Node Analysis for Academic Networks

A[1], **B**[1], **and C**[1]

[1]Shanghai Jiao Tong University

Many structures and systems in real world can be abstracted into graphs, yet random graph models are indispensable for the description of network systems due to various factors of inaccuracy. The discussion in this article is focused on academic networks, a type of complex networks concerning optimization and estimation of the proposed algorithm's performance. On the basis of the results obtained, we have further explored the similarities and differences between major academic relationships represented by citation ratio among corresponding domains/organizations in the entire Internet. The association between these relationships and h-index is also derived to evaluate academic influence.

h-index | academic networks

Academic networks are social networks built upon academic activities along with corresponding relationships. They are generally huge networks composed of papers, authors, conferences and research fields, as well as the relations between them. The citation network for papers and the collaboration network for authors are two common types of academic networks. Relying heavily on citation dependencies, academic influence is one of the most popular performance metrics measuring the scientific status of a paper, an author or a conference among some certain scientific communities. For a paper, the more it is referenced, the more influential it will be. Similar reasoning can be applied to an author or a conference as well.

Provided that node degrees in academic networks represent the number of citations each node has, studying these quantities allows us to further discover the relations between academic nodes and even to unravel the potential patterns hidden underneath. Furthermore, in cases where academic nodes symbolize authors, we are able to analyze entire citations of his/her publications for drawing the conclusion on his/her overall scientific influence.

The above reasoning and observations motivate us to analyze the relations between author's citation statistics and academic influence.

1. Academic evaluation index

Traditional academic evaluation indexes of an author include peer review, total cited number, important paper cited numbers, etc. They all have merits and drawbacks. In 2005, Hirsch, physicist from University of California San Diego, proposed a new evaluation index: **h-index**. Since then, h-index has been widely used in various databases.

If an author has an h-index of h, it means that he/she has at least h papers that have been cited no less than h times. By taking into account both the number of total publications and the corresponding citations, h-index evaluates an author comprehensively. Those who have published a lot yet are rarely cited by others, as well as those who have only few papers with high cited numbers, fail to maintain satisfying h-indexes. In addition, h-index is simple and easy to compute. Last but not the least, each paper can only be cited by another paper for once and once at most, facilitating an equality between citations and papers. We therefore adopt h-index as one of the academic evaluation indexes of author in our paper.

Apart from h-index, we need an additional metric to evaluate the citations. We decide to use the **proportion of cited numbers for a scholar's 5 most popular papers in his/her total cited numbers** because the ratio carries information concerning citations of important papers. Furthermore, the cite ratio always has a value ranging from 0 to 1 and thus weakens the negative impact of extreme cited numbers. Meanwhile, the cite ratio also makes it possible to study other aspects of the academic networks such as the impact of eliminating some significant citations. In an effort to avoid potential negative influence brought by extreme data, we only analyze authors with 10 or more publications.

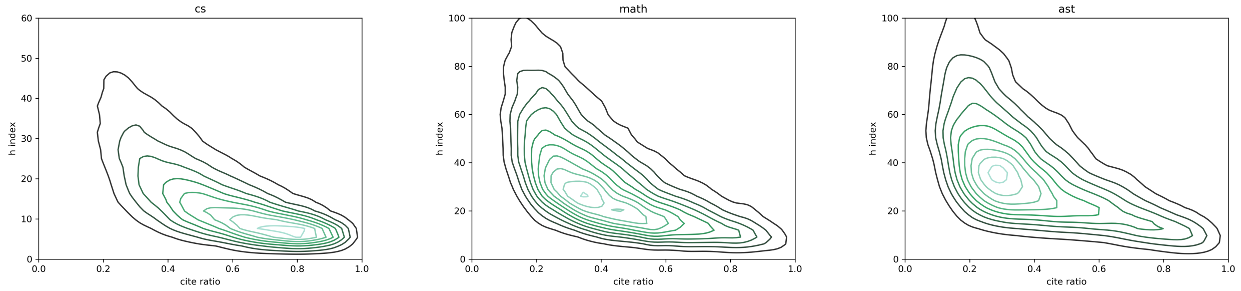
2. Database and web crawling

Before analyzing the relations between different evaluation metrics, we have to acquire the following information of every author in the first place: h-index, total cited number and cited numbers for each paper. Since no existing datasets provide the desired information collectively, we try to get the data from online databases by means of web crawling.

