degrees of freedoms. We also observe that adding institution rank and field of study overall improves the propensity score prediction performance introduced in the next paragraph. To study the Average Treatment Effects (ATE) with confounders, we need to compute propensity score $p = P(T|X)$. We first need to operationalize treatment to be binary by defining a binary threshold for two groups of high / low reputation authors. For the outcome citation growth, we also experimented with both binary and continuous formulation. Since the distribution of H index is skewed (shown in Fig. 10, so we experiment with both mean and median as the threshold. Using mean would make low-reputation authors easier to classify and the other way around for median,



(a) Computer Vision (b) Algorithms (c) Computer Network (d) Normalized Reputation

Figure 7: We verify internal validity by conditioning and normalizing based on 34 subfields. (a)-(c) are same plot as Fig. 2, with only authors in the subfield of Computer Vision, Algorithms, and Computer Network, respectively. (d) is same plot as Fig. 2 with authors' reputation normalized according to his/her field of study.

Average Treatment Effects (Mean +- Std), 50 runs, 20000 samples				
ATE.	Unadjusted	Nearest Neighbors Matching	IPW Matching	Stratification Matching
Continuous Y	16.528	6.800 +-0.632	46.654 +-32.331	-0.054 +-0.003
Discrete Y	0.502	0.327 +-0.067	1.517 +-1.177	0.531 +-0.008

Table 1: The Average Treatment Effects table. We experiment three different propensity matching methods with confidence interval and observe that there is a citation growth difference between high and low reputation authors, which agrees with our H1 conclusion.

shown in Fig. 8. As expected, the unadjusted ATE, computed as the citation growth sample difference between two groups, is large. We use a few methods in propensity score matching to disentangle the confounding variables. To lower the variance of our fitting, we use bootstrapping of 20000 samples over 50 runs to compute confidence interval for each method. For the nearest neighbor matching, the gap between two groups seems to decrease, which indicates that the confounding variables might have played in role in causing the correlation between T and Y. The inverse propensity weighting (IPW) matching has high variance due to the outliers of high and low propensity score, i.e. there are authors that are very easy to classify. Finally, we adopted 20 uniform layers for the stratified matching, and it appeared that some further tuning would be required for interpretable results. Overall, the causal inference on our observed graph helped validate our H1 conclusion.

4.5 External Validity

Our result (Fig. 9) on a Level-2 subfield (Deep Learning) suggests that positive correlation of reputation and future citation growth holds both before and after the field development. This indicates that the conclusion generalizes to a different set of situation.

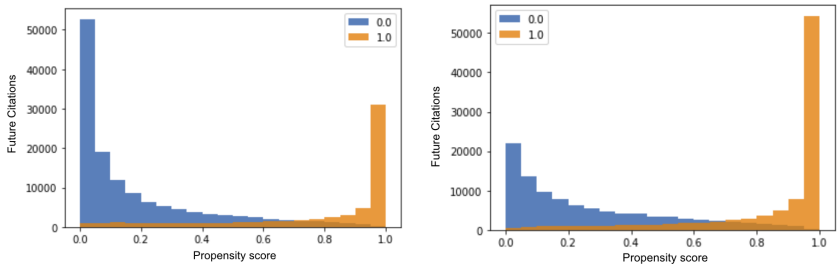


Figure 8: Propensity score fitting result. Left) Mean as the binary threshold for treatment; Right) Median as the binary threshold for treatment. We show that using different threshold to operationalize the binary treatment makes classification results different but we observe similar conclusion.

5 Discussion

5.1 Interpretation

We are interested in the citation pattern of the CS field, which is largely a combination of two factors of the paper: internal contribution and external advertisement. In our H1, we study the correlation between reputation and future citation growth. Based on the findings and validity analysis, we conclude that future citations is correlated with reputation to certain extent. This infers that the field pays attention to renowned authors, which is not necessarily a bad idea since famous researchers often tend to continue publishing good works. Therefore, in our H2, we study that an author's career research potential, which is measured by T-Split H-Index. Our preliminary finding is that an author with lower H-Index might have higher research potential. Therefore, there might be some discrepancy in the attention of the field and the actual work quality. When a researcher cites work, they should consider its paper quality and future research influence in addition to the author reputation. Overall, our methods use three different types of validity analysis to study the causal relationship, and arrive at positive results.

