KB Quick Notes

Dr. Stephan Gauch [iD]

Robert K. Merton Zentrum für Wissenschaftsforschung

Humboldt-Universität zu Berlin

Table of contents

What is all this?

In the legal professions, there is a concept that in German carries the wonderful name "Lose-blattsammlung" (aka looseleaf binder, often as part of a subscription then called looseleaf service). You pay for the joy of receiving bundles of notes on some legal topic or issue, that you then may neatly file away in a special folder in order to be "up to date" on all the relevant facts and concerns surrounding said topic. In a way, what you are currently looking at is also some sort of a looseleaf binder, in that it aims to inform you on things. Yet, in contrast to being "all up to date all the time," it is rather something that should get you "up to speed" in that it gives you a starting point for a subset of topics and concepts in bibliometrics. Also, there is no real service to subscribe to. Rather, these notes are a starting point of starting points.

But… why?

"Why not write a proper textbook?" you might say. And yeah, sure, thanks, great suggestion! I probably could give a number of excuses. But rather than those, I give you this. Also, I thought it would be best to have something to just jump in without prior training in order to set a basis for an inclusive discourse about the concepts and ideas discussed. Having recently screened the state of bibliometrics in use, I see a lot of contributions that would probably not qualify for publication in a bibliometric core journal. I don't think that is a problem in itself. I also don't subscribe to the notion that authors of such contributions are per se "incompetent" or just have to "see the light" or should be prevented from trying new things. Not necessarily (only) for moral reasons - the latent elitism should be obvious - but for pragmatic ones. There is a lot of talk about being "responsible" when working with metrics. That's true for sure, right? Yet, it also is also quite funny in my book, since one probably would have a bit of a hard time to find vocal advocates for a) the "irresponsible" use of metrics, or, for that matter, b) "meaningless" metrics or c) measuring what is "irrelevant". So these pleas don't really do much on their own. Anyway! Rather than thinking of commandments, I opted for this: Providing a simple starting point to allow interested newcomers to join the party.

Who (I think) might enjoy reading this?

Well, first, foremost and obviously, anyone who wants to know what bibliometrics is about.

Obvious candidates are:

- Students not having access to an introductory bibliometrics course and want to start learning.
- Students that do have access to an introductory bibliometrics course, but struggle with getting a grip on the basic concepts and how they relate to each other.
- Practitioners and administrators planning to use bibliometrics in their work and need a starting point (sic!) on what's what.
- Policymakers who want to engage in a discourse with the bibliometric community and connect on the level of language use, what concepts bibliometricians believe to be relevant and why they believe that.
- Anyone who wants to better understand a bibliometric study.

Where to start with the Quick Notes?

Is there a structure to the individual notes? Do they have to be read in a particular order? Yes, and no. No, you may read this in any order you see fit. Yes, starting with the basics might be a good idea, but I guess you would have figured that out on your own and would not need a cue for that. Also yes, I tend to bracket bibliometrics into the following territories and did so in the Quick Notes as well.

Evaluative bibliometrics, the territory of the vertical, where the tideous repetitiveness lives. What is "countable", based on an argument of shared characteristics, e.g. one citation being just like any other citation, is stacked up to find out who are the "best of the best of the really very, very bestest". For better or worse as numerous reform initiatives suggest.

Explorative bibliometrics is the territory of the horizontal. The realm of mapping and charting the old and the new or finding some path from one peculiarity to another. How do topics evolve? Who cooperates with whom? These sorts of things. The Explorative and the Evaluative are quite neighborly, since you can try to translate many horizontal phenomena into vertical signifiers of "value" or "quality". A bibliometric mapping exercise can be both speak for "success in community building" as well as "representation of a community". Hence, borders here are sometimes rather loosely defined. Yet, both can have rather different ideas of what is "relevant" or "adequate".[1]

Both of these territories share that they work "with" databases in some way or another.

Curative bibliometrics is the territory responsible for keeping things in order. Pruning, weeding things out, naming all the things properly, etc. In practical terms: Making sure that mapping and counting can be done "adequately" and in a proper orderly fashion. Obviously, this is less working with the database but rather "on" the databases.

---

[1]Using a "citation window" means to count the incoming citations to a contribution of a pre-defined period of time in order to prevent unfair comparison of old and new publications. After all, the old ones had more time to get cited. Makes sense in evaluation, but probably less so when the aim is to explore or map something. The difference between a) "We need to compare fairly and therefore have to control for unfair advantage of old publications!" does not play well with b) "Why cut off a part of the past of a field, when your interest is in the past of a field?". There is, of course, a Quick Note on Citation Windows.

Finally, **reflexive bibliometrics**, a territory that I consider myself the most with nowadays, is a home for those who do not necessarily(!) work with databases or not necessarily(!) work on databases - usually they can do both, too - but try to understand how working with and working on databases shapes what bibliometricians do, what others do with what bibliometricians do, and what others, e.g. researchers, do when observing what other others do, e.g. funders, with things that bibliometricians do and how all of this plays together. So, inhabitants are rather interested in the idea of working on the profession of bibliometrics and performativity of bibliometrics, i.e. how measures or maps rather "produce" what they seek to "show".

Anyway. I digress, again. Let's move on.

How the Quick Notes are structured

The notes themselves definitely have an internal structure. All documents start with a "What is this about" section, giving a rough idea what concept the note is about. Next, you will find some sort of a situating passage. Why is that concept important in bibliometrics? Where applicable, you will find a passage on "How does this work?" giving some verbalized idea of the procedures involved.

The Quick Notes do NOT contain code. Preposterous! Where is the tutorial? Isn't bibliometrics also a craft? Of course it is! And there might be more to come. Eventually. I guess. Anyway!

The last section contains limitations and critique towards or at least related to the concept. These are not the result of a careful screening exercise but rather me channeling the field. Last but not least: References. Another starting point. Obviously, that selection is illustrative and not exemplary.

Another aspect that structures the Quick Notes is their length. They are almost insultingly short given the glorious purpose an ambitious reader might throw at them. Needless to say: Each is a rabbit hole. A bottomless pit of sorts. Each and every one of the topics makes for a quite wonderful fetish. Consider these notes a shovel, at maximum a very tiny and short ladder. The absolute minimum that these should achieve though is getting you, dear reader, to be able to discuss these issues with those that are more versed and have already dug a deeper hole for themselves. The absolute maximum is getting you hooked on one or more of these fetishes.

That's an awful lot of starting points and not much of a conclusion

Indeed. This is intentional and pretty much by design. These notes are supposed to leave you, dear reader, a tad unfulfilled. So it's not an all-inclusive trip but rather a badly drawn map on a napkin to instill the recklessness of, dare I say, adventure. Also, you might have realised that the Quick Notes don't have a DOI. This is not a coincidence! You should, for multiple reasons, not reference these notes. First of all, they are, for reasons argued above and didactic purposes,

shallow, dramatically short, on the brink of being oversimplified. The aim is to have you understand the basics. So, rather than referencing these notes, I would rather have them shared.

Some final remarks: Why did you not make this a proper Wiki? Well, all of this is licensed CC-BY. Go forth and wreak havoc. Why did you not cover [insert your favourite fetish here]? Great question! Thank you! Can I contribute to this or write a Quick Note myself? Color me delighted! For the moment, I guess you might just contact me about this: stephan.gauch@hu-berlin.de.

Anyway.

Enjoy the ride.

Berlin, 18 Jul 2025

Acknowledgments

Basics

Bradford's Law of Scattering

What is Bradford's Law about?

Bradford's Law, formulated by Samuel C. Bradford in 1934, is a pattern often found in bibliometric studies, namely that in bibliometrics pretty much everything is distributed *log-normal*, i.e., among other characteristics, heavily right-skewed. The law describes the scatter or dispersion of scientific literature, that, if scientific journals are arranged in order of decreasing productivity of articles on a given subject, they may be divided into a nucleus of journals (so-called core journals) more particularly devoted to the subject and several groups or zones containing the same number of articles as the nucleus when the journals are increasingly less representative to the subject as a whole, i.e. they contain less works related to a specific matter. Bradford's law then predicts the number of journals within the nucleus and zones. The relationship of size of the nucleus and the zones is, at least according to the original paper by Bradford: $1 : b : b^2$. This may not be universally applicable today due to changes in the scientific publication landscape. Yet, in principle, the concept still holds up in the way Bradford's Law is famously summarized as:

> A small core of journals will account for the majority of significant scientific papers.

Why is Bradford's Law important?