We choose *Google Scholar* as the major data source given the fact that it enjoys high academic prestige and provides detailed citation statistics. Each profile page in Google Scholar collects all the publications of an individual author, which facilitates the computation of h-index and the extraction of citation related information to a great extent. However, Google Scholar doesn't provide any classified databases, which indicates that we have to get supplementary information concerning an author's research fields from other sources.

We extract supplementary information from *DBLP* and *WOS(Web of Science)*. DBLP provides entries of all publications in Computer Science and the corresponding authors since 1993. It has approximately 2 million author entries at present. WOS is another authoritative database which covers 5 different scientific fields: life science, natural science, engineering science, social science and humanities. It collects data from more than 12 thousand journals and 160 thousand conferences since 1950. In addition, WOS supports field related searching, which enables us to obtain information about authors in each research field.

After getting the author information from DBLP and WOS, we are ready to retrieve citation statistics from Google Scholar. Web crawling is an automatic process that extracts useful information from numerous websites without pausing. Web crawling can yield large quantities of information, but meanwhile it increases the load and pressure posed on servers. Too much web crawling can result in DDoS (Distributed Denial of Service). In order to avoid such unpleasantness, we ameliorate



(a) Computer Science (b) Mathematics (c) Astronomy

Fig. 1. author h-index and cite ratio distribution in various fields

our web crawling from the following 3 perspectives:

1. Forge user-agent. We modify the user-agent field in web crawler's requests and ensure that they look exactly the same as those issued by explorers. In this way, we disguise the web crawler as a regular user.

2. Use IP agents. We search and use appropriate IP agents on the Internet to guarantee that a new IP is ready the moment the current IP is blocked by anti-web crawling mechanisms.

3. Send randomized requests at low frequencies. Our web crawler pretends to be a polite user and sends requests to the server every several seconds. For better simulative effects, we use random numbers to generate the time intervals between two consecutive requests.

Once the web pages are successfully retrieved, our web crawler filters and analyzeds the contents contained. Since all target web pages have the same format, all the web crawler needs to do is to extract specific elements. Our web crawler is implemented by Python and the information extraction from web pages is facilitated by a Python library: *pyquery*. Pyquery supports selective web page visits according to element attributes.

3. Data analysis

Our web crawling fails to exhibit excellent performance because not every scholar has a profile page in Google Scholar. We got in the end more than 60 thousand valid entries from mainly three scientific fields: Computer Science, Mathematics and Astronomy. The authors specialized in Computer Science come from DBLP, whereas scholars majoring in Mathematics and Astronomy are from WOS. There are 100 thousand entries for both Mathematics and Astronomy, with 100 thousand the upper limit for WOS entry exportation. We are not aware of the data generation mechanisms for WOS, thus we assume without loss of generality that the derived entries are randomly generated.

A number of aspects concerning the collected data are observed and studied:

1. The relation between h-index and cite ratio of authors.

2. The distribution pattern of authors' cite ratio.

3. Distributions of h-index and cite ratio for all scholars in an institute/organization with different academic status.

4. Lateral comparison of distributions for scholars in different research fields of interest.

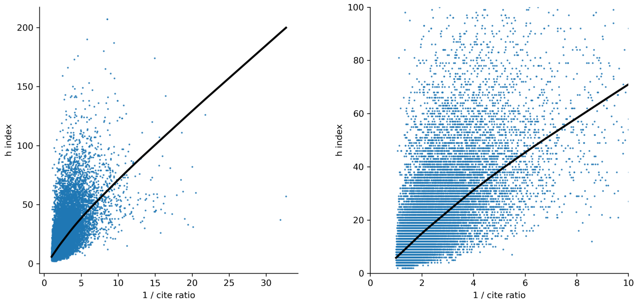
5. Characteristic features for the distributions of scholars with top tier awards and prizes.

A. The relation between h-index and cite ratio.

Three-dimensional distribution graphs are first computed to visualize the relation between h-index and cite ratio. *Kernel Density Estimation* plots of three fields are illustrate in 1.

Despite minor differences, all graphs reveal a negative correlation between author h-index and cite ratio. The correlation coefficients for Computer Science, Mathematics and Astronomy are respectively -0.63, -0.56 and -0.54. The two academic indexes have fairly strong negative correlations. In fact, by the observation from graphs, it looks like they tend to exhibit inverse proportion. We carried out simulations to verify our suspicion.