5.2 Limitation

There are several assumptions made in our work that might lead to limitations. Our work is also by no means complete and thorough by relying on partial data from the third source and subjective interpretations. We will discuss these issues below.

Field: Our dataset is filtered based on only Computer Science papers, so cross-field interactions are not considered. Moreover, the conclusions for Computer Science as a recently developing field might limit our extensions to other fields such as Physics and Biomedicine.

Metric: H-index and its variants are imperfect proxy to represent reputations or measure research performance [13]. At the core of the dataset and our analysis is paper citations, which itself is also flawed since paper semantics and authorship [1] are ignored. An extreme case is that the paper mentioned in [13] with 2896 "authors" affiliated to 228 institutions, had received 1631 citations within a year. All of the institutions received full credit for this which negatively affects the validity of citations. Similarly, while we adopted field-based normalization, source-based normalization technique [12] can also be important by taking into account the quality of citing sources.

Causal Analysis: While we addressed four observed variables, we still have many unobserved factors such as the author's actual research ability, paper quality, race, gender, etc.

Semantics: The semantic information for paper text, or field confidences, and institution rankings are ignored in our work. These information are critical for more advanced analysis as well as bias reduction.

H2: Our experimental H2 formulation is an interesting direction but our data and analysis are limited. This limitation is also related to the difficulty of measuring author research performance and paper quality, and a smaller scope of topic might help.

5.3 Future Work

The immediate future work would be to work on the limitations in the previous sections. The systematic bias inside the AI technology and ML field is also a general future direction that this work can go into, as in this very insightful NeurIPS workshop this year [6].

Field: We can consider extend the analysis to other neighbor field such as Electrical Engineering.

Metric: We can consider more subjective metrics to compensate the observed quantity. For instance, open peer-review conference has public scoring information [18] and many websites (including MAG dataset) also contains ranking for institutions and fields. Another way to process field of study data is to normalize across certain tie windows and considering confidence when counting fields.

Causal Analysis: In addition to consider the unobserved variables and build a more complete graph, we can also consider other parametrization and models for feature fitting. Our graph also admits more interpretation, for instance, author's H-Index can be an observation instead of treatment.

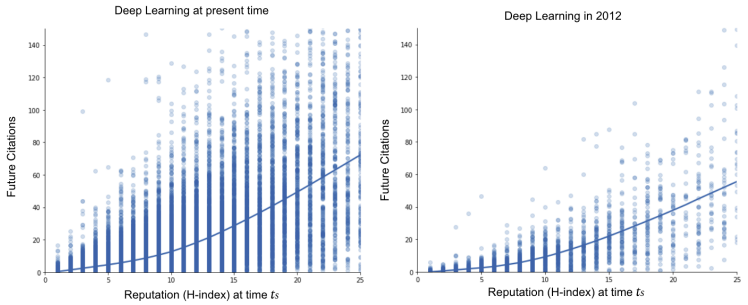


Figure 9: Left) Hypothesis 1 Plot for the Deep Learning subfield. Right) Hypothesis 1 Plot for the Deep Learning subfield before 2012. We observe much a difference in field development but the external validity also holds across time. The x-axes the indexes at time T and the y-axis is the citation growth $\vec{c}_{+}^{t+}$.

Semantics: Considering semantic information such as paper quality can require NLP techniques that are typically used in these large-scale academic search engine.

H2: We can adopt similar techniques used in H1 for H2's construct, internal, and external validity. In future work, we can also think of other ways to measure the same phenomenon with actual paper quality and research ability.

6 Related Works

Our research questions belong to the subfield of scientometric, which is defined at 1971 by [14]. It is defined as developing "the quantitative methods of the research on the developmentof science as an informational process". Scientometrics is closely related to Bibliometrics, infometrics, altmerics, etc. In a nutshell, it is a study that concerns with the analysis of citations in the academic literature. In the review paper [13], it considers the historical development of scientometrics, sources of citation data, citation metrics and the "laws" of scientometrics, normalisation, journal impact factors and other journal metrics, visualising and mapping science, evaluation and policy, and future developments. All of these are relevant in our topic, especially the discussion over citation, H-Index, normalization, and modeling of author career. Our topic focuses narrowly on Computer Science, and is motivated by an observation and hypothesis that reputations is correlated to paper citations. We therefore use different scientometrics to validate our hypothesis.

