Evaluating evaluative bibliometrics: a case study of two research groups

Jimi Thaule
University of Agder (Norway)

Marte Strand
University of Agder (Norway)

Evaluating evaluative bibliometrics: a case study of two research groups

By Jimi Thaule and Marte Strand
University of Agder, Norway
jimi.thaule@uia.no, marte.strand@uiano

Abstract
Bibliometrics is a method increasingly used to perform evaluations of scientific output and impact, in particular in order to distribute means, such as research grants. But also internally within universities and other research institutions. Various performance and impact meaures are used to establish quality of research.

This can be highly problematic, not only in terms of ethics, but also with regards to method. Especially considering the proliferation of tools to perform bibliometric analysis, which means that analyses are increasingly performed without actual understanding of bibliometrics as a scientific method.

In our research we have compared two research groups in the same field of research, both from Norwegian universities, and with a similar size and goal. We have used a variety of methods to normalize between them, in order evaluate the ethics and methodical reliability of the results.

We found that in order to compare the two groups for benchmarking purposes we needed to perform a number of normalizations, to the point where it rendered the results largely useless. Too many individual strengths of each group had to be left out of the evaluation in order to compare the two in equal terms.

In our case these problems were increased by the fact that one of the two groups is multidisciplinary, which in turn demanded methods to correct for differing publication patterns within the same group. Without knowledge of the researchers background this is an element that could easily be overlooked, and in turn skew the results in the group's disfavor.

This is turn means that evaluative bibliometrics is in danger of either skewing the results in favor of a certain type of research or group of researchers, or type of publication. In the worst case research funds can be alloted, or entire research groups lose funds based on unsound comparisons and prejudice against certain types of publications.

Introduction
As noted by the authors of *the Leiden Manifesto* there is a tendency to use bibliometrics to compare bodies of publications against each other, to establish which of them has greater impact, and increasingly in order to allocate funding or other resources (Hicks and Wouters 2015, p. 430). Such evaluative bibliometrics can be defined as producing rankings, either of institutions, groups of researchers, publications or publishers (Furner, 2014, pp. 85-86). In some cases this type of evaluation can be misuse of bibliometrics, as important ethical aspects are overlooked, and quality and quantity can be conflated (Furner, 2014 p. 88).

While bibliometrics as a means of evaluation has been present since the seventies, it has rapidly become more noticeable in the last decade, as institutions see bibliometrics as an "objective" way to rank and benchmark, between institutions (such as through university rankings) or internally among groups and individuals (Gingras, 2014 pp. 109-110). At the same time the Internet has made bibliometric data accessible to people with no background in bibliometrics, in turn causing a proliferation of indicators and a rising interest in bibliometrics (Gingras, 2014, pp. 110-111). As evaluative bibliometrics, as a field, has grown in scope and complexity the need for precise, ethical and valid methods of comparison has become more evident, and there are numerous potential sources of erroneous analysis. For benchmarking purposes, it is necessary to find a way to compare *like with like* and to assess whether or not this is meaningful with regards to understanding and evaluating impact, especially in terms of quality (DeBellis, 2009, pp. 199-200). Is it at all possible to say something about the quality of the body of publications produced by the groups, based solely on bibliometric data?

This paper aims to explore common methods for citation analysis used for this purpose, and the results these methods yield. We selected two research groups within the same area of

computer science, with a similar goal, size and research area. Both groups are research groups at Norwegian universities. They will be referred to as group 1 and group 2.

Making the two groups comparable requires some processing and preparation. First of all, although the two research groups have similar research areas they consist of researchers with various backgrounds. Group 1 is a multidisciplinary group, including researchers with background in the humanities, while group 2 consists exclusively of IT-researchers and engineers and has a slightly larger field of interest than group 1. This translates to different patterns of publication: Many of the researchers have published primarily in conference proceedings, especially those in group 2, while some of the researchers in group 1 have published in books.

Also, we must consider that group 2 is slightly older than group 1, and while researchers in both groups have published in the same within the time frame it is certainly possible that the more established history of group 2 can skew the results.

In summary our objective is to explore a variety of methods for citation analysis to evaluate these two research groups, in order to assess whether or not these methods have validity for benchmarking purposes.

Method

Framework:

The primary goal of our selection of methods is to ethically and soundly adjust for the various problems presented in our introduction, while taking the background and composition of group 1 and group 2 into consideration.

In addition to established researchers both groups include master and PhD students, research assistants and chief engineers who have none or few publications. These have been removed from the search results in order to make the groups more comparable on even grounds, however. When removing these both groups consists of 10 researchers. Out of these, nine researchers in group 1 and seven in group 2 have published articles in journals analysed by Web of Science and SCOPUS.

Due to the difference in publication patterns, some researchers publishing in conference proceedings and others in books, we have removed these results from the list in order to normalize for publication patterns, thus focusing on scientific articles only.

As the research area is relatively young, and several of the researchers did not produce results before 2011, a publication window from 2011 to 2016 was deemed most meaningful. By expanding the window to before 2011 we would risk basing our analysis solely on a handful of researchers, and not the groups. Since group 2 was established before group 1 it was also important to choose a citation window that would not favour older publications, and thereby skew the results. Likewise, extending beyond 2016 would entail less complete data, and therefore less reliable results, as not all the publications will be registered in SCOPUS yet.

A synchronous moving citation window of three years was used to analyse the two groups' output. Consequently, citations counted in 2014 are based on articles published in the time period 2011 – 2013, citations counted in 2015 are based on articles published in the time period 2012 – 2014, citations counted in 2016 are based on articles published in 2013 – 2015, and finally citations counted in 2017 are based on articles published in the time period 2014 – 2016.

Data Selection:

We initially pulled data from three sources: Web of Science (WoS), SCOPUS and CRISTin (Current research information system in Norway, a national system for registering academic publications). Our examination relies primarily on data from SCOPUS as SCOPUS returned a higher yield than WoS, while covering all the results from WoS, and because CRISTin does not include citation scores.

However, in order to check the completeness of our data we checked how many of the articles we found in CRISTin were also retrievable in SCOPUS. For group 1 72% of the articles could be found in SCOPUS, and for group 2 the number was 87%. We take this to indicate that the tendencies we found in SCOPUS are largely reliable, and comparable, and the final results are based on SCOPUS findings, as these were the most complete.

Normalization and Fractionalization of the dataset.

From the onset it is evident that attempting to normalize these results based on subject categorization in SCOPUS would prove difficult. Not only because the categories are in

themselves too broad, but also because the two groups have somewhat different scopes (DeBellis, N. 2009, p. 195). The different categories were never developed for comparison, but rather for retrieval, and using them for comparison does not work very well, even though they are increasingly used for this purpose (Leydesdorff, L & Bornmann, L. 2016 p. 707-708).

An attempt to define fields using the journals in which the two groups have published would presumably prove as futile. Considering both the multidisciplinary nature of most journals (Zhou, P. Leydesdorff, L. 2011 p. 361), as well as the multidisciplinarity of this specific field of research itself, and in particular the composition of group 1. Another possible approach could be applying Garfield's *historiographical method* to the corpus of articles produced by group 1 and group 2 (DeBellis 2009 p. 151-153). By mapping all the references used in the articles in our selection we might presumably able to dynamically trace a history of citations and a genealogical development of the subject. This would be quite time demanding compared to the actual need for an accurately defined subject, and the question remains however what value this map of subjects would have.

Therefore, in order to compare group 1 and group 2 we found it necessary to correct for different traditions of authorship, with regards to number of authors per article. This was done by fractional counting of authors in each article. As such each author is given a fractional score equal to one over total numbers of credited authors. The objective of fractionalization is to "control for the in-between field differences caused by different citation potentials" (Zhou, P and Leydesdorff, L. 2011 p. 361). In our case this is in order to correct for the fact that group 1 has authors that primarily publish single author papers, while group 2 has none of these. Normalization by fractionalization has been shown to decrease "in-between group variance in the impact factors by 81%" (Zhou, P and Leydesdorff, L 2011 p. 362). Considering that group 1 includes researchers from a different academic background than the majority of researchers, fractionalizing will presumably correct for different traditions within the group, as well as between the groups.

In terms of field normalization this leaves us with three problems. One of having to deal with a heterogenous group compared to a less heterogenous group, who both define their work as being in the same field. Secondly that the categories in SCOPUS cannot be used for this task, and thirdly a logistical problem of possibly defining a corpus of texts for comparison in order to normalize based on publications. The latter is well beyond the scope of our task as well as highly dubious, as noted above. Consequently, we move away from normalizing through field weighted citations and normalize and fractionalize our data set using author fractionalization.

Limitations:

As noted in the Leiden Manifesto it is important to evaluate based on "a suite of possible indicators", according the field one is studying. (Hicks, Wouters et al p. 430) Doubts can be cast on the validity of the results after removing conference proceedings, as many researchers in computer science rely heavily on this type of publication (DeBellis, N. 2009 p. 196). It must be noted that the results can only be used as a form of benchmarking and comparison of tendencies, not as a complete description of publication patterns. In both groups two of the researchers have published in conference proceedings exclusively.

There are two reasons we decided to remove the conference proceedings. One is a question of validity. While we can find numerous proceedings in Web of Science and SCOPUS, the CRISTin database's categories for registering conference proceedings are heterogenous, and it would be difficult or impossible to compare the two. Additionally, proceedings and articles are cited and distributed in two very different ways, and in order to make a valid comparison we have elected to include only the articles, since they are more numerous in both groups. In spite of the importance of proceedings.

We were left with datasets consisting of a relatively small number of researchers, with 68 and 36 articles in each group respectively. 36 articles is below a recommended threshold of fifty articles. We chose to go forwards with the analysis regardless of this, because the point of our examination is not to conclude about the groups, but to evaluate and explore the methods.

Results:

Based on results in Scopus on author name and author address, considering institutional changes in the time-period 2011 – 2016, the dataset used for the analysis consisted of 68 articles for group 1, and 36 articles for group 2.

Figure 1 illustrates the two groups' total number of citations within the citation windows, with and without self-citations. Due to the multidisciplinary nature of the groups' results removing self-citation was included as both different disciplines and individual researches might have varying traditions of self-citations, potentially skewing the results. Figure 1 demonstrates that although removing self-citations naturally decreases the overall number of citations for both groups, it does so proportionally for each group within all the citation windows. As the proportional number of self-citations is similar within both groups, it does not seem to have a significant effect when comparing the groups.

Figure 1 exhibits an overall steady increase in number of citations for each group, though group 1 displays a slight decline in 2017. As there can be up to 18 months delay in Scopus registrations, this could affect the 2017 citation window.

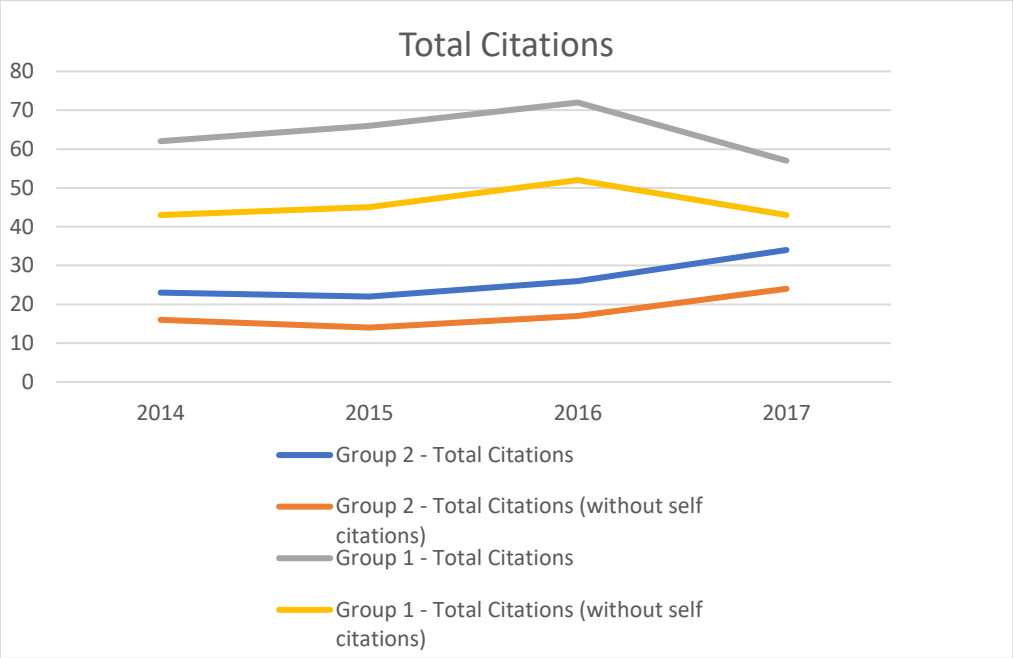

Fig.1 Total number of citations 2014 – 2017, based on a moving citation window of three years.

Figure 1 further illustrates the overall difference in total number of citations between the groups, group 1 receiving more citations than group 2 in all four citation windows. This could be a result of group 2 publishing 32 articles less than group 1, making up only 52% of group 1s total output in the given time-period. In order to see potential tendencies between the two groups less dependent on total output of published articles, the results are also illustrated as average citation per article exhibited in figure 2 (Waltman, L. 2016 pp. 371-372).

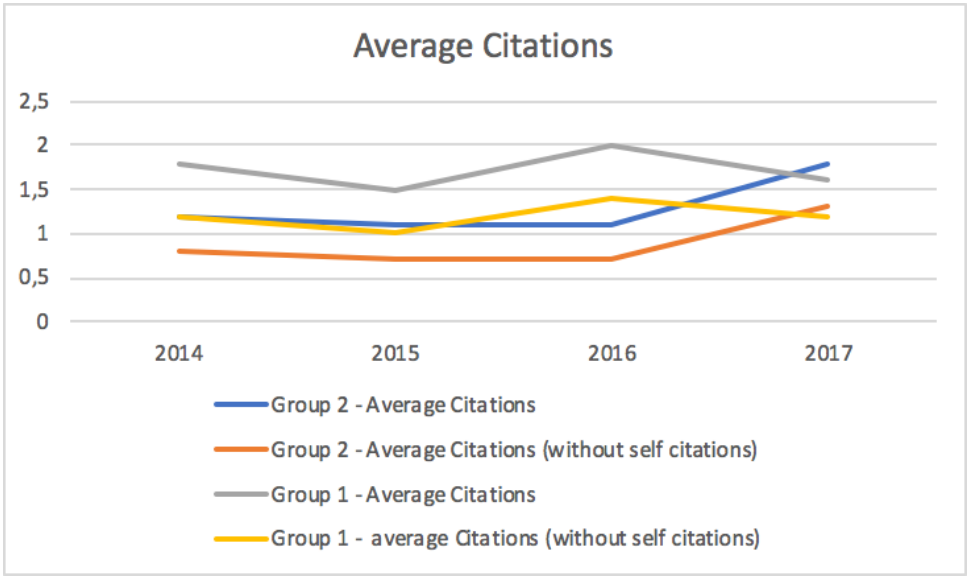

Fig. 2. Average number of citations per articles with and without citations, based on a moving citation window of three years.

Figure 2 illustrates that although group 1 has a much larger number of total citations, the difference between the two groups decrease when looking at average citations.

However, total number of citations and average citations does not account for the groups' potential for different publication, citation and authorship traditions that could influence the finale results.

Zhou, P. and Leydesdorff, L. (2011 p. 367) argued that the different traditions of authorships would have an adverse effect for those disciplines more reliant on single authorships, often applying to social science and the humanities. Thus Figure 3. exhibits total number of citations which has been fractionalized based on number of authors per individual article, to avoid bias.

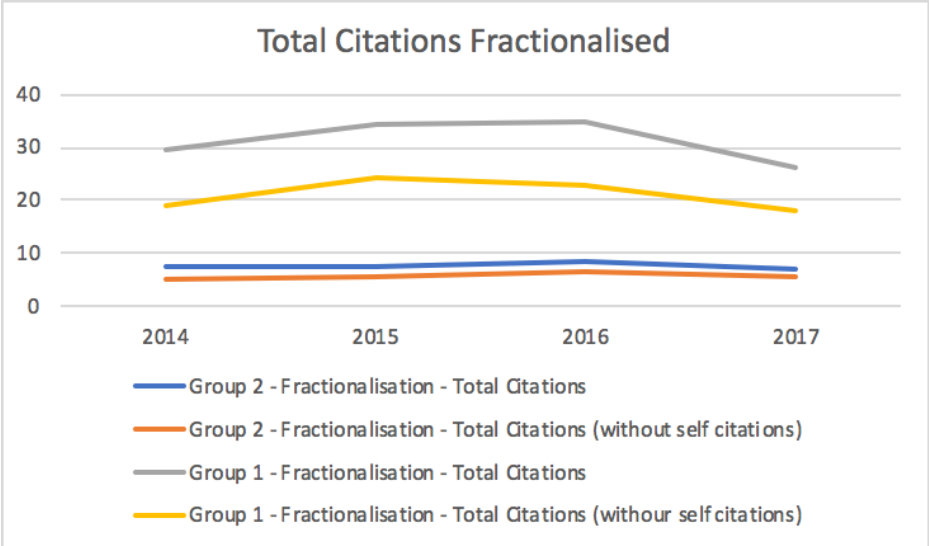

Fig.3 Total number of citations per article fractionalized for number of authors per article, based on a moving citation window of three years,

Though, similar to the tendencies in figure 1, figure 3. exhibits an even greater gap between the two groups favouring group 1. This indicates a difference in authorship tradition within the two groups, where group 1 has a combination multiple authored articles and single authorships, group 2 include only the former.

The difference becomes even more evident when applying the fractionalized numbers on average citation per article as seen in figure 4. The tendencies exhibited in figure. 4

demonstrates a drastic difference from figure 2. Figure 2 showing that the two groups intersecting while figure 4 does not.

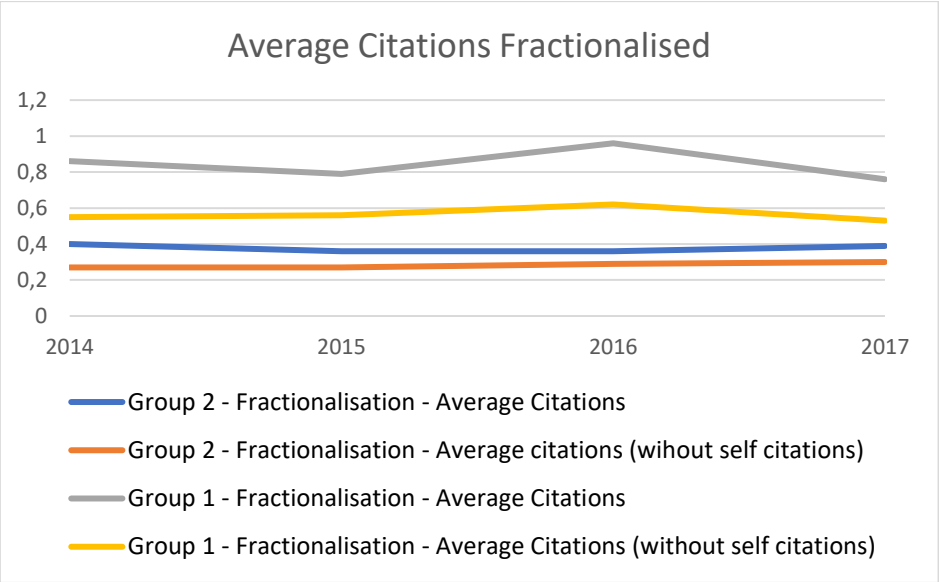

Fig.4 Average number of citations per article fractionalized for number of authors per article, based on a moving citation window of three years.

Discussion

As seen by the results different methods produce different numbers and even tendencies, despite being based on the exact same dataset. This brings us back to the initial question, of whether these methods hold validity as evaluative tools. And in what way they can be used.

Notably, the multidisciplinary nature of group 1 shifts the fractionalized results in their favour, as several of the researchers in group 1 have exclusively or primarily published single author papers. In group 2 single authorship is non-existent. This is in line with the results Zhou and Leydesdorff found in their study of multidisciplinary units in 2011, where fractionalization was seen to "upgrade" the status of social sciences, compared to natural sciences (Zhou, P. and Leydesdorff, L. 2011 p. 367). This can be taken as an indication that group 1 has greater academic impact than group 2. This tendency is especially notable if we calculate the statistical average of the fractionalized citation scores for both groups, thereby exaggerating tendencies for illustrative purposes, as group 1 fluctuates between 0.76 and 0.96, while group 2's average is between 0.36 and 0.4.

It can be argued that the results show that group 1 has a greater academic impact than group 2, despite having existed as a group for a shorter time. The tendency is significantly stronger with fractionalized results. However, this can very well be a result of how the two groups publish differently, and the scope of their work rather than in indication of actual impact. Considering the more practical nature of group 2's research this seems likely, albeit difficult to confirm within the scope of this paper. Regardless, the findings should not be taken as a confirmation that group 1 has more impact than group, but rather as an artefact of their publication patterns.

In order to compare the two groups, various methods for normalization were utilized, and these produced numbers that are theoretically possible to use for comparison. The question that arises is whether or not this process has also rendered some of the numbers meaningless, and that a comparison for comparison's sake overlooks the individual strengths and qualities of group 1 and group 2. It is also quite evident how different methods produce different results, as shown by comparing figures 2 and 4, and these are subject to possible manipulation. Both of these exhibit average citations, one of which adjusted for author fractionalization. Thereby displaying two opposing tendencies. For purposes of evaluation this creates a problem, especially considering how easily obtainable bibliometric data has become, in turn making faulty analyses more likely.

Several other methods of normalization could be applied, in turn producing other numbers and tendencies. The multidisciplinary nature of group 1 could possibly benefit from other types of fractionalizing, for instance looking at primary authorship or citing side citation

analysis. These methods could potentially remove some bias. As pointed out in the Leiden Manifesto this underscores the need for qualitative analysis in addition to quantitative examinations.

As described in the section on limitations of our method we decided to leave out conference proceedings, and focus entirely on articles. Our reasoning behind this is threefold: conference proceedings are cited differently from articles, we were unable to verify our findings with CRISTin due to how papers are registered in CRISTin and finally because academic impact is primarily seen through articles and both groups published most of their output in the form of articles. As such we wound up with numbers that are comparable to each other between the groups, but the level to which they reflect the real output and impact of group 2, in particular, is unclear. In order to more fully understand their impact, we would have had to develop a method for including conference proceedings and correcting for different citation and publication patterns of proceedings.

An intrinsic question that needs to be posed, with regards to the general ethics of bibliometric evaluation, is whether or not what we measure is what we seek to measure. It can be assumed that academic output equals productivity in the academic sector, but this might not be the case (Furner. 2014 p. 88). In addition to producing articles and other academic papers the members of group 2 carry out practical experiments, whose impact necessarily cannot be measured by bibliometrics. As such, a bibliometric analysis can contribute to obscure the actual productivity and significance of a research group. We managed to compare like with like, as we set out to do, but may have lost some meaningful aspects of group 2 in order to do so.

While the various methods we have used to compare group 1 and group 2 have been useful as a study of the groups, we believe caution is to be advised. There is much to be learned from the output and patterns, but perhaps not as a means of comparing impact. To fully understand the impact group 1 and group 2 have in their field a more qualitative study is necessary, and not a solely relying on metrics. Significantly our study has no way of showing the impact of the practical experimental work carried out by group 2, in particular.

In conclusion we might say that although anything can, and maybe even will, be compared, a responsible and meaningful comparison is much harder to perform. In this case a responsible comparison gives more meaning in terms of the sociology of publications, rather than an evaluation of impact. Ultimately our paper shows how easily numbers can be manipulated and skewed in one direction or the other, and therefore how easily bibliometric analysis can be misused. This does not mean that bibliometrics as a method has no role in evaluations. It does however mean that bibliometrics alone should not be used to make decisions.

Literature:

DeBellis, N. (2009). *Bibliometrics and citation analysis: from the science citation index to cybermetrics.* Toronto: Scarecrow Press

Furner, J. (2014). The Ethics of Evaluative Bibliometrics B and Sugimoto, C.R. *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact*. Cambridge, MA: MIT press pp. 85-108

Gingras, Y. (2014). Criteria for evauating indicators Cronin, B and Sugimoto, C.R. *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact.* Cambridge, MA: MIT press pp. 109-126

Hicks, D. and Wouters, P .(2015). The Leiden Manifesto for research metrics *Nature vol. 520 pp. 429.431*

Leydesdorff L, and Bornman, L .(2016). The operationalization of “fields” as WoS subject categories (WCs) in evaluative bibliometrics: The cases of “library and information science” and “science & technology studies” *Journal of the association for information science and technology* (67) 3 pp. 707–714

Waltman, L .(2016). A review of the literature on citation impact indicators. *Journal of informetrics* 10(2): 365-391

Zhou, P and Leydesdorff, L. (2011). Fractional Counting of Citations: A cross- and interdisciplinary assessment of the Tsinghua University in Beijing *Journal of Informetrics (5) 3 p. 360-368*