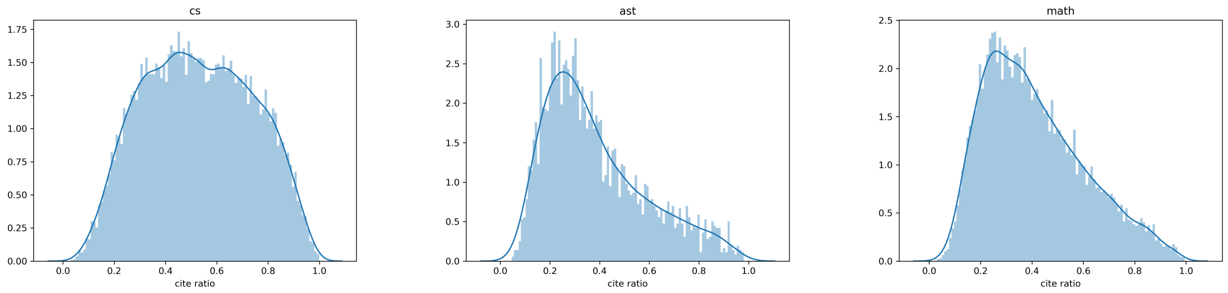
We transform the abscissa of the graphs into the inverse of cite ratio before simulation since most existing simulations are either linear or polynomial. We use the *seaborn* library in Python for local regression fitting. Seaborn performs simulations by local weighting and the corresponding result resembles a smooth line. The simulation result and its partial enlargement are illustrated in 2:



(a) original simulation result (b) partial zoom of simulation result

Fig. 2. relation between author h-index and cite ratio

The simulated result is almost a straight line, which con-



(a) Computer Science (b) Astronomy (c) Mathematics

Fig. 3. author cite ratio distribution in various fields

forms to observations and verifys our suspicion. In the following paragraphs, we demonstrate mathematically **the inverse proportion relation between h-index and cite ratio** under the *simple h-index model*.

According to the assumptions of simple h-index model, each author publishes p papers annually and each published paper will be cited c times annually. Hence, the total cite number and the important paper cited number ratio for an author is:

$$N = \sum_{i=1}^{n} pci = pc\frac{n(n+1)}{2} \quad [1]$$

$$N_5 = \sum_{i=1}^{n} 5c = 5cn \qquad or \quad [2]$$

$$N_5 = 5cn - (5-p)c - (5-2p)c - ...(until 5 > kp) \quad [3]$$

Note that there are at most 5 items in equations 2 and 3, we can therefore neglect the last terms. Assume that an author's h-index is entirely contributed by the papers published before year y, we have:

$$(n-y)c = ph = h \quad [4]$$

The first term is the minimal citation number of a paper which contributes to h-index and the second term is the total number of papers contributing to h-index. By combining the equations above, we have:

$$\frac{N_c}{N} = \frac{10}{p(1+\frac{h(c/p+1}{c})} \sim \frac{10}{h(c+p)} \quad [5]$$

It is clear that **under ideal conditions, the cite ratio is inversely proportional to h-index**. There are inevitable fluctuations of author's annual publication numbers in reality, which explains the dots scattered on both sides of the simulated line.

The authors with high academic achievements usually have low cite ratios because many of their papers have great citation counts. Inversely, authors with little academic influence only publish a limited number of high quality papers, which yields high cite ratios.

B. Cite ratio distribution.

Visualized cite ratio distributions for each field are illustrated in 3:

Apart from visualizations, seaborn also calculated several important parameters for cite ratio distributions as well (see 1):

Table 1. cite ratio distribution parameters in different fields

Field	Mean	Variance	Skewness	Kurtosis
Computer Science	0.5295	0.0441	0.0399	-0.9236
Mathematics	0.4100	0.0381	0.6418	-0.2615
Astronomy	0.3846	0.0414	0.8454	-0.0410

It can be observed both from graphs and tables that the three distributions have similar variances, which suggests a comparable degree of deviation for the three distributions. The cite ratio distribution in Computer Science community is relatively flat and symmetric. The cite ratio distribution in Mathematics is more concentrated, with an obvious peak between 0.2 and 0.4. The distribution in Astronomy has a sharper peak than that for Mathematics and is close to a normal distribution which has a kurtosis of 0.