The Microsoft Academic Graph (MAG) that we used have been analyzed extensively in previous work [16; 7; 10; 8]. Although we work extensively on the dataset and often check with the published statistics on the dataset, We do not focus on understanding such a giant graph. Instead, we focus on studying and analyzing the impact of reputation on research performance, by leveraging the dataset information.

H-Index [9] and its variants [2; 11; 5] are proposed in previous works. They are often presented in the format of analyzing the representation power of these indexes for authors' producitivity, reputation, and actual research performance. Our work use these indexes as constructs to measure reputations and focus on their convergent validity on the CS field.

7 Conclusion

In conclusion, we study the research question on the correlations of future citation growth with previous reputation and relative research potentials, inspired by the observation on recent citation in CS field. Our hypothesis and result suggests that there is a positive correlation between reputation and future citation growth. We then study the construct validity, internal validity, and external validity of our method and further verify the conclusion. Answering this question has some important implications including the bias over author reputation in CS citations. However, our work has many limitations such as the studied field and metrics and future directions such as causal inference and the measurement of authors' research potentials.

References

[1] Nasir Ahmad Aziz and Maarten Pieter Rozing. Profit (p)-index: the degree to which authors profit from co-authors. *PLoS One*, 8(4):e59814, 2013.

[2] Lutz Bornmann, Rüdiger Mutz, and Hans-Dieter Daniel. Are there better indices for evaluation purposes than the h index? a comparison of nine different variants of the h index using data from biomedicine. *Journal of the American Society for Information Science and technology*, 59 (5):830–837, 2008.

[3] Marshall Copeland, Julian Soh, Anthony Puca, Mike Manning, and David Gollob. Microsoft azure and cloud computing. In *Microsoft Azure*, pages 3–26. Springer, 2015.

[4] Rodrigo Costas and María Bordons. The h-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level. *Journal of informetrics*, 1(3):193–203, 2007.

[5] Leo Egghe. Theory and practise of the g-index. *Scientometrics*, 69(1):131–152, 2006.

[6] Charles Isbell et. al at NeruIPS 2021 Workshop. You can't escape hyperparameters and latent variables: Machine learning as a software engineering enterprise, 2020. `https://nips.cc/virtual/2020/public/invited_16166.html`.

[7] Michael Färber. The microsoft academic knowledge graph: A linked data source with 8 billion triples of scholarly data. In *International Semantic Web Conference*, pages 113–129. Springer, 2019.

[8] Drahomira Herrmannova and Petr Knoth. An analysis of the microsoft academic graph. *D-lib Magazine*, 22(9/10), 2016.

[9] Jorge E Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences*, 102(46):16569–16572, 2005.

[10] Sven E Hug, Michael Ochsner, and Martin P Brändle. Citation analysis with microsoft academic. *Scientometrics*, 111(1):371–378, 2017.

[11] Bihui Jin, LiMing Liang, Ronald Rousseau, and Leo Egghe. The r-and ar-indices: Complementing the h-index. *Chinese science bulletin*, 52(6):855–863, 2007.

[12] Loet Leydesdorff and Lutz Bornmann. How fractional counting of citations affects the impact factor: Normalization in terms of differences in citation potentials among fields of science. *Journal of the American Society for Information Science and Technology*, 62(2):217–229, 2011.

[13] John Mingers and Loet Leydesdorff. A review of theory and practice in scientometrics. *European journal of operational research*, 246(1):1–19, 2015.

[14] Vasilliĭ Nalimov. Measurement of science. study of the development of science as an information process. Technical report.

[15] Paul R Rosenbaum et al. *Design of observational studies*, volume 10. Springer, 2010.

[16] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Hsu, and Kuansan Wang. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web*, pages 243–246, 2015.

[17] Richard S Sutton, Andrew G Barto, et al. *Introduction to reinforcement learning*, volume 135. MIT press Cambridge, 1998.

[18] David Tran, Alex Valtchanov, Keshav Ganapathy, Raymond Feng, Eric Slud, Micah Goldblum, and Tom Goldstein. An open review of openreview: A critical analysis of the machine learning conference review process. *arXiv preprint arXiv:2010.05137*, 2020.