Bradford's Law has been important for understanding and managing the expanse of scientific literature, assisting librarians and information scientists in identifying influential journals in a particular field, thus enabling allocation of resources for library collections. In the context of practical applications it may therefore support resource management in libraries by focusing on core journals. It also aids researchers in targeting the most focused journals for their work, both for reading and publication purposes. In this sense, it is also important for both evaluative and exploratory bibliometrics. Especially when aiming to evaluate or explore a thematic field, be it a discipline or a topic, using, at least in part, a journal-based strategy, Bradford's Law will highlight both the strength and the weakness of a journal-based approach. First and foremost, Bradford's Law provides an excellent argument to focus on core journals to produce an effective query for delineating a field or topic as it helps in identifying the most productive journals quickly and at the same time provides an argument, why there are limits to journal-based strategies; especially in the 2nd zone, the number of journals is squared compared to the 1st zone and this 1st zone already contains a factorised number of journals compared to the original core journals. The strengths in relying on Bradford's Law to design effective queries are that a journal-based strategy will produce a vast number of

results very quickly by identifying the core journals, perhaps together with domain experts on the subject in question. This benefit is usually in part qualified by publisher practices aiming to bundle journals into subscription packages providing marginal reductions when choosing said packages over individual subscriptions. Finally, Bradford's Law may be used to give a rough prediction of a field's size. Following the general rule of relationship between the 1st and 2nd zones, very rough assumptions can be made about how *large* a scientific field might be.

Limitations

Sticking to Bradford's Law will make your query more *mainstream*

The main weaknesses lie in coverage. Over-reliance on Bradford's Law will produce results that may over-represent the *mainstream* of a topic or field. In a similar vein, optimizing search strategies gearing toward Bradford's law may not be possible in highly emerging topics or fields, or with high degrees of interdisciplinarity, where the publication strategies of authors may be geared toward more generalized outlets due to the lack of specialized journals not being available in early stages.

Bradford's Law might be driven by multidisciplinary applications

Another limitation of Bradford's Law lies in the subjectivity of both article classification and journal selection. Especially in the long-tail of the distribution a single article in a journal will contribute to increasing the skewness of the overall distribution. In the case of bibliometrics, which is increasingly being used as a basis for tracing actual developments of fields or topics, e.g. in the context of systematic reviews, a large number of papers are *applications* of bibliometrics rather than bibliometric research in the stricter sense.

Bradford's Law is (in some sense) field dependent

Bradford's Law may not apply uniformly across different subjects or disciplines. How it diverges from the ideal constants and coefficients may be dependent on a number of characteristics, i.e. a field's age, the application orientation of a field, tendencies for interdisciplinarity, rhythms of scientific outputs, funding etc. The digital age and open access movements have further altered the publication landscape and access patterns, which may also further affect the applicability and usefulness of Bradford's Law.

Bradford's Law abstracts from quality

Bradford's law makes absolutely no claim about quality. Scholars might jump to the conclusion that articles in the fringes, i.e. the 2nd zone, may always be of *low quality*. Even though anecdotal evidence might reflect this, especially in fields of high application-orientation, this can not qualify as a general rule. One prominent example might be the Hirsch-Index, a prominent measure of the productivity and impact of a researcher in their field. The article introducing the Hirsch-Index was published in Proceedings of the National Academy of Sciences (PNAS). Definitely not one of the core journals, an influential paper nonetheless.

Further Reading

Bailón-Moreno, R., Jurado-Alameda, E., Ruiz-Baños, R., & Courtial, J. P. (2005). Bibliometric laws: Empirical flaws of fit. *Scientometrics, 63*(2), 209–229. https://doi.org/10.1007/s11192-005-0211-5

Bradford, S. C. (1934). Sources of information on specific subjects. *Engineering, 26*(4), 85–86.

Glänzel, W., & Thijs, B. (2012). Using 'core documents' for detecting and labelling new emerging topics. *Scientometrics, 91*(2), 399–416. https://doi.org/10.1007/s11192-011-0591-7

Mutschke, P., & Mayr, P. (2015). Science models for search: A study on combining scholarly information retrieval and scientometrics. *Scientometrics, 102*(3), 2323–2345. https://doi.org/10.1007/s11192-014-1485-2

Nicolaisen, J., & Hjørland, B. (2007). Practical potentials of Bradford's law: A critical examination of the received view. *Journal of Documentation, 63*(3), 359–377. https://doi.org/10.1108/00220410710743298

Shenton, A. K., & Hay-Gibson, N. V. (2011). Bradford's Law and its relevance to researchers. *Education for Information, 27*(4), 217–230. https://doi.org/10.3233/EFI-2009-0882

Citation Windows

What is a Citation Window?

A citation window refers to the specific time period during which citations to a contribution are counted and analyzed. This window can range from a few years to several decades. In practice, however, the periods for citation windows typically range between 2 and 5 years.

Why is it Important?

Using citation windows is important because they directly influence the calculation of impact of a contribution. They achieve this by addressing a specific notion of *fairness* that relates to the comparison of recent and old research. The first argument for citation windows is that old articles had more time to accumulate citations and therefore have an unfair advantage over recent ones. The second argument is that in order for a scientific article to be referenced at all, it usually needs a certain amount of time to get recognised in the publication landscape. The rather idealized idea is that an article is being read, has some relevance in some research process, the research itself is being performed, the article is being referenced in a manuscript, which again is submitted to a journal and goes through rounds of peer review until finally, after publication, and some additional time to be included in a bibliographic database, the original contribution we are interested in, has attracted a citation. Obviously, different fields have different speeds at which research is cited. For instance, in rapidly evolving fields like information technology, shortening the citation window may be more helpful, while in disciplines with longer research cycles, such as history, longer windows may be necessary.

How Does it Work?

The length of a citation window strongly depends on the task at hand, the requirements toward immediacy or the research question as well as the practices of the field in question. Citations are then collected and analyzed only within this predefined period. For instance, a 5-year citation window starting from the year of publication means that only citations received within those five years are considered in the analysis. This information is not always part of publicly available data.

Limitations

Does not account for slow burns

One of the primary issues with citation windows is the risk of misrepresenting the impact of research. Short windows may not capture the long-term influence of a contribution, especially in fields where citations accumulate slowly. Such *sleeping beauties* might receive little attention at first but a surge in reception after some time. If this period is no longer within the citation window this reception is being ignored.

The longer the citation window - the older (and less relevant) the results

Longer windows may include citations that are less relevant to the current state of the field, due to different rhythms of scientific output. In some fields publication propensity favors shorter time frames from research to publishing. This effect may even be catalyzed further when considering differences in project or research process length. In fields where rhythms are tight, longer citation windows might cover things that are no longer relevant. Yet, context matters, i.e. distinctions have to be made between citation windows as a means to discount relevance vs. a means of preventing measures from being overly influenced by matters of temporality.

Citation windows might not fit the notion to be evaluated against

Even though citation windows are useful for matters of comparison it may not be sensible to use when temporality is explicitly part of the concept to evaluate against. One example can be the Hirsch-Index, which puts a strong premium on *breadth of impact*, which in turn makes it a bad measure to use in the context of evaluation of junior scientists. In the case of the Hirsch-Index the notion of *breadth* does not match up to the measurement logic of restricting temporality. Similarly other evaluations that somehow address notions of *long-termedness* will not benefit from citation windows.

Citation windows might be counterproductive for exploratory purposes

Citation windows make intuitive sense in the context of evaluative bibliometrics. In exploratory bibliometrics they might be counter-productive. For instance, when questions of evolution or dynamics of a field are relevant to the purpose of exploring. More often than not, the exploratory bibliometrics involve understanding the *origin story* of a field or topic. Where did it come from? How did it emerge? In this case limiting analysis by using citation windows may eliminate just these interesting classics, e.g. from co-citation analyses.

Citation windows only address one issue

Time isn't the only factor that influences citation counts. Document type is another very plausible candidate with review papers usually receiving a considerable premium on citations flowing in. Moreover, the choice of a citation window can introduce bias, as it might favor certain types of publications or disciplines over others. This might not be a strong limitation or argument against the use of citation windows but should just be a reminder that citation windows do not fix all potential biases.

Citation windows may be inherently inaccurate

When calculating citation windows analysts have to be aware of what information specifically the citation window is calculated on. Sometimes, when calculation is performed using the publication year, publications from early months in the year may receive a substantial premium. Using the actual publication date might change the story quite a bit.

Be aware of difference in citation windows

There is no universal standard when it comes to citation window length. When interpreting or merging data the mere information that a citation window has been applied may therefore be insufficient. There's also the challenge of comparing studies using different citation windows, which can lead to inconsistencies in interpretation of indicators.

Further Reading

Campanario, J. M. (2011). Empirical study of journal impact factors obtained using the classical two-year citation window versus a five-year citation window. *Scientometrics, 87*(1), 189–204. https://doi.org/10.1007/s11192-010-0334-1

Donner, P. (2018). Effect of publication month on citation impact. *Journal of Informetrics, 12*(1), 330–343. https://doi.org/10.1016/j.joi.2018.01.012

Glänzel, W. (2004). Towards a model for diachronous and synchronous citation analyses. *Scientometrics, 60*(3), 511–522. https://doi.org/10.1023/B:SCIE.0000034391.06240.2a

Glänzel, W., Schlemmer, B., & Thijs, B. (2003). Better late than never? On the chance to become highly cited only beyond the standard bibliometric time horizon. *Scientometrics, 58*(3), 571–586. https://doi.org/10.1023/B:SCIE.0000006881.30700.ea

Wang, J. (2013). Citation time window choice for research impact evaluation. *Scientometrics, 94*(3), 851–872. https://doi.org/10.1007/s11192-012-0775-9

Coverage in Bibliometrics

What does Coverage mean in Bibliometrics?