The differences between Computer Science and other two research fields may be due to the following points:

1. Data source. The Computer Science data are obtained from DBLP whereas the other two are from WOS. The graphs and parameters for Mathematics and Astronomy share the most similarities by pairwise comparison. Thus, we conclude that different data sources may lead to differences in data collection stage.

2. Research field. Since we assume that the exported entries from WOS are randomly generated, differences between databases cannot fully explain the differences in results. Research fields may also account for the differences in results. We have already known that authors with high academic influences have low cite ratios. As they get older, authors generally tend to have a greater academic influence. For fields like Mathematics and Astronomy which have come into existence for a long time, many influential authors have already completed their research. They usually have great ages and consequently cluster on the graph where the cite ratio is low. Furthermore, there were no detailed subject classifications until recent decades. Scientists back then were at a time both Mathematicians and Astronomers. Hence, similar results are

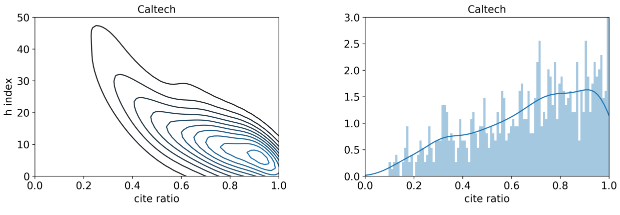
observed in these two fields. Although similar, Mathematics lays the foundation for many other branches of science. As a result, there are constantly new scholars dedicating themselves into the research field of Mathematics. Astronomy, on the contrary, is one of the applications of Mathematics in some way. It doesn't have much overlapping with other scientific fields and consequently there are fewer scholars in this field.

Computer Science is a new research field with only a history of around 100 years. With the prosperity of Artificial Intelligence, Block Chain and many other novel technologies, Computer Science is drawing the attention of countless scholars. The brief history of Computer Science suggests that there are not many outstanding scientists who have completed their great explorations, which explains the relatively flat cite ratio distribution.

3. The components of cite ratio distribution. According to academic citation dynamics and our observations, we have good reason to believe that the cite ratio distribution for authors of about the same age follows a normal distribution. Since the 60 thousand authors obtained by web crawling are from all ages, the resulting cite ratio distribution for each field is most likely to be a combination of several normal distributions. For newly emerged research fields such as Computer Science, the resulting distribution is symmetric. For classic research field like Mathematics and Astronomy, elderly researchers whose birthday dates back to more than 100 years ago are probably deceased or no longer have new publications. They are located near peak area, with cite ratio between 0.2 and 0.4.

C. Comparison between distributions among various organizations.

Apart from analyzing the cite ratio distributions of a variety of fields, we further gather comprehensive data from quantities of universities and institutes. As have been claimed beforehand, we collect statistical of h-index and cite ratio from every scholar with no less than 10 publications in an organization. The range of institutes mainly covers North America, Europe and Asia.



(a) kernel density estimation plot of h-index and cite ratio (b) distribution plot of cite ratio

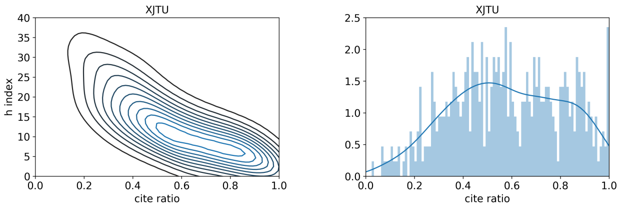
Fig. 4. distributions of h-index and cite ratio for a typical organization

The analysis of a typical institute is visualize in figure 4. The results illustrate a sharp contrast to the distributions in previous analysis. It is obvious that the peak of the cite ratio distribution is heavily right-skewed compared with field-based rivals and the center of kernel density estimation is close to the bottom-right corner. These observed phenomena reveal a rather contradictory conclusion that a large proportion of scholars in this particular institute have low influence. Nonetheless, this difference is prevalent in most of the organizations analyzed. Based on abundant observations and some appropriate inferences, the sharp contrast of domain-based distributions and institute-based distributions can be attributed to the following factors.

1. The coverage of Google Scholar is more comprehensive than DBLP and WOS, recording inferior publications and the corresponding unimportant scholars. Unlike the former databases which only collect data from certain first tier journals and conferences, the later aims to obtain content from arbitrary publications. In this way, numerous scholars win a place and establish their own profiles which in other databases seems possible. Needless to say, this additional crowd of scholars achieve less progress and tend to have higher cite ratios, leading to a large 'tail' in the distribution plot. The above characteristic concerning this distribution is in analogy to a power law distribution of the influence among scholars in an organization, namely only a small fraction become well known for their achievements while the vast majority stay unknown.