8 Appendix

8.1 Data Schema

1. PaperReferenceYears
   This table contains information about citations between paper, detailing the cited and the citing paper with the year they are published.
2. AuthorSplitYear
   This table contains information about the Split Year of each Author where Split Year follows the definiton defined on this paper.
3. PaperAuthorAffiliationYearCitations
   This schema contains the paper and affiliation of each author per year with the citation count of that paper for that year.
4. PaperAuthorYearCitations
   Filtered version of PaperAuthorAffiliationYearCitations.
5. PaperYearCitations
   Contains information about the Paper, the year it is published and the citation is has for that year.
6. PaperNormalizedYearCitations
   Contains the citation count for a number of years after the original publication. This is useful to track the citation growth across time.
7. AuthorYearCitations
   Contains information about the author, the year, and the citation for that year.
8. AuthorPaperCitationsAtSplit
   Contains information about the Author's citation count for each paper at the Author's split year, where split year is defined in the paper.
9. AuthorHIndexFutureCitations
   Contains inforamtion about Author's HIndex at split year as well as the Average citation the author get per year after split year T.
10. AuthorHIndexFuturePaperCitations
    Contains information about Author's HIndex as well as the Average Future Citation per year after split year T.
11. AuthorHCoreCitations
    Contains informations about the paper published by author that is in his/her H-core at split year as well as the paper that will contribute to the author's H-core after time split.
12. AuthorHcorePaperInfoSplit
    This table contains information about the various information of each Paper's statistics for each particular year after the paper is published. For example, it contains the citation count for the paper on and before the year, whether it contributed to the H-index of the author on and before the year T, the future citation count of this paper after the year T, the year difference between T and the publishing year, as well as the year difference T from the current year 2020.
13. AuthorHindexCitationInfo
    This table contains information about each Atuhor's statistics. Namely, for each year T of the author's career, it contains the Previous H-Index on year T, the previous total citation on year T, the H-Index on year T onwards(Not including paper before T), the citation count accumulated by the author after year T(not including any paper published before year T).
14. AuthorFieldNormalizedHindexCitationInfo
    Similar to AuthorHindexCitationInfo but all the citations as well as H-index are normalized according to the averages of the field.
15. AuthorPaperSubField
    All information about the author, the paper they publish and the fields they published in.
16. FieldStatistics
    This table contains the overall field statistical information for each particular year. For

example, it contains the total H-index of all author up until the year T, the average citation of author on year T, the number of paper published up until year T, the number of author in that field for year T.

17. FieldHindexCitationInfo
    This contains other citation and H-Index related information per field.
18. AuthorAcceptedCareerYearInfo
    This table contains information regarding the author's career year, such that for each particular year T, it contains information such as the current career year of the author, the number of accepted paper, the total number of paper published before T.
19. AuthorPaperConferenceInfo
    This schema contains information about the paper and its associated conference and journals where it is published. For each conference or journal, it contains information such as the name of the conference/journal and the rank of the conference and journals.
20. AuthorAffiliationOverYear
    This table contains information such as the author's affiliation over years.
21. AuthorAindexCitationInfo
    This is similar to AuthorHIndexCitationInfo, except we are measure in terms of A-Index instead.
22. AuthorRindexCitationInfo
    This is similar to AuthorHIndexCitationInfo, except we are measure in terms of R-Index instead.
23. AuthorMindexCitationInfo
    This is similar to AuthorHIndexCitationInfo, except we are measure in terms of M-Index instead.
24. AuthorGindexCitationInfo
    This is similar to AuthorHIndexCitationInfo, except we are measure in terms of G-Index instead.