In bibliometrics, *coverage* refers to the extent to which a bibliographic database (or tool) includes relevant publications, journals, conferences, and other scholarly outputs within a specific field or across multiple disciplines. Coverage can be evaluated in terms of the breadth (range of subjects, disciplines, or publication types) and depth (historical range, level of detail in indexing) of the included materials. Effective coverage is crucial for comprehensive bibliometric analyses, as it directly impacts the accuracy and reliability of the results derived from the data. Besides the notion of coverage from a disciplinary perspective, it can also refer to the occurrence of missing data on the level of individual records in a database, which may also be referred to as *completeness*.

Why is it Important?

There is no complete database of scientific publications including all the features that would qualify current bibliometric analyses. In that sense all databases are defined by a focus, some might say bias, towards specific criteria of inclusion. Both notions of coverage have substantial implications for bibliometric analyses. For instance, adequate coverage ensures a more complete and accurate representation of the scholarly landscape, which in turn enables more valid comparisons across different fields, institutions, or time periods. Despite its importance for bibliometrics, balancing quality, i.e., more precisely, and more problematically, citation counts, vis-a-vis quantity or completeness might not be as straightforward as one might expect in matters of coverage. Among the current considerations are matters of post-colonial concern, e.g. contributions by the *periphery* of the science system such as the global south, with some scholars rejecting the notion of a primacy of the center vs. the periphery. Other aspects that problematize coverage in this way is identification of trends and gaps, where a higher coverage assists in identifying emerging research trends and potential gaps in the literature. The issue therefore is more complex than *the more, the merrier*.

Evaluating and ensuring coverage usually involves multiple aspects, such as:

1. Assessment of Bibliometric Sources: Examining the scope and extent of bibliometric databases and tools to determine their coverage.
2. Transparent Data Inclusion Criteria: Defining criteria for what types of publications and sources are included.

3. Regular Updates and Expansion: Continuously updating and expanding databases to include new publications, journals, and other relevant scholarly outputs.
4. Cross-Database Comparison: Comparing data across multiple bibliometric sources to identify coverage overlaps and gaps.

How Does it Work?

Coverage is a concept that is more a matter of consideration than something that can be *performed* as a method in the stricter sense. More often than not, due to the commercial nature of bibliographic data and available resources. With databases from the Web-of-Science-Family the criteria usually will be a derivative of citations over time relative to publication volume on the level of journals taking into account the disciplinary focus of said journals, a.k.a *quality*. Scopus seemingly aims for more of a balance between quality and quantity. Yet, in both cases, how, when and why a journal is included is not overly transparent. Moreover, which journals are being *put to review* is unclear as well, rendering the overall composition a bit of a mystery at times. Descriptions of working groups or committees in part exist, yet these are also only moderately informative toward the issue. The prevalence of focus is not just present in established databases but also extends to current contenders. Sometimes at the cost of the second notion of coverage. The platform *Dimensions* features a larger collection of articles leading to an overall larger citation network but has issues with completeness.

Limitations

Adequate Coverage is pricey

Achieving and maintaining extensive coverage requires significant resources and effort. This is especially true when understanding coverage as a dynamic concept rather than a static silo logic at a definite point in time. Main drivers are changes in the topic, field and publication landscape.

Coverage is politics by other means

Coverage can be biased towards certain languages, regions, or disciplines, leading to skewed perspectives both from an evaluative as well as explorative perspective. Keeping up with the rapidly evolving landscape of scholarly publications is a continuous challenge.

Coverage is exclusion

Coverage should not be confused with unreflected inclusion. Selecting journals, repositories, hubs or articles always involves a judgment, even if done *100% algorithmically*. In this sense understanding coverage can (and maybe should) be understood more along the lines of curation. Extensive coverage may result in an overwhelming amount of garbage data, complicating analysis and interpretation.

Further Reading

Daniel, B. Klein, & Chiang, E. (2004). The Social Science Citation Index: A Black Box—With an Ideological Bias? *Econ Journal Watch, 1*(1), 134–165.

Harzing, A.-W., & Alakangas, S. (2016). Google Scholar, Scopus and the Web of Science: A longitudinal and cross-disciplinary comparison. *Scientometrics, 106*(2), 787–804. https://doi.org/10.1007/s11192-015-1798-9

Larsen, P. O., & Von Ins, M. (2010). The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics, 84*(3), 575–603. https://doi.org/10.1007/s11192-010-0202-z

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: A comparative analysis. *Scientometrics, 106*(1), 213–228. https://doi.org/10.1007/s11192-015-1765-5

Stahlschmidt, S., & Stephen, D. (2020). Comparison of Web of Science, Scopus and Dimensions databases. Berlin: Deutsches Zentrum für Hochschul- und Wissenschaftsforschung. https://bibliometrie.info/downloads/DZHW-Comparison-DIM-SCP-WOS.PDF

Institutional Disambiguation in Bibliometrics

What does disambiguation mean in Bibliometrics?

Disambiguation of affiliation data refers to the process of accurately identifying and distinguishing the institutional affiliations of authors in scholarly publications. This involves resolving ambiguities and variations in the way institutions are named or represented in publication data. For instance, an institution could be referred to by different names or acronyms, or multiple departments within a single institution could be listed separately. The aim is to ensure that each publication is correctly attributed to the right institution. This also includes diachronic changes, i.e. affiliations splitting up or merging, like in the case of the Karlsruhe Institute of Technology or (at least temporarily) the Berlin Institute of Health. Apart from institutional disambiguation, similar principles are being applied to disambiguate author names, which usually proves to be a significantly harder challenge for various reasons.

Why is it Important?

Disambiguation is the prerequisite for accurate attribution. This aspect might be the simplest factor why disambiguation is important in bibliometrics: to ensure correct attribution of research output to institutions, crucial for institutional rankings, reputation, and funding. The notion of accuracy of attribution is not limited to evaluative bibliometrics such as productivity assessments and impact evaluations. Rather, it also is relevant, maybe even more relevant, in exploratory bibliometrics, e.g. in the context of analyses of collaboration networks. Yet, disambiguation also provides further benefits beyond bibliometrics, e.g. in the context of in-house library and resource management of organizations. Finally, disambiguation supports overall research visibility of an organization by accurately showcasing an institution's research contributions as well as enhancing overall addressability, aiding in visibility and recognition in industry, policy and the academic communities. All in all, disambiguation provides the basis for fairer and more meaningful comparisons between institutions and is indispensable for providing clean and high quality data for policy-making and strategic decisions in research management.

How Does it Work?

There are numerous disambiguation approaches. Some will focus on the use of structure-detection using multiple pieces of information. Others will focus on fuzzy matching incorporating spelling errors. Other approaches again might focus on so-called *master list approaches* that are basically extensively curated thesauri of spelling variants. Other approaches again aim to use complementary data, such as WikiData, to clean institution strings. In the Competence Network Bibliometrics the current approach is to use regular expressions, basically matching for delicately definable patterns rather than *plain* searches, to define an enormous set of positive and negative rules on how institution strings are consolidated. These rules are applied to clean, if possible, to the level of institutes and then aggregate back to the level of the institution as a whole including start and end dates to account for changes in institutional setups. Most of the approaches mentioned have a quite similar workflow, which includes the following steps.

1. Data Collection: Gathering affiliation data from consolidated data sources or publication records.
2. Identification of Variations: Recognizing different variations and representations of the same institution (differences based on the approach chosen).
3. Standardization and Matching: Standardizing the names and details of institutions and matching different variants to a single, standardized form.
4. Continuous Updating: Regularly updating the disambiguation process to accommodate new institutions, mergers, name changes, etc.

The recursiveness and circular nature of this idealized approach should make it clear that disambiguation usually is not a *one-shot* task but rather a continuous effort. Yet, for specific approaches and questions using smaller datasets it may be admissible to use combined approaches including structure detection and manual inference to arrive at high quality data.

Limitations

All data is dirty all the time

Disambiguation can quickly become a highly complex and resource intensive task. Assuming that commercial, or non-commercial for that matter, databases are sufficient for evaluatory or exploratory purposes might be a gross lapse of judgment. Even though newer approaches aim to mitigate the problem and some providers try to shift cleaning either into algorithms, communities or clients, there currently is no database that is perfectly clean in the regard described above, even though they might claim so in sales pitches and promotional material.

Disambiguating data is like feeding a dragon

The process can be complex and resource-intensive, requiring sophisticated algorithms and expert intervention. All these things are costly and will require continuous and/or distributed efforts to achieve. As soon as manual inference comes into play knowledge about the national or local organizational landscape both within and beyond stereotypical *research organizations* is an absolute must! At least for now high quality data simply won't come cheap. Also, affiliation data is continually changing, making ongoing maintenance a challenge as well as an imperative.