2. The diversity of the collection of fields is worth considering as well. An organization includes numerous scholars while each of then specializes in one or more fields. The collective union is thus a complex combination of domains concerning every aspect of life, ranging from natural science to social science. Fields like Mathematics and Chemistry bear a long period of research, leading to professional being limited to the elderly and little space for the late comers, while new fields such as Computer Vision and Artificial Intelligence are still early on the rise, leaving behind abundant issues to be put forward and solved. A single domain may exhibit a left skewed peak to some extent, but the statistics of an organization is based on the entire coverage of fields that its faculty work in. Since it is not a surprise that the quantity of 'new' fields far exceeds that of 'old' fields, the fact that organizational distributions exhibit right skewed peaks seems reasonable.

Apart from the differences between institutes and domains, the distributions of each independent organization vary from individual to individual as well. Figure 5 visualizes the statistical data and corresponding distributions of a less influential university, exhibiting a sharp contrast from the previous institute.



(a) kernel density estimation plot of h-index and cite ratio (b) distribution plot of cite ratio

Fig. 5. distributions of h-index and cite ratio for a less influential organization

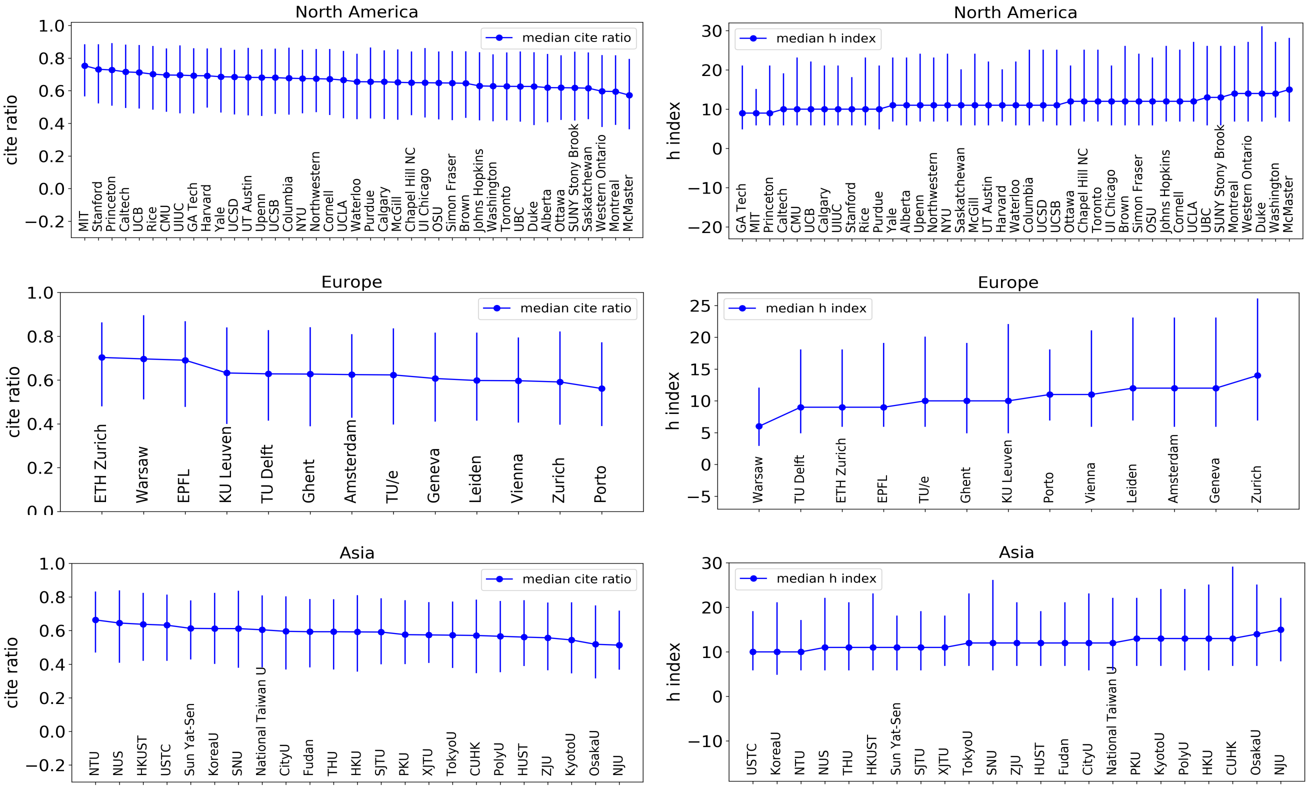


Fig. 6. Boxplots of cite ratio(left) and h index(right) for organizations in North America(6(a)&6(b)), Europe(6(c)&6(d) and Asia(6(e)&6(f)). The distribution of each organization is represented by a line spanning from the first quartile Q_1 to the third quartile Q_3 with the median marked and connected in a line. Distributions of cite ratios are arraged in ascending order concerning the median of cite ratio and that of h index in descending order according to the median.

By observation the conclusion can be drawn that the center for kernel density estimation stands farther from the bottom-right corner and the peak of the distribution plot tends to be more centered than right skewed. In other words, a less influential organization holds a larger proportion of influential scholars. This contradictory inference points out the existence of more 'ordinary' people in a top tier university. Indeed, the unified influence of an institute is not merely determined by the minority of experts, but rather depends on the achievements of the majority. **An organization maintains its high status not because of the achievements of one or a few specific scholars, but because of the progress made by the entire faculty in every realm of scientific research**.

The gap between a more influential institute and a less influential one can as well be explained from a field based perspective. Naturally, it is reasonable to assume that a superior organization holds more scholars than an inferior one. Nevertheless, the well-known are always the minority, occupying a negligible amount of the entire faculty. Relatively, the above pattern suggests the existence of more ordinary scholars in an outstanding university, which recognized the achievement of every individual for making up the status of the organization they