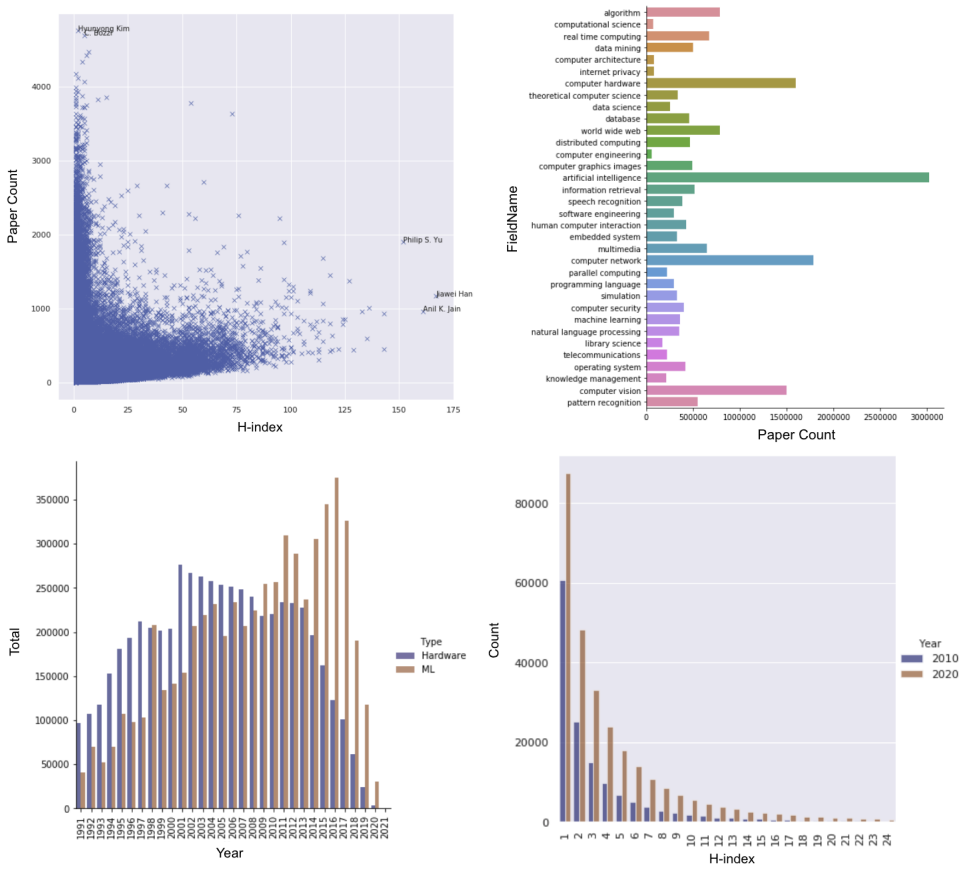


Figure 10: Top Left). Paper Count vs H-Index for authors. Right). Top Right). Paper Count vs SubField. Bottom Left) Machine learning Subfield vs Hardware field paper count change in recent 20 years. Bottom Right) Machine Learning subfield author H-Index change.

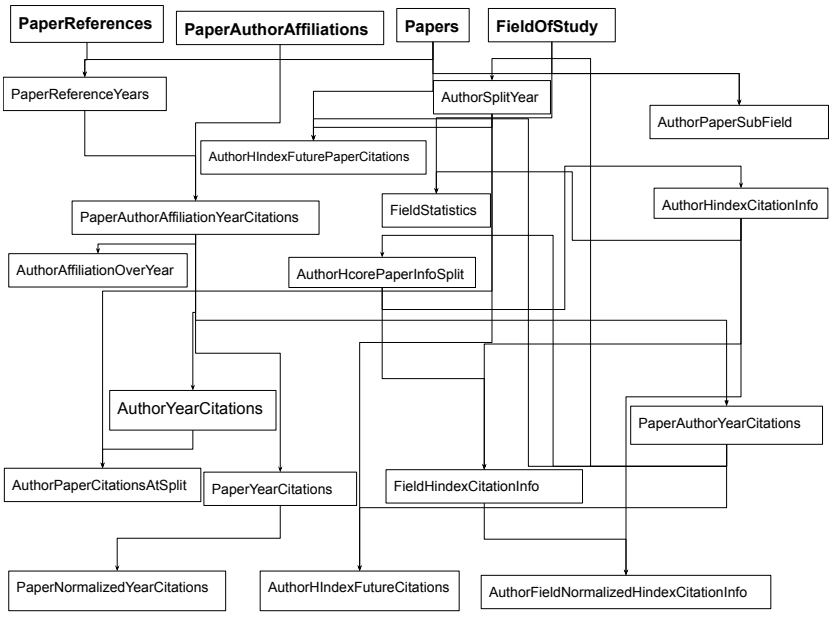


Figure 11: Figure shows the dependency graph of key tables. Bold text represents the base table that we start working from.