Don't expect disambiguation to be standardized (yet)

Disambiguation will sometimes feature an obscene variability across sources. How affiliation data is processed by the different data providers is not consistent or comparable. This lack of standardisation can make integration of data and harmonization efforts a complex task.

Further Reading

Daraio, C., Lenzerini, M., Leporelli, C., Naggar, P., Bonaccorsi, A., & Bartolucci, A. (2016). The advantages of an ontology-based data management approach: Openness, interoperability and data quality. *Scientometrics, 108*(1), 441–455. https://doi.org/10.1007/s11192-016-1913-6

Donner, P., Rimmert, C., & Van Eck, N. J. (2020). Comparing institutional-level bibliometric research performance indicator values based on different affiliation disambiguation systems. *Quantitative Science Studies, 1*(1), 150–170. https://doi.org/10.1162/qss_a_00013

Müller, M.-C., Reitz, F., & Roy, N. (2017). Data sets for author name disambiguation: An empirical analysis and a new resource. *Scientometrics, 111*(3), 1467–1500. https://doi.org/10.1007/s11192-017-2363-5

Rimmert, C., Schwechheimer, H., & Winterhager, M. (2017). Disambiguation of author addresses in bibliometric databases - technical report. Bielefeld: Universität Bielefeld, Institute for Interdisciplinary Studies of Science (I²SoS).

Tang, L., & Walsh, J. P. (2010). Bibliometric fingerprints: Name disambiguation based on approximate structure equivalence of cognitive maps. *Scientometrics, 84*(3), 763–784. https://doi.org/10.1007/s11192-010-0196-6

Precision and Recall

What are Precision and Recall

One of the main findings of the Cranfield Experiments was the development of two measures that help to formulize how improvement in several areas of information science can be measured, e.g. how good a search query is, how well a classification algorithm corresponds to expert judgment or how to evaluate the effectiveness of search systems. In a way it shares characteristics with Alpha and Beta errors, which are prominent measures in inferential statistics about asserting or rejecting a Null hypothesis and whether this is right or wrong. To understand Precision and Recall it is important to understand two separate issues. First, if documents in a result set are assigned as relevant or not relevant and second if that assignment is true or false, resulting in 4 (relevant vs. not relevant x true vs. false) conditions.

> True Positive (TP): A document is deemed relevant and this assessment is true

> False Positive (FP): A document is deemed relevant and this assessment is false

> True Negative (TN): A document is deemed irrelevant and this assessment is true

> False Negative (TN): A document is deemed irrelevant and this assessment is false

Precision refers to the proportion of retrieved documents that are relevant. This is calculated in the following way: Take the number of relevant documents found (True Positives) and divide those by the total number of documents retrieved (the sum of True Positives and False Positives), or, in mathematical terms: $\frac{TP}{(TP+FP)}$.

This can be understood as a measure of *correctness*, i.e. a perfect Precision means:

> All documents found are relevant!

Recall, on the other hand, refers to the proportion of relevant documents found. This is calculated in the following way: Take the number of relevant documents found (True Positives again) and divide those by all relevant documents there are (the sum of True Positives and False Negatives), or, again in mathematical terms: $\frac{TP}{(TP+FN)}$.

This can be understood as a measure of *completeness*, i.e. a perfect Recall means:

> All relevant documents have been found!

Precision and Recall sometimes are reported as percentages, e.g. 90%, or in decimal notation, 0.9.

Why is it Important?

In bibliometrics, Precision and Recall are crucial for assessing the performance of information retrieval systems, such as search engines and digital libraries, the quality of classification algorithms, procedures of disambiguation of authors or affiliations, quality of search queries and much more. It also helps to balance optimization strategies when doing bibliometric research. For instance, when the potential of manual post-processing is given, e.g. in an exploratory analysis when the corpus of publications will be screened for publications fitting the intended field or topic delineation, a *high recall strategy* will overall yield better results because the initial corpus will be larger and irrelevant things can be weeded out.

A similar argument can be made for classification algorithms that are being post-processed manually. Over-indexing on Recall of course could be problematic, too, leading to a search strategy that simply returns, in the most extreme cases, all records in a database. High precision strategies may be more appropriate when defining a seed, e.g. when post-processing a search strategy by incorporating structural aspects such as citation networks. In this case, optimizing for precision might be worthwhile as moving back and forth in the citation tree will automatically increase recall. Overdoing high-precision strategies is not optimal either. In the most extreme case a single document might be returned which is really close to the topic but several other documents are not being found, making the seed highly dependent on this single paper leaving out substantial components, e.g. competing paradigms.

Limitations

You usually can't have it all

There usually is a substantial trade-off between Precision and Recall. Often, improving Precision leads to a decrease in Recall, and vice versa, making it challenging, if not impossible, to optimize both to very high values.

Relevance does all the work

Sometimes, especially when doing explorative work, assigning relevance may be a very challenging task, especially when a bibliometrician has little or no domain knowledge about the field to be analyzed. Furthermore, relevance is at times elusive and highly context dependent subjecting relevance, and hence, Precision and Recall, depending on the user's need and the specific application.

The recursive problem of completeness

When working exploratively a complete set of relevant documents IS the actual goal. Yet, Recall implies that we know this in advance, framing the goal to attain a complete set of relevant documents being in part dependent on having information about the complete set of relevant documents. This is obviously recursive and therefore problematic. It also shows the limits of Precision and Recall, i.e. carefully constructed (sic!) sets of documents and assessment of relevance. In many real-world scenarios, it's therefore difficult to know the total number of relevant documents, complicating the calculation of Recall. It is usually more helpful to understand Precision and Recall as *flavors of optimization* and general *guideposts* on what to achieve overall, namely correctness and completeness.

Further Reading

Baeza-Yates, R., & Ribeiro-Neto, B. (2011). *Modern information retrieval: The concepts and technology behind search* (2nd ed.). Addison-Wesley Publishing Company.

Cleverdon, C. W. (1972). On the inverse relationship of Recall and Precision. *Journal of Documentation, 28*(3), 195–201. https://doi.org/10.1108/eb026538

Manning, C. D., Raghavan, P., & Schütze, H. (2009). *Introduction to information retrieval*. Cambridge University Press. https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf

Powers, D. M. W. (2008). Evaluation: From precision, recall and F-Measure to ROC, informedness, markedness and correlation. *Mach. Learn. Technol., 2*. https://doi.org/10.48550/ARXIV.2010.16061

Evaluate

Composite Indicators in Bibliometrics

What is it?

In the field of bibliometrics, composite indicators (CIs) are sophisticated metrics formulated by integrating multiple individual indicators into a single, overarching index. In this sense CIs resemble other statistical concepts found in general statistical applications, e.g. additive indices. Usually composite indicators are designed to encapsulate a broader and more nuanced understanding of a complex phenomenon, e.g research performance, impact, or productivity aiming to integrate multiple facets of a phenomenon. They typically combine diverse bibliometric dimensions such as citation counts, publication volumes, journal impact factors, and collaboration metrics. In other cases bibliometric metrics can be part of composite indicators. The overall aim of CIs is to synthesize these varied elements to offer a holistic measure that captures the complex nature of scholarly activities.

Why is it important?

The significance of composite indicators lies in their ability to provide a holistic evaluation of a phenomenon. They overcome the limitations of singular bibliometric measures by offering a balanced perspective that encompasses various aspects of scholarly work. This comprehensive approach is particularly attractive for policy-makers, institutional administrators, research managers etc. as aid in informed decision-making, strategic planning, and resource allocation as CIs linearize a multi-dimensional concept into a singular metric. It is something that both has the flair of *holisticity* and *multidimensionality* and at the same time features the convenience of being able to sort it (most of the time descending). Furthermore, these indicators are instrumental in facilitating comparative analysis across researchers, institutions, or countries, allowing for a more equitable assessment by normalizing diverse bibliometric measures into a unified framework. One example of a composite indicator in bibliometrics is the Hirsch-Index, even though its computation is rather unique and it does subsume composite indicators more in thought than as actual reference to a type of metric. Another example for a composite indicator are the different altmetric indicators combining different outlets of (social) media into a single metric.

How does it work?

The development and application of composite indicators in bibliometrics involve a multi-step process that is intricately documented in the *OECD Handbook on Constructing Composite Indicators*. In a very simplified way the construction follows steps.

Initially, relevant bibliometric indicators that reflect different facets of research performance, or any other phenomenon for that matter, are selected. These indicators are then polarity-harmonized (same direction), normalized to ensure comparability, typically by converting them to a common scale using a measure (e.g. Z-Transformation, rescaling etc...) that limits the range to a predefined scale of possible values or at least aims to produce a more uniform distribution. The next steps involve the careful weighting and aggregation of these indicators, with each component assigned a specific weight based on its perceived importance in the overall assessment. Producing such weights can be very challenging and depends in part on the outcome of the previous steps. In principle it is an assessment of contribution to the overall phenomenon to be measured, or, in more simple words, answering questions such as "How many Tweets are worth a feature in a national newspaper?". Sometimes weighting is achieved by transforming all the contributing metrics into a single *currency*, e.g. monetary values, even though such a process is, at the very best, tedious and complex. The weighted aggregation then results in the composite index, which then should be subjected to rigorous validation to ensure its reliability and validity both quantitatively, e.g. by integrating it into inferential models, but also semiotically by aiming to place it within the canon of previously established indicators.