belong to. Moreover, influential institutes aim at cutting-edge topics and issues and focus mainly on newly-emerged domains or even discover **new fields** while inferior ones can only follow the footsteps of the former ones. Above all, scholars in a more influential institute tend to concentrate on their own research subjects and discover new issues in new fields. Conversely, a member of the faculty in a less influential organization is more willing to work in an established domain for his recognized position. In this way, the differences of peaks in distributions of various organizations can be further explained.

To verify the assumption that abundance of scholars with high cite ratios determines the influential positions of organizations, the distributions of each institute is further displayed by means of boxplots. From observations of cite ratio boxplots in 6 the conclusion can be drawn that **organizations with high influence tend to have high cite ratio medians**. In other words, these institutes hold large proportions of ordinary scholars, thus the assumption beforehand can be verified. Observations on h index boxplots further solidate the conclusion by showing that **more influential universities exhibit lower h index medians** and vice versa. Recall that the derivations of 5 suggest an inverse proportion relation between cite ratio and h index, this phenomenon may as well indicate the existence of quantities of low influence faculties in a top tier institute.

D. A comprehensive view of distributions in various research fields of interest.

Information from DBLP and WOS facilitated the collection and visualization of data concerning fields of Computer Science, Mathematics and Astronomy. Now that the underlying patterns and relationships between h-index and cite ratio are revealed, more research fields are to be analyzed and visualized to verify and solidify the obtained conclusions. We have collected data from the following fields:

1. Chemistry
2. Economics
3. etc.

E. Distributions of academic award winners.

H-index is already sufficient to tell that a scholar has high academic influence, but having great influence is not exactly equal to winning great prizes. By looking into the cite ratio and h-index of these groups of scholars, it is possible that we can draw a clue about why they are able to stand out among peers. Winners for the following awards/prizes are analyzed:

1. Turing award
2. etc.

4. Conclusion

This paper mainly studies the academic relations. We evaluate authors' academic influence by two different indexes: h-index and cite ratio. Author data from three different research fields are first obtained by web crawling. The indexes are then computed and analyzed with the help of Python. In particular, the relation between h-index and cite ratio along with the similarities and differences among cite ratio distributions are intensely studied. Upon drawing conclusions based on observations and derivations, more research fields are analyzed to verify our suspicions. We also look into distributions of various organizations and institutes, along with those of academic prize winners.

ACKNOWLEDGMENTS. Please include your acknowledgments here, set in a single paragraph. Please do not include any acknowledgments in the Supporting Information, or anywhere else in the manuscript.