H-index Sequences across Fields: A Comparative Analysis

Marlies Olensky
Institute of Information Science
Academia Sinica
marlies.olensky@iis.sinica.edu.tw

Tsung-Han Tsai
Institute of Information Science
Academia Sinica
zark912@iis.sinica.edu.tw

Kuan-Ta Chen
Institute of Information Science
Academia Sinica
swc@iis.sinica.edu.tw

ABSTRACT

This study presents the first analysis of h-index sequences on a larger scale. Exemplarily, we investigated researchers from three different fields within Computer Science. We use Google Scholar citation profiles as data source to construct the h-index sequences of individual researchers. Our ultimate goal is to develop a self-evaluation tool, to assess one's own development of the h-index in comparison to other researchers in the same field, maybe identify career role models in the field and assess career development with future chances of success. The results of this study show that the average h-index sequences behave differently for the datasets, which is partly due to the different sample sizes. Hence, further research will be needed to confirm if every research field behaves differently. In addition, we applied the algorithm developed by Wu et al. [22] to our data to classify the h-index sequences of individual authors according to five different shape categories. The majority of researchers has an S-shaped h-index sequence, followed by IS-shaped and linear sequences. Purely concave or convex sequences hardly ever occur. The researchers with the highest h-indices after 10 career years respectively belong to the S-shaped and IS-shaped categories with a few linear category occurrences. Hence, having a linear h-index is not only very hard to achieve, it is also not a guaranty to be the researcher with the highest h-index in a field.

Keywords

Research evaluation; h-index sequences; Google Scholar; large scale analysis

1. INTRODUCTION

Research assessment, and individual research assessment in particular, has been a research topic discussed for a long time. Indicators have been developed that should convey an impression of a researcher's impact at a glance but when it comes to career advancement, promotion and assignment of research funds, a single indicator might be misleading and not contain the whole picture. Different research evaluation initiatives rely on different approaches and not all of them rely on scientometric or bibliometric indicators to quantify research output. However, particularly for individual research assessment, one of these indicators has become a prominent indicator of scientific impact, because it provides a balance between productiveness and citation impact: the h-index. A lot of research has been conducted on the h-index since its introduction in 2005 [13] and derivative indicators have been developed and tested [5, 9]. Only a few studies [16, 22, 23] have investigated the development of the h-index over time, i.e. h-index sequences, and only one study [17] on a larger scale. However, we believe that h-index sequences can be an informative and adequate source for individual assessment since they are normalized by the career years of the individual researcher.

The contributions of this study are fourfold:

1. We performed the analysis on a larger scale to classify researchers' h-index sequences.

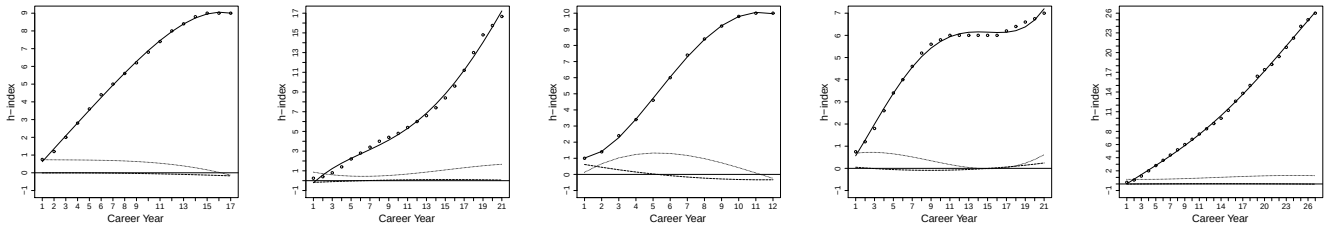
2. The h-index sequences can be different across fields, but we found that the distribution of researcher types might be similar, which needs to be investigated in more detail in future work.

3. Counterintuitively, a linear trend in an h-index sequence does not guarantee a more successful career, in terms of a higher h-index. The researchers with the highest h-indices mostly have S-shaped sequences or IS-shaped sequences.

4. We found that the classification according to Wu's algorithm is not really balanced, since there is a large tendency toward S-shaped sequences.

Our ultimate goal is to develop a self-evaluation tool, to assess one's own development of the h-index in comparison to other researchers in the same field, maybe identify career role models in the field and assess career development with future chances of success. This study, therefore, explores the feasibility of using the h-index sequences for this purpose, by conducting the first analysis of h-index sequences on a larger scale. The paper is organized as follows: Section 2 presents previous studies on h-index sequences, which influenced the research design; Section 3 discusses the methodological approach; Section 4 presents and discusses the results of the study and Section 5 concludes the paper.

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2890520.



(a) convex (b) concave (c) S-shaped (d) IS-shaped (e) linear

Figure 1: 5 types of h-index sequences, as classified by Wu et al. [22] illustration based on our data

2. RELATED WORK

Studies have shown why the h-index gives an adequate picture of one's research output and impact, but also why we do not see the full picture when looking solely at the number of publications that have received at least the same number of citations [1, 8, 12]. The h-index has also been researched over time to determine whether it is a good means to predict one's scientific future [6, 7, 10, 22, 23]. However, the problem has been so far that large-scale data collection is a cumbersome process, because citations received in the past need to be cut off for each additional citation year in order to determine the correct h-index. Hence, we only know about a handful famous scientists, usually Nobel Prize winners or special award winners that have outstanding citation records, and what their h-index sequences look like over time. Wu et al. [22] have identified five different h-index development curves that may describe the career of a scientist. Their study has been carried out on 47 Nobel Prize winners in Medicine (16), Chemistry (14) and Economics (17) and showed that all of them have one of the five types of h-index sequence curves: convex, concave, S-shaped (=first concave, then convex), IS-shaped (=first convex, then concave) and linear (Fig. 1). However, what do we know about *average* researchers? What kind of career curve do they typically have? How do we know at a certain point in our career if there is still a chance to become a top tier scientist or how big that chance is?

The only analysis on h-index sequences on a larger scale has been carried out by Liu et al. [17]. They developed a tool to crawl the citation profiles of computer scientists in Microsoft Academic Search[1], but only used a fraction of these citation profiles to carry out three experiments. They focused on researchers with a high h-index, compared their h-sequences and studied rising stars. Additionally, they selected 50 scientists and performed latent semantic analysis to identify trends in research interests. However, it seems they did not tap the full potential of their collected data, because a quick search for authors in Computer Science on Microsoft Academic Search provides 1,598,575 results (as of 26 November 2015).

With regard to the data sources used for bibliometric analysis, usually one or a combination of the big three databases (Web of Science[2], Scopus[3] and Google Scholar[4]) is employed in their function as citation index [4, 8]. Many studies have investigated Google Scholar as a source for citation analysis and also compared it to the commercial data sources Web of Science and Scopus [2, 3, 11, 18]. Coverage, overlap and citation counts have been compared for various subject domains. The results showed that Google Scholar is a clear winner with respect to coverage of publications, even in the social sciences and humanities where both, Web of Science and Scopus, have some deficits. The studies also concluded that in order to use the databases for research evaluation purposes they should be used in a complementary way and apply data cleaning methodologies. Apart from Google Scholar having the most problems with duplicate publications (and, therefore inflated citation counts [14]), also Web of Science and Scopus struggle with data quality issues, such as incorrect extracted citation information that lead to non- or incorrect matches between cited and citing article [19]. Hence, there is no ready-to-use bibliometric data source for citation analysis. It therefore depends on the goal of the particular study to decide which data source fits the purpose best.

3. METHODOLOGY

Our study presents the first analysis on a larger scale of h-index sequences as a first step towards a self-evaluation tool that will provide time-dependent information and a comparison of the h-index of individual researchers. The research questions we have identified in this context are the following:

- What does the average h-index sequence of a group of researchers from the same field regarding shape and variance look like?
- Can we find differences in the average h-index sequences between researchers from different fields?
- Applying the classification algorithm developed by Wu et al. [22], what category is found most among researchers? What h-index sequence do the researchers with the highest h-index have?

The first and the second research questions deal with the visualization of the average h-index sequences over time. It is rather easy to calculate the current h-index of a researcher, if you have all publication and citation data available. It is mathematically defined as the intersection of the 45 ° line with the curve of number of citations versus paper number, where papers are numbered in order of decreasing citations [13]. However, to calculate the h-index as a sequence in

[1] http://academic.research.microsoft.com/
[2] http://webofknowledge.com/
[3] http://www.scopus.com/
[4] http://scholar.google.com/

Table 1: Descriptive statistics per field

	Computer Vision	Algorithms	Machine Learning
# citations	242,838	90,726	276.985
# authors	234	85	131
h-index			
mean	9.10	8.94	12.11
sd	10.36	8.23	13.72
median	6.00	7.00	8.00
min	1.00	1.00	1.00
max	84.00	49.00	81.00
career year			
mean	11.55	12.00	11.82
sd	7.62	7.04	8.22
median	9.50	11.00	9.00
min	1.00	1.00	1.00
max	41.00	33.00	39.00

time, the citation data needs to be cumulatively concatenated for each additional year, starting in the year of the author's first publication (= career year 1). Because of this normalization of the data, a comparison between different researchers from the same field but also between different fields (given the fact that they are similar in their publication behavior) is valid. We investigated three exemplary fields of Computer Science: *Computer Vision*, *Algorithms* and *Machine Learning*. However, the methodology could be applied to any field of interest. The third research question applies the classification algorithm developed by Wu et al. [22] to the h-index sequences of authors from both fields. We are interested in the distribution of researchers per category and investigate if we can find differences between the fields.

In order to conduct such a large-scale analysis, we needed to collect the complete citation data of all researchers in question. The citation data consists of cited and citing articles and their bibliographic data, including most importantly the publication years of the citing articles. A cited article is one that has been referenced by one or more articles and is also sometimes referred to as target article. In our study target articles are the publications listed on the citation profiles of the researchers. An article citing another article is called a citing article or source article. It holds a reference to usually more than one target article.

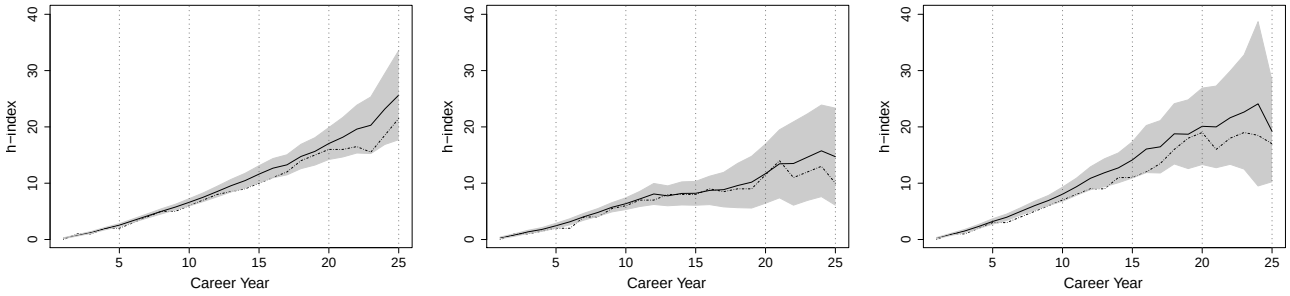
3.1 Choice of data source

In contrast to the study of Liu et al. [17], we opted for Google Scholar as a data source. Ortega et al. [20] have evaluated Google Scholar and Microsoft Academic Search citation profiles and conclude they both could be used for evaluation purposes if applied alongside other data sources. However, when comparing their coverage and quality they found that Google Scholar provides the better results: Google Scholar citation profiles include more documents and citations than those in Microsoft Academic Search with a strong bias toward the information and computing sciences. This bias works actually as an advantage for us, since we are interested in computer scientists. In addition, Microsoft Academic Search shows a higher number of duplicated profiles and a lower updating rate than Google Scholar citation profiles. Because of its comprehensive coverage Google Scholar is an adequate data source for large scale analysis, especially when not only focusing on the elite of a research field. Since we are interested in evaluating individual researchers, disambiguating authors was the most important aspect and this can be rather easily achieved by looking at the individual Google Scholar citation profiles. We can assume that because researchers have to sign up with their academic e-mail address in order to create a profile that they will also invest a certain amount of time to make sure the publications listed on their profiles are close to complete and that they also at least superficially check for duplicate publications or publications by homonymic authors. Since Google Scholar citation profiles and the links to the bibliographic data of cited and citing articles are publicly available and do not require the access to a commercial database, such as Web of Science and Scopus, they are often the first impression of scientific impact, that, for example, hiring committees look at. However, we are aware of Google Scholar's data quality problems that have been investigated and discussed in several studies [14, 15, 11] and these issues are far from being solved [19]. Nevertheless, we cannot ignore the popularity of Google Scholar as a scholarly search engine and its use as easily accessible reference point when checking a researcher's citation profile. Therefore, we decided to still base our study on Google Scholar data and applied data cleaning and harmonization methods (cf. subsection 3.2).

3.2 Data collection

The reason why no large-scale analysis of h-index sequences has been conducted so far, has been addressed in previous studies as "because it is a cumbersome process". Nevertheless, we opted for Google Scholar for the reasons described in the previous section. We developed a tool to crawl Google Scholar citation profiles of computer scientists, who had a public profile on Google Scholar and stated *Computer Vision* or *Algorithms* as one of their areas of interest or contained these strings in their affiliation. We collected the author-specific data (name, affiliation and areas of interest) and the bibliographic data we needed to calculate the h-index sequences. Hence, we not only collected the publication data and the citation counts, but also collected the data of the citing articles in order to be able to calculate the cumulative citation counts of the cited articles for each citation year. Even though Google Scholar is a public service which offers bibliographical information to all the users, it poses some technical challenges when we crawl researcher profiles. The main obstacle is that it prevents automated crawling by presenting a CAPTCHA test [21] so that users need to pass a "Are you human?" test in order to continue using the service. Such CAPTCHA tests will present themselves randomly, partly correlated with the rate and number of pages crawled. Therefore, we resort to use a dozen of PCs to crawl data from Google Scholar simultaneously with a rate acceptable by Google yet providing a good throughput rate in order to collect a large number of researcher profiles with minimum time.

In total we investigated 234 researchers in the field *Computer Vision*, 87 in the field *Algorithms* and 133 in the field *Machine Learning*. A total number of 610,549 citing articles was processed (cf. Table 1 for the detailed descriptive statistics). We found 7 duplicated citation profiles (even with verified e-mail addresses), which we of course excluded



(a) Computer Vision (b) Algorithms (c) Machine Learning

Figure 2: Average h-index sequences. Solid line = mean, dashed line = median, grey area = 0.95 confidence interval

from the analysis. We eliminated articles without publication and/or citation year from the analysis and also checked for plausible publication and citation years (e.g. the publication year of the target article must be the same or smaller than the publication year of the source article). We had to completely remove one profile from the analysis, since it was a profile where the researcher confirmed every publication by homonymic authors (yet with different second initials) as his own. We also eliminated duplicate source and target article pairs, if existing, from the analysis. However, we did this with exact string matching only, which means that different versions of the same article with article title variants have not been de-duplicated.

4. RESULTS AND DISCUSSION

4.1 Average h-index sequences

Answering the research question about the average h-index sequence, we can observe similarities but also differences between our data sets. Fig. 2 shows the average h-index sequences (solid line) complemented by the 0.95 confidence interval area (grey area). In addition, the illustrations show the median (dashed line). Only around 10% of the researchers have a career of 25 years that is why we decided to cut off the comparison at 25 career years. In general, the average h-index sequence increases over time. However, this increase can be observed in the *Computer Vision* and *Machine Learning* datasets almost linearly for the entire career year line and in the other dataset only until career year 12 and with a smaller yearly increase. Interestingly, all three datasets show approx. the same mean and median h-index at 5 and 10 career years, which can be a sign that in the first ten career years the h-indices of researchers within a certain research field behave the same. Note, that only 50% of the researchers have a career with more than 10 years in all three datasets. Hence, the accuracy of the calculation decreases with increased career year and the findings need to be corroborated in additional future analyses. The median is predominantly higher than the mean in the two bigger datasets and they drift further apart after approx. career year 12. This could be an indicator that at approximately 12 career years, there is a turning point: either, a career takes off, which is expressed in an increased h-index, or the career stagnates and shows only small increases in the h-index in the following years. Therefore, the median h-index is a better means to describe the "average" researcher in both datasets.

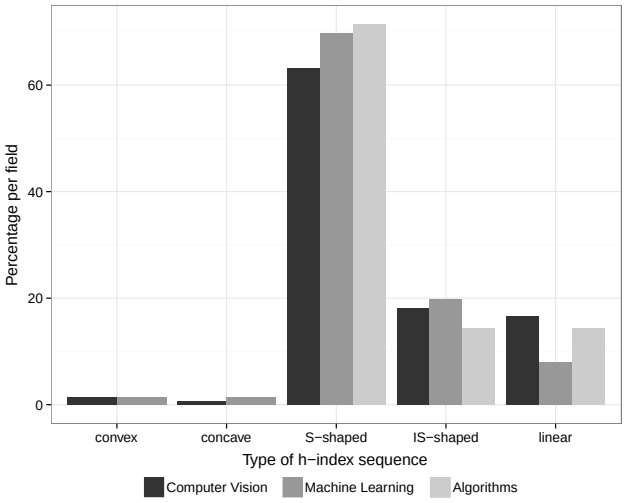
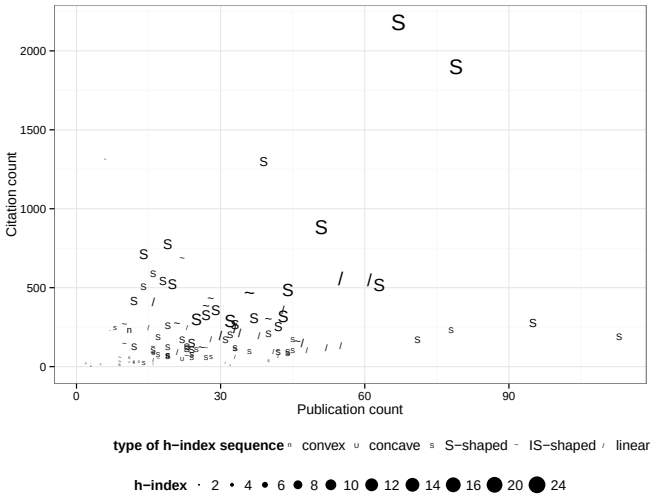


Figure 3: Comparison of h-index sequence categories by field

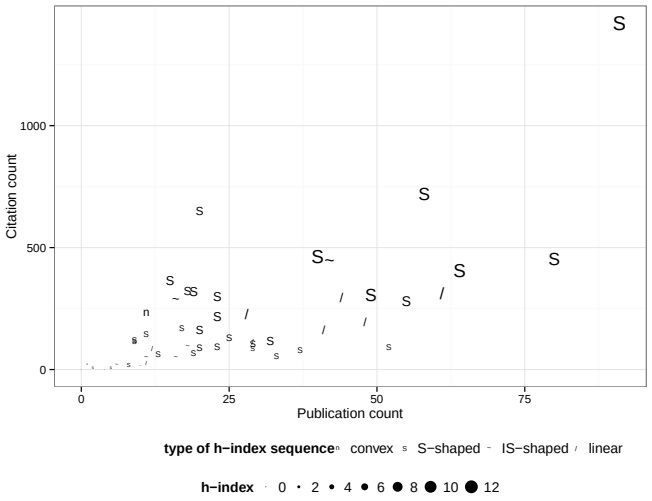
However, we also need to take into account the different sample sizes. The *Machine Learning* was only about half the size of the *Computer Vision* dataset and the *Algorithms* dataset was even smaller, meaning the *Algorithms* dataset is less crowded. The differences caused by the sample size are predominantly expressed by the increased area of the confidence interval. From around career year 15 in *Computer Vision*, and a little earlier in the other two datasets, the area of the confidence interval starts getting larger. This can be explained by the decreasing number of researchers that have a career with more than 15 years. This suggests that each field has its own dynamics after career year 15.

4.2 Classification of h-index sequences

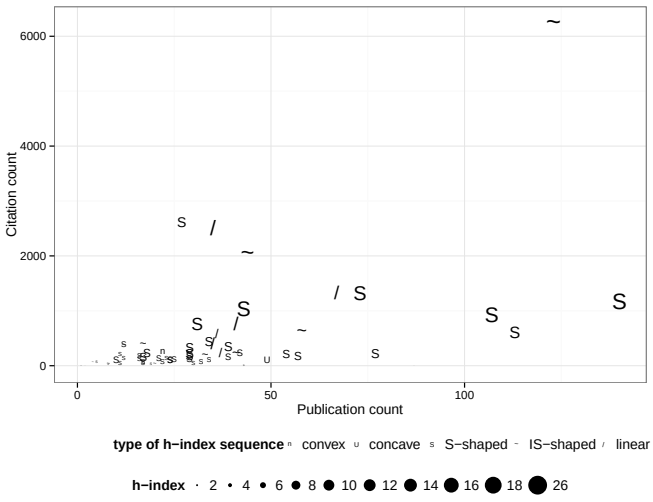
To answer research question 3, we applied the algorithm developed by Wu et al. [22] to classify the h-index sequences into one of the five categories: convex, concave, S-shaped (=first concave, then convex), IS-shaped (=first convex, then concave) and linear (cf. Fig. 1. We found that more than 8 career years are necessary to classify the h-index sequences with the algorithm. Therefore, of the 234 *Com-*



(a) Computer Vision, career year 10



(b) Algorithms, career year 10



(c) Machine learning, career year 10

Figure 4: Publication count, citation count, h-index per author and type of h-index sequence

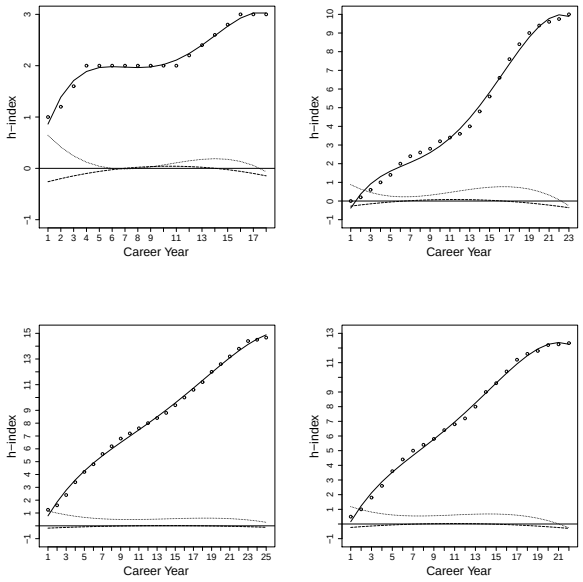


Figure 5: Examples of h-index sequences classified as S-shaped by the Wu algorithm, that are not unambigously S-shaped

puter Vision (CV) researchers 56.84%, of the 131 *Machine Learning* (ML) researchers 57.14% and of the 87 *Algorithms* (AL) researchers 64.37% could be classified. Of those, the majority shows an S-shaped h-index sequence (CV: 63.16%, ML: 69.74%, AL: 71.43%), followed by slightly more IS-shaped sequences (CV: 18.05%, ML: 19.74%, AL: 14.29%) than linear ones (CV: 16.54%, ML: 7.89%, AL: 14.29%) (Fig. 3. Purely convex or concave sequences only occur once or twice in CV and ML. This means the majority of researchers have a rather slow start in their early career years, expressed by a slow increase in the h-index, followed by an increase in the middle of their career and at some point the h-index stagnates again.

Fig. 4 shows a career year comparison of all three datasets after 10 career years. Each researcher is represented by their type of h-index sequence (= shape and colour) and the size of the different shapes represent the size of the h-index at this specific point in their career. On the x-axis we find the total number of publications and on the y-axis the total number of citation received. Looking for the researcher with the highest h-index in all three fields, we can find different h-index sequence categories. Counterintuitively, not all researchers classified as linear h-index sequence have the highest h-indices. In *Computer Vision*, the researchers with an S-shaped h-index sequence have the highest h-index, but there are two researchers with an IS-shaped and a linear sequence that are as high. In contrast, in *Algorithms*, the researcher with the highest h-index belongs to the IS-shaped category. *Machine Learning* also paints a different picture. The researchers with the highest h-index have an IS-shaped, S-shaped or linear h-index sequence. Hence, having a linear h-index is not only very hard to achieve, it is also not a guaranty to have the highest h-index in a field.

5. CONCLUSION

This research investigated the h-index sequences of researchers of three exemplarily chosen research fields within Computer Science. We performed a large-scale analysis to analyze the average h-index sequence as well as classify researchers according to their h-index sequence shape. The average h-index sequence curves can be very different across areas, but were similar for the first 10 career years for all three research fields. However, this finding needs to be corroborated in future research by an even larger data sample. The majority of researchers has an S-shaped h-index sequence, followed by IS-shaped and linear sequences. Purely concave or convex sequences hardly ever occur. Counterintuitively, a linear trend of an h-index does not guarantee a higher h-index. Researchers with a high h-index mostly have S-shaped h-index sequences. Since Wu et al.'s algorithm was developed using Nobel Prize winners' careers, we suspect that this is the reason why the classification of average researchers may have been unbalanced toward one category. The examples in Fig. 5 show some h-index sequences that may not be classified unambiguously as S-shaped. Therefore, in our future work we will work on refining the classification of h-index sequences as well as adding more subfields to our investigation of Computer Science, in order to carry out a true large-scale analysis of h-index sequences.

6. REFERENCES

[1] G. Abramo, C. A. D'Angelo, and F. Viel. A robust benchmark for the h- and g-indexes. *Journal of the American Society for Information Science and Technology*, 61(6):1275–1280, 2010.

[2] L. S. Adriaanse and C. Rensleigh. Web of Science, Scopus and Google Scholar: A content comprehensiveness comparison. *The Electronic Library*, 31(6):727–744, 2013.

[3] É. Archambault, D. Campbell, Y. Gingras, and V. Larivière. Comparing bibliometric statistics obtained from the Web of Science and Scopus. *Journal of the American Society for Information Science and Technology*, 60(7):1320–1326, 2009.

[4] J. Bar-Ilan. Which h-index? A comparison of WoS, Scopus and Google Scholar. *Scientometrics*, 74(2):257–271, 2008.

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

[6] L. Egghe. Comparative study of h-index sequences. *Scientometrics*, 81(2):311–320, 2009.

[7] L. Egghe. Mathematical study of h-index sequences. *Information Processing and Management*, 45(2):288–297, 2009.

[8] M. Franceschet. A comparison of bibliometric indicators for computer science scholars and journals on Web of Science and Google Scholar. *Scientometrics*, 83(1):243–258, 2010.

[9] F. Franceschini, M. Galetto, D. Maisano, and L. Mastrogiacomo. The success-index: An alternative approach to the h-index for evaluating an individual's research output. *Scientometrics*, 92(3):621–641, 2012.

[10] R. Guns and R. Rousseau. Simulating growth of the h-index. *Journal of the Association for Information Science and Technology*, 60(2):410–417, 2009.

[11] A.-W. Harzing. A preliminary test of Google Scholar as a source for citation data: A longitudinal study of Nobel Prize winners. *Scientometrics*, 94(3):1057–1075, 2013.

[12] M. Henzinger, J. Suñol, and I. Weber. The stability of the h-index. *Scientometrics*, 84(2):465–479, 2010.

[13] J. E. Hirsch. 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):16569–16572, 2005.

[14] P. Jacsó. Deflated, inflated and phantom citation counts. *Online Information Review*, 30(3):297–309, 2006.

[15] P. Jacsó. Metadata mega mess in Google Scholar. *Online Information Review*, 34(1):175–191, 2010.

[16] L. Liang. H-index sequence and h-index matrix: Constructions and applications. *Scientometrics*, 69(1):153–159, 2006.

[17] Y. Liu and Y. Yang. Empirical study of L-Sequence: The basic h-index sequence for cumulative publications with consideration of the yearly citation performance. *Journal of Informetrics*, 8(3):478–485, 2014.

[18] L. I. Meho and Y. Rogers. Citation counting, citation ranking, and h-index of human-computer interaction researchers: A comparison of Scopus and Web of Science. *Journal of the American Society for Information Science and Technology*, 59(11):1711–1726, 2008.

[19] M. Olensky. *Data Accuracy in Bibliometric Data Sources and its Impact on Citation Matching*. PhD thesis, Humboldt-Universität zu Berlin (Germany), 2015.

[20] J. L. Ortega and I. F. Aguillo. Microsoft Academic Search and Google Scholar citations: Comparative analysis of author profiles. *Journal of the Association for Information Science and Technology*, 65(6):1149–1156, 2014.

[21] L. Von Ahn, B. Maurer, C. McMillen, D. Abraham, and M. Blum. recaptcha: Human-based character recognition via web security measures. *Science*, 321(5895):1465–1468, 2008.

[22] J. Wu, S. Lozano, and D. Helbing. Empirical study of the growth dynamics in real career h-index sequences. *Journal of Informetrics*, 5(4):489–497, 2011.

[23] F. Y. Ye and R. Rousseau. The power law model and total career h-index sequences. *Journal of Informetrics*, 2(4):288–297, 2008.