Limitations

Designing composite indicators involve careful evidence and processing

The construction of these indicators is a complex task that requires meticulous consideration in selecting components and determining their respective weights. This process can introduce subjectivity, particularly in the weighting stage, potentially skewing the final outcome.

Compressing is not unfolding

Over-reliance on a single composite score may also lead to oversimplification, failing to capture the nuanced and multifaceted nature of a phenomenon fully. Even though composite indicators are argued to produce a holistic view of a subject, they in the end linearize the mutli-dimensionality into a single vector of values. In this regard, there are two fractions in bibliometrics. There are those that embrace composite indicators for their merits and argue that at the very least the overall values reflect a multi-dimensional phenomenon and still provide the potential for rankings. Opposed to this perspective are those that argue that such compression may help to produce simple rankings but provide no guidelines for improvement as changes in the CIs are not reflected by

underlying changes of the corresponding dimensions. At the very least these changes are over-shadowed by the methodology. The proponents of this position favor capturing multidimensionality through visualization techniques, e.g. by using radar or spider plots, sacrificing the benefits of simple ranking.

Garbage in - composited garbage out

Accuracy and reliability of composite indicators are heavily dependent on the quality and availability of the underlying data.

Further Reading

El Gibari, S., Gómez, T., & Ruiz, F. (2022). Combining reference point based composite indicators with data envelopment analysis: Application to the assessment of universities. *Scientometrics, 127*(8), 4363–4395. https://doi.org/10.1007/s11192-022-04436-0

Johnes, J. (2018). University rankings: What do they really show? *Scientometrics, 115*(1), 585–606. https://doi.org/10.1007/s11192-018-2666-1

Makkonen, T., & Van Der Have, R. P. (2013). Benchmarking regional innovative performance: Composite measures and direct innovation counts. *Scientometrics, 94*(1), 247–262. https://doi.org/10.1007/s11192-012-0753-2

Moon, H. S., & Lee, J. D. (2005). A fuzzy set theory approach to national composite S&T indices. *Scientometrics, 64*(1), 67–83. https://doi.org/10.1007/s11192-005-0238-7

Nasir, A., Ali, T. M., Shahdin, S., & Rahman, T. U. (2011). Technology achievement index 2009: Ranking and comparative study of nations. *Scientometrics, 87*(1), 41–62. https://doi.org/10.1007/s11192-010-0285-6

OECD/European Union/EC-JRC. (2008). *Handbook on constructing composite indicators: Methodology and user guide*. Paris: OECD Publishing. https://doi.org/10.1787/9789264043466-en

Vinkler, P. (2006). Composite scientometric indicators for evaluating publications of research institutes. *Scientometrics, 68*(3), 629–642. https://doi.org/10.1007/s11192-006-0123-z

Field Normalization

What is Field Normalization?

Field normalization is a process used in bibliometrics to adjust for variations across different scientific fields when comparing bibliometric indicators, such as citation counts. This is necessary because different scientific fields have distinct publication and citation practices.

Why is it important?

Imagine comparing a biology paper and a mathematics paper based solely on their citation counts. This might be misleading because biology papers, in general, tend to receive more citations due to the nature of the field. Field normalization aims to level the playing field, allowing for a fairer comparison.

How does it work?

Determining expected values: Normalized indicators operate on a notion of the *average*, the *usual*, the *expected* etc... In simple terms: What we can use to assess if something is *a lot*. In order to normalize indicators, we need to calculate a baseline for a specific field. This usually involves calculating the average number of citations for papers in that field, published in a particular year. The median, despite being more robust toward the outliers and right-skewed distributions we find in citation counts, is often not a good measure as the median often does not vary. More often than not, the median might simply be 0. Other percentiles such as the Top 10% or Top 1%, i,e, the 90th-percentile or 99th-percentile, are used in another context.

Accounting for the expected values: In the next step, the citation count of an individual paper is compared to this baseline. We achieve this by simply dividing the observed value by the expected value calculated in step 1. For example, if the average citation count in field X is 10 and a paper in field X has 20 citations, it is performing above the average for its field. Most normalized indicators will therefore be centered around 1 as the *neutral* outcome of normalization. The indicator designer might choose to center an indicator at 0. In this case usually a) 1 is subtracted from the ratio or b) a logarithm is applied to the ratio. In the latter case, there is also an element of diminishing returns built into the indicator.

Limitations

Not sensitive enough

Some critics of this approach argue that normalized indicators as they are described above are not useful as they do not vary over time.

The distribution of citations is too skewed

The arithmetic mean is sensitive to outliers. In the case of the distribution of citation counts this is usually skewed to the right, i.e. there are a lot of papers rarely being cited and a few that are cited disproportionately often. Critics argue that this makes the mean a bad measure in such cases. They propose to use different measures such as the so-called *Excellence rate* or ppTOPX (ppTop10, ppTop5...) as it a) focuses on the politically relevant issue of *excellence* and ignoring *noise* and b) is not prone to distortion by the mean.

(Often) too reliant on classifications

This critique is mostly oriented toward field- or category-oriented normalization. It does not account for normalization based on average citation counts of journals such as the JCR (Journal Citation Rate). The main point is that classifications may mis-represent what is happening in science and underestimate the impact of inter- and transdisciplinary research or at the very least over-represent what is happening in *core disciplines*. Among other limitations the Source Normalized Citation Impact (SNIP) by Moed & Colleagues addresses the problem of classification by constructing the *field* as a set of journals citing a journal in question, thereby contextualizing the journal by its reception.

Practices of referencing can be biased against new entrants

Some more fundamental criticism is put forward by critics relating to the Matthew Effect. In a nutshell, the Matthew effect describes the phenomenon that prestigious authors receive a premium on their reputation, i.e. they get more citations simply because they are more well known and have a higher status or established reputation. This argument usually calls into question the validity of evaluatory citation analysis in general and is an issue of theories of citation.

Citations reflect simplified notions of impact.

A similar critique to the aforementioned argument calls into question the validity of citations as measures of quality. This critique is oriented toward the notion that impact should not be reduced to what is often referred to as *simplified counting*. At best, multiple indicators should be used.

Bibliometrics and Scientometrics have no theory of scientific fields

Some critics argue that the scientometrics community has no convincing theory of what delineates topics from scientific fields. Hence, normalization efforts reflect mere categorization efforts of journals or scientific articles that reflect rather artificial boundaries and neither account for emerging scientific fields nor for phenomena of interrelatedness such as inter- and transdisciplinarity.

Further Reading

Bornmann, L., Haunschild, R., & Mutz, R. (2020). Should citations be field-normalized in evaluative bibliometrics? An empirical analysis based on propensity score matching. *Journal of Informetrics, 14*(4), 101098. https://doi.org/10.1016/j.joi.2020.101098

Bornmann, L. (2020). How can citation impact in bibliometrics be normalized? A new approach combining citing-side normalization and citation percentiles. *Quantitative Science Studies, 1*(4), 1553–1569. https://doi.org/10.1162/qss_a_00089

Heintz, B. (2010). Numerische Differenz. Überlegungen zu einer Soziologie des (quantitativen) Vergleichs / Numerical difference. Toward a sociology of (quantitative) comparisons. *Zeitschrift für Soziologie, 39*(3), 162–181. https://doi.org/10.1515/zfsoz-2010-0301

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for Research Metrics. *Nature, 520*(7548), 429–431. https://doi.org/10.1038/520429a

Merton, R. K. (1968). The Matthew effect in science: The reward and communication systems of science are considered. *Science, 159*(3810), 56–63. https://doi.org/10.1126/science.159.3810.56

Moed, H. F. (2010). Measuring contextual citation impact of scientific journals. *Journal of Informetrics, 4*(3), 265–277. https://doi.org/10.1016/j.joi.2010.01.002

Fractional Counting

What is Fractional Counting?

Fractional counting relates to one of the major branches of evaluative bibliometrics, i.e. the evaluation of productivity or output volume. It is a method used to attribute credit for scholarly publications, particularly those with multiple authors. The overall goal is to account for a notion of distributive fairness matching efforts.

Why is it important?

In many fields, collaborative research leading to multi-author papers is common. If the assessment of productivity is aimed to be relative to the efforts involved, doubts about the validity of simply summing up the number of contributions may emerge. Shouldn't it make a difference how many authors are involved then? This is where fractional counting comes in. The basic idea of fractional counting is to account for differences in distribution of labor, i.e. accounting for authors sharing the work when compiling a scientific contribution. Assigning full credit to each author can inflate the apparent productivity and impact of individuals or institutions involved. Unlike full counting methods, where each author or institution gets full credit for a publication, fractional counting divides the credit among the contributors based on their number.

How Does It Work?

Credit Allocation on different levels

In a paper with multiple authors, instead of each author receiving one full count for the publication, the credit is fractionally divided. For example, if a paper has four authors, each author might receive 0.25 (1/4th) of a count for that paper. Yet, the level of allocation as well as the level of analysis are important. Depending on the notion of what *contributes* to the paper the level of fractioning might vary. Despite author-level fractioning, other levels may be used, e.g. country level, organization level, address level, i.e. address of the institution, which might be relevant for institutions with multiple sites or locations, etc... Specific forms may also involve pooling the effort to one author, e.g. the principal investigator, last author position in medicine, or the first author representing the *main author*. Yet, these follow a completely different understanding of fairness.

Impact on Metrics

Fractional counting affects various bibliometric indicators. This impact is strictly not limited to analyses of publication volume and productivity but also influences the way citation analyses are being conducted.

Variations in Practice

The exact method and data source of fractional counting can vary. Comparing data from different sources, e.g. WoS, Socpus, OpenAlex or Google Scholar, as well as different levels of fractioning is strictly not advised, let alone, merging them into one unified dataset for analysis.

Limitations

Simple fractioning is not accounting for divisions of labor

Fractioning rests on the rather plausible idea that productivity is relative to the number of authors involved. Yet, dividing by the number of entities might not represent the actual efforts each author or organization has invested in the paper. Recent developments in the publication landscape try to account for this by introducing contribution sections. In these sections, the authors state in what parts of the research and publication process they have been involved in.

Fractional counting can put a penalty on cooperation

Using fractional counting introduces a penalty on cooperation. This might lead to misrepresentation of outputs. Imagine an organization that a couple of years ago made it their mission to cooperate more with other organizations. If that mission is being put to work, it is rather plausible that the number of co-authored papers with authors from other organizations will increase. This could lead to a situation where full-counts, i.e. whole counting of papers, increase but the number of fractional papers level out as more authors are involved due to cooperation activity. In extreme cases the fractional count could be declining while full-count increases. Always use full counts as a litmus test to understand your fractional counts.

Further Reading

Aksnes, D. W., Schneider, J. W., & Gunnarsson, M. (2012). Ranking national research systems by citation indicators. A comparative analysis using whole and fractionalised counting methods. *Journal of Informetrics, 6*(1), 36–43. https://doi.org/10.1016/j.joi.2011.08.002

Cronin, B. (2001). Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices? *Journal of the American Society for Information Science and Technology, 52*(7), 558–69. https://doi.org/10.1002/asi.1097

Donner, P. (2024). Remarks on modified fractional counting. *Journal of Informetrics, 18*(4), 101585. https://doi.org/10.1016/j.joi.2024.101585

Gauffriau, M., Larsen, P. O., Maye, I., Roulin-Perriard, A., & Von Ins, M. (2008). Comparisons of results of publication counting using different methods. *Scientometrics, 77*(1), 147–76. https://doi.org/10.1007/s11192-007-1934-2

Leydesdorff, L., & Opthof, T. (2010). Normalization at the field level: Fractional counting of citations. https://doi.org/10.48550/ARXIV.1006.2896

Sivertsen, G., Rousseau, R., & Zhang, L. (2019). Measuring scientific contributions with modified fractional counting. *Journal of Informetrics, 13*(2), 679–94. https://doi.org/10.1016/j.joi.2019.03.010

Waltman, L., & Van Eck, N. J. (2015). Field-normalized citation impact indicators and the choice of an appropriate counting method. *Journal of Informetrics, 9*(4), 872–94. https://doi.org/10.1016/j.joi.2015.08.001

Hirsch-Index (h-index)

What is the Hirsch-Index?

The Hirsch-Index, or h-index, is a metric developed by physicist Jorge E. Hirsch in 2005 to quantify the combined impact and productivity of a scholar's published work. In contrast to other evaluative metrics focusing either on productivity or reception of scientific publications by balancing the quantity (number of publications) and quality (citations) of scientific output seeking to combine both perspectives giving an overall notion of the *eminence* of a scholar or research organization. The h-index is defined as the number of papers (h) that have been cited at least h times. For instance, if a researcher has an h-index of 10, it means they have published 10 papers that have each been cited at least 10 times. The set of publications for which these conditions are true, i.e. they definingly contribute to the h-Index due to their received number of citations, is sometimes labeled as the *h-core* and is sometimes analyzed separately.

Why is it Important?

The h-index is important because it provides a single figure that can suggest the significance and broad impact of a researcher's cumulative research contributions. In that respect it is a rather small scale version of a composite indicator. It's widely used in academia for evaluating the influence of a researcher's work, considering both the quantity and quality of their publications. It's also employed by funding agencies, universities, and research institutions for hiring, promotion, and funding decisions. Its composition makes it resistant to outliers, as e.g. one or two highly cited papers, or a large number of scarcely cited papers, have little effect on the Hirsch-Index. What counts is continuous performance above a certain level. This also means that with increase of the Hirsch-Index raising the index by 1 involves increasingly higher efforts if trying to *optimize* it. Finally, it is exceedingly easy to compute, making it a measure that can be performed with considerably little data as no field normalization is involved. Another benefit of the Hirsch-Index is its scalability to higher levels of aggregation making it a measure that is also applicable on the level of organizations. Despite its prominence and widespread application it is also an object of strong criticism in the current evaluation landscape.

How Does it Work?

Calculating the h-index involves the following steps:

1. Listing Publications and Citations: Compile a list of the researcher's publications, along with the number of citations each paper has received.
2. Ranking by Citation Count: Rank the publications according to the number of citations.
3. Find the h-index threshold: Find the point where the highest rank of the paper in the list is equal to or greater than the number of citations it has received. This number is the h-index.

Limitations

The Hirsch-Index favors senior researchers

The Hirsch-Index is only meaningfully interpretable when used to understand differences within the group of established researchers for the simple fact that the output is the first limiting factor for the Hirsch-Index. It needs a minimum of X papers to have a Hirsch-Index of X. Junior and early career researchers will be largely discriminated against as they may only have a few papers.

The Hirsch-Index is not helpful to compare researchers from different fields

The strength of the Hirsch-Index, namely its integration of output and reception and the fact it does not, per definition at least, require field-normalization, also produces a weakness making it not universally applicable across different academic disciplines due to variations in publication and citation practices.

The Hirsch-Index can be a metric garbage compactor

As the Hirsch-Index is neither field nor journal normalized it produces another issue as it ignores almost all context of citations. For instance, it does not distinguish between citations from high-quality sources and less reputable ones. In theory, a high Hirsch-Index can be achieved by a combination of high levels of self-citation and strategic publication in low impact journals. Or any other means of gaming citation counts for that matter.

The Hirsch-Index features strong fluctuations between databases and fields

The Hirsch-Index can vary enormously depending on the database used with the impact of database selection depending on the field, or, more precisely the overall publication culture prevalent in a field and the tendency to publish pre-prints. Usually, differences between Web of Science (WoS) and Scopus are rather small with large differences being mostly between the aforementioned and Google Scholar with strong pre-publishing fields such as computer science receiving very high premiums in Google Scholar.

Further Reading

Bar-Ilan, J. (2008). Which h-index? — A comparison of WoS, Scopus and Google Scholar. *Scientometrics, 74*(2), 257–71. https://doi.org/10.1007/s11192-008-0216-y

Bornmann, L., & Daniel, H.-D. (2007). What do we know about the h index? *Journal of the American Society for Information Science and Technology, 58*(9), 1381–85. https://doi.org/10.1002/asi.20609

Egghe, L. (2010). The Hirsch index and related impact measures. *Annual Review of Information Science and Technology, 44*(1), 65–114. https://doi.org/10.1002/aris.2010.1440440109

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences, 102*(46), 16569–72. https://doi.org/10.1073/pnas.0507655102

Van Raan, A. F. J. (2006). Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. *Scientometrics, 67*(3), 491–502. https://doi.org/10.1556/Scient.67.2006.3.10

Percentile-Based Indicators (Excellence Rate, ppTopX)

What is it?

In bibliometrics, percentile-based indicators are statistical measures used to evaluate and compare academic publications. These indicators place publications within the context of a larger set of data, typically by showing where a particular work falls in the distribution of all works in the same field or subject area. The ppTop10 (percentage of publications in the top 10%) is a specific type of percentile-based indicator. It measures the proportion of an entity's (such as an author's, institution's, or country's) publications that are in the top 10% of the most cited papers in their respective fields. This is determined by comparing the citation count of a paper to that of other papers in the same field and year.

Why is it important?

Percentile-based indicators follow two main arguments. The first refers to the issue of log-normality in the citation distribution, i.e. a lot of uncited and rarely cited papers and a very low number of very highly cited papers leading to a right-skewed distribution, and the discussion if the arithmetic mean (average) of the number of citations can provide a useful value in the light of these highly cited outliers. The usual reply to such problematisations is to use the median as a central value. In citation distributions the median, by nature being an integer value rather than a metric number, is often ranging between 0 and 2 (on a good day), which makes it a very bad metric to capture expectations of *the usual number of citations in a field*. While the median is rarely used, the percentile being used is rather the top 10%, 5% or 1%. Which leads to the second argument that also highlights its colloquial name: The excellence indicator. The argument in this case is that using the average as a benchmark, e.g. in field normalization, is misleading and does not provide substantial and good advice to policy makers. Rather, focus should be put on the upper parts of the citation distribution denoting notions of *excellence* rendering everything else as rather irrelevant *noise*. The share of a unit of observation in these excellent publications relative to the total amount produced captures the notion of: "How much excellence is there relative to the total output?"

How It Works

Step 0: Field and Year Normalization

Each publication is compared to others within the same field and publication year. This ensures that the comparison is fair and accounts for variations in citation practices across different disciplines and over time. This is crucial as it controls for publication and citation practices between the different fields.

Step 1: Ranking the total corpus by citation count

ALL publications, or more precisely, all citable items, in a field for a specific publication year in a given database are ranked based on their citation count.

Step 2: Identifying Top 10%

The top 10% of publications in terms of citation count are identified.

Step 3: Calculating ppTop10

The percentage of an entity's publications that fall into this top 10% category is calculated. For example, if an institution has 100 papers and 15 of them are in the top 10% most cited in their fields, the institution's ppTop10 score would be 15%.

A higher ppTop10 score suggests that a significant portion of an entity's work is highly recognized in its field.

Limitations

Excellence rates shift observation and discussion

The strength of ppTop10 also contributes to its major weakness as it pushed discursive positions toward over-indexing on excellence as the only meaningful category. This might be useful in some discussions but shifts the discussion away from discussing emerging fields, topics or delayed reception. Overall, while ppTop10 indicates high citation levels, it doesn't account for the overall impact or quality of the research. Some important works may not be highly cited.

Further Reading

Bornmann, L., & Mutz, R. (2011). Further steps towards an ideal method of measuring citation performance: The avoidance of citation (ratio) averages in field-normalization. *Journal of Informetrics, 5*(1), 228–230. https://doi.org/10.1016/j.joi.2010.10.009

Bornmann, L., & Williams, R. (2020). An evaluation of percentile measures of citation impact, and a proposal for making them better. *Scientometrics, 124*(2), 1457–1478. https://doi.org/10.1007/s11192-020-03512-7

Pudovkin, A. I., & Garfield, E. (2009). Percentile rank and author superiority indexes for evaluating individual journal articles and the author's overall citation performance. *Collnet Journal of Scientometrics and Information Management, 3*(2), 3–10. https://doi.org/10.1080/09737766.2009.10700871

Tijssen, R. J. W., Visser, M. S., & Van Leeuwen, T. N. (2002). Benchmarking international scientific excellence: Are highly cited research papers an appropriate frame of reference? *Scientometrics, 54*(3), 381–397. https://doi.org/10.1023/A:1016082432660

Van Leeuwen, T. N. (2003). The Holy Grail of science policy: Exploring and combining bibliometric tools in search of scientific excellence. *Scientometrics, 57*(2), 257–280. https://doi.org/10.1023/A:1024141819302

Waltman, L., & Schreiber, M. (2013). On the calculation of percentile‐based bibliometric indicators. *Journal of the American Society for Information Science and Technology, 64*(2), 372–379. https://doi.org/10.1002/asi.22775

AoR (Averages of Ratios) vs. RoA (Ratios of Averages) in Bibliometrics

What is it?

Averages (or Means) of Ratios (AoR) and Ratios of Averages (or Means) (RoA) are two distinct statistical approaches used in bibliometric analysis and other fields where comparative analysis of data sets is required. The issue is most prevalent in evaluatory bibliometrics when using average-based indicators, i.e. indicators that somehow aim to normalize output or citations against something to be considered *the usual*, i.e. an average. In simple terms, the AoR approach involves first calculating the ratio for individual pairs (!) of data points across two data sets, and then finding the average (mean) of these ratios. This gives equal weight to each data point, focusing on the individual level of comparison before aggregating. In RoA, by contrast, this approach calculates the average (mean) for each data set (!) separately and then takes the ratio of these two means. This method emphasizes the collective characteristics of each data set, treating the data sets as aggregated wholes from the onset.

Why is it important?

The choice between these two methods can significantly influence the outcome and interpretation of bibliometric analyses, particularly when comparing groups or categories. Both methods reflect different aspects of the data and what to learn from it. AoR can highlight individual-level differences or similarities, while RoA provides a comparison at the aggregate level. The selection of the appropriate method therefore depends on the research question and the nature of the data, making it crucial to understand both approaches for accurate and meaningful analysis. AoR typically results in higher scores than RoA, particularly for entities with fewer papers or lower impact scores. RoA tends to underestimate the impact, especially for departments, institutions, and countries with lower scores. RoA eventually leads to inconsistencies in global averages, which should ideally sum to unity but don't always do so in practice. The choice between AoR and RoA impacts the evaluation of research performance, with AoR being generally more favorable for entities with fewer publications.

How does it work?

AoR:

1. Calculate the ratio for each corresponding pair of data points across the two sets.
2. Sum up all these individual ratios.
3. Divide the total by the number of ratios to get the mean.

RoA:

1. Calculate the mean for each data set separately.
2. Divide the mean of one data set by the mean of the other to get the ratio.

Limitations

AoR:

- Can be heavily influenced by outliers or extreme ratios.
- May not accurately represent the overall characteristics of the data sets if there's significant variability within them.

RoA:

- Can mask individual data point variations, leading to oversimplified conclusions.
- May not be appropriate in cases where data sets have widely differing distributions or sizes.

Further Reading

Egghe, L. (2012). Averages of ratios compared to ratios of averages: Mathematical results. *Journal of Informetrics, 6*(2), 307–317. https://doi.org/10.1016/j.joi.2011.12.007

Egghe, L., & Rousseau, R. (1996). Average and global impact of a set of journals. *Scientometrics, 36*(1), 97–107. https://doi.org/10.1007/BF02126648

Larivière, V., & Gingras, Y. (2011). Averages of ratios vs. ratios of averages: An empirical analysis of four levels of aggregation. *Journal of Informetrics, 5*(3), 392–399. https://doi.org/10.1016/j.joi.2011.02.001

Vinkler, P. (1996). Model for quantitative selection of relative scientometric impact indicators. *Scientometrics, 36*(2), 223–236. https://doi.org/10.1007/BF02017315

Explore

Weightedness and Directedness of a Graph in Bibliometrics

What is it?

In bibliometrics, and more broadly in network analysis, a graph is a visual representation of relationships between different entities (like authors, publications, or institutions). The *weightedness* and *directedness* of a graph are two important characteristics. Weightedness refers to whether the edges (connections) in the graph have weights that signify the strength or intensity of the connection. For example, in a co-authorship network, the weight could represent the number of papers co-authored by two researchers. Directedness indicates whether the edges have a direction, representing asymmetric relationships. In a citation network, for example, an edge from paper A to paper B would signify that A cites B, not the other way around.

Why is it Important?

Weightedness and directedness are two characteristics that inform what can be done analysis-wise with network data. For instance, these characteristics determine what type of community detection algorithm may be useful and which may produce meaningful results. Directedness refers to the presence of directed edges in a graph, indicating the directionality of flows, e.g. of information, between nodes. In bibliometrics or altmetrics, directedness can be an important factor when identifying communities. For example, in a directed social network where interactions are one-way (e.g., following a user on Twitter), community detection algorithms may prioritize densely connected groups of users who receive and share information more frequently with each other. Weightedness refers to the presence of edges with different weights in a graph. In the context of community detection, weightedness can be used to capture the significance or strength of connections between nodes within a community. For instance, in bibliometrics or altmetrics, where some users have more followers or higher levels of influence than others, algorithms may assign higher weights to these connections and prioritize communities that are densely connected with influential individuals. Similarly, the analysis of co-authorship structures on author or institutional level will benefit from considering the weight (i.e. the number of co-published papers) rather than just relying on information if co-authorship occurred in a binary fashion, i.e. yes or no. The relevance of directedness and weightedness in graph analysis becomes evident when considering different algorithms for community detection. For example, the Louvain method is a popular community

detection algorithm that incorporates both relatedness, via centrality measures, and weighted-ness, by using the weights to edges based on their impact or influence. This approach allows for the identification of communities with not only high relatedness but also strong ties to influ-ential nodes within those communities. In a similar vein, the Leiden algorithm, a rather elegant extension of the Louvain method, works the same way in this respect. Another community detec-tion algorithm, the Girvan-Newman algorithm, leverages edge removal to identify communities based on maximizing betweenness centrality (or eigenvector centrality for that matter). By consid-ering both directedness and weightedness, this algorithm can detect densely connected groups of nodes while also accounting for the significance of connections within those communities.

Overall, understanding weightedness and directedness is important in exploratory bibliometric analysis for several reasons. First of all, the characteristics allow for a more nuanced and accu-rate representation of the relationships in a scholarly network and thereby also a more informed interpretation. Again, it does make a difference if a network is directed or not, e.g. if a paper ref-erences another paper or is cited by it. It does make a difference if a network is weighted, i.e. if a co-publication between two organizations is framed as simply existing or not or if a strength (e.g. the Salton Cosine of co-publications) is being observed. In short, honoring these charac-teristics can enable researchers to conduct more detailed and specific analyses, like identifying influential authors (in directed graphs) or strong collaboration networks (in weighted graphs). Fur-thermore, directed and weighted graphs provide deeper insights into the nature and intensity of scholarly interactions, which can be crucial for understanding scientific communication and collab-oration patterns. Weighted networks sometimes offer clear and informative visual representations of complex relationships by integrating the strength of ties as width of the connecting edges. Fi-nally, positionality algorithms will benefit from weighted graphs as this helps in integrating weights into the calculator, thereby putting strongly related nodes closer together than less strongly related nodes.

How Does it Work?

Usually the two graph characteristics can be deduced from the way the network data has been col-lected and constructed. When using data from a third party it is always advisable to double check if the assumption being made, e.g. an author collaboration matrix being undirected-weighted, is actually true. In the case of a collaboration network based on authors researchers may make the decision to give the relationship between first author (or last author *PI* position depending on the field) and all the other authors a direction to, for instance, signify some sort of managerial role. The way such networks are constructed thereby follows a) the potentiality of what weightedness and directedness can actually be achieved based on the underlying data, and b) the research question in mind and the rationale behind the construction process. Some characteristics can be checked. If weights greater 1 are present in a dataset, then it is not unweighted. If only weights of 0 or 1 exist this does NOT imply that the graph is weighted. It simply can only have relationships of 0 and 1. For all intents and purposes it CAN be treated as unweighted, though. When cell-wise subtracting the matrix representing a graph with its transposed form (rows and columns are switched, i.e. the

matrix is tipped to the side, and the result for all cells is NOT 0, then the graph should be directed, assuming no mistake has been made constructing it. If all cells feature 0, then the graph CAN be undirected. Yet, in the case of directedness the graph should still be treated as directed.

Limitations

Always check, never assume

The thing with these characteristics is that sometimes developers of an algorithm implementation assume that the user knows what they are doing and take some measures to optimize. For instance, developers might choose to only use the upper or lower triangle of a matrix representing a graph as they hold the same information in an undirected graph anyway and because, assuming N number of nodes, doing $N * \dfrac{(N-1)}{2}$ calculations is less than doing $N * N - N$ calculations. In short: An implementation can produce results even IF the data does not follow the requirements of the method.

Community Detection Algorithms in Bibliometrics

What is it?

Community detection algorithms in bibliometrics are computational methods used to identify clusters or groups within scientific networks. These networks can be based on various relationships such as co-authorship on author, organization, regional or national level, citations, co-citations, shared keywords among scientific papers etc... Said clustering can reveal hidden patterns, subfields, or research trends within a larger scientific domain.

Why is it Important?

Visual inspection of networks above a certain size is challenging at best, especially if networks are unweighted, i.e. like with the current way we use citation data, which is largely showing if a paper cites another paper or not and does not show *how often* a paper mentions another. Visualizing such networks often leads to the so-called *hairball problem*. Finding structures in such networks visually(!) seems an almost impossible task with dense networks. Also relying too much on attributes of nodes (degree, betweenness...) can usually be instructive when it comes to identifying *influential nodes* (authors, organizations), but is of very little use when the question relates to identifying structures. To achieve this, one way to move forward is using algorithms that identify *clusters* or *agglomerations*, in the loosest sense of the word, based mostly on either using matrix algebra, discrete mathematics, or graph theory. Common algorithms include modularity optimization methods, hierarchical clustering, and spectral clustering. Such algorithms come by a plethora of names: Structure-detection algorithms, community-detection algorithms, clustering algorithms etc... The common denominator is that these algorithms aim to uncover the structure of these networks by grouping nodes (again... such as authors, papers, or journals) into communities or clusters that are more densely connected with each other than with the rest of the network. By doing that, they can facilitate the analysis of the development, interaction, and interdisciplinary nature of various scientific fields or help identify emerging research areas and trends. As a strategic tool, identifying such *clusters* can guide organization and funding bodies in making informed decisions about resource allocation, based on the identification of key research areas and collaboration networks. Finally, using such algorithms can guide formative evaluation as it can continuously assist researchers in identifying potential collaborators or relevant research groups.

How Does it Work?

Implementing community detection algorithms involves a number of steps and there are certain flavors how these can be used in detail. Yet, quite usually, the following steps have to occur.

Step 1: Network Construction

Building a network from bibliometric data, where nodes represent entities like authors or papers, and edges represent relationships like co-authorship or citations.

Step 2: Deduce the type of network and select algorithm

Not all algorithms make sense for all kinds of networks. Select an appropriate algorithm based on the graph being directed or undirected as well as the graph being weighted or unweighted. Some algorithms require a fully connected graph not featuring multiple *components*, i.e. any node can be reached by any other node, or in other words, the graph has no *un-connected subgraphs*.

Step 3: Run analysis

Applying community detection algorithms to the network.

Step 4 Determine stable number of clusters/groups

This usually involves an algorithm that optimizes towards a certain measure maximizing edge characteristics within clusters, such as average degree of nodes in a cluster, number of edges within a cluster etc... and minimizing this characteristic between clusters. There are a good dozen of means to do this (Elbow method, gap statistics, the Calinski–Harabasz index, the Davies–Bouldin index...)

Step 5 Labeling the clusters, Analysis and Interpretation

Analyzing the resulting communities to draw insights about the underlying structure of the scientific network.

Limitations

Algorithm Complexity

Some community detection algorithms can be complex and computationally intensive depending largely on the size and density of the network. The more sparse and the smaller a network, the faster computation can go. Also, some algorithms of this sort can be *NP-hard*, i.e. solvable in Nondeterministic Polynomial time leading to an exponential increase in computation relative to increasing the extent (number of nodes and edges) of the problem at hand making them *tricky* to solve quickly and easily. (To provide an understanding of what NP-hard means in practice: Problems being NP-hard is why most non-trivial cryptography *works*.)

Community detection is no silver bullet for interpretation

The results can sometimes be challenging to interpret, especially in highly interconnected or multidisciplinary fields or situations of complex interconnectedness, for instance, when the connectedness is a phenomenon in and of itself like in the case of interdisciplinarity.

Structure Detection algorithms detect structures

That's what they do. This does not mean that different algorithms produce the same conclusions or that one conclusion has more *validity* than another. There are some common metrics that resemble quality measures from inferential statistics, such as modularity, but they only are interpretable within very limited confines. Furthermore, humans are excellent pattern recognition machines, leading to usually a stance of expressing their own ideas about structures onto the results, which in turn usually leads more to a *see what you already know* situation.

Dependence on Data Quality

The effectiveness of community detection is heavily reliant on the quality and completeness of the bibliometric data.

Further Reading

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 2008*(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Clauset, A., Newman, M. E. J., & Moore, C. (2004). Finding community structure in very large networks. *Physical Review E, 70*(6), 066111. https://doi.org/10.1103/PhysRevE.70.066111

Luke, D. A. (2015). *A user's guide to network analysis in R*. Use R! Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-23883-8

Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E, 69*(2), 026113. https://doi.org/10.1103/PhysRevE.69.026113

Traag, V. A., Waltman, L., & Van Eck, N. J. (2019). From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports, 9*(1), 5233. https://doi.org/10.1038/s41598-019-41695-z

Traag, V. A., & Šubelj, L. (2023). Large network community detection by fast label propagation. *Scientific Reports, 13*(1), 2701. https://doi.org/10.1038/s41598-023-29610-z

Bibliographic Coupling

What is Bibliographic Coupling?

Bibliographic coupling (BC) occurs when two scientific papers reference the same third paper in their bibliographies. In simple terms, bibliographic coupling is the number of shared references for a pair of papers. The concept was introduced by M.M. Kessler in 1963 and has developed into a key tool in bibliometrics for analyzing the structure and development of scientific research, especially in terms of shared intellectual heritage contributing to scientific papers suggesting a direct relationship in subject matter or methodology.

Why is it Important?

Bibliographic coupling is important as it helps in identifying research trends and the development of scientific fields. Papers that share many references are likely to be closely related, offering insights into how researchers build upon existing knowledge. This analysis is particularly useful in identifying research clusters, understanding the evolution of topics, and detecting emerging areas of study. Also, sometimes it can help to limit down the size of large networks by using only such publications that feature a minimum bibliographic coupling of 1.

How Does it Work?

Computing bibliographic coupling is rather straightforward. Usually reference lists from a set of relevant scientific papers are compiled using a bibliographic database. The preferred method is to produce a citing paper x cited paper matrix (citation network). This network is an unweighted-directed graph. Unweighted directed graphs can be transformed into weighted-undirected graphs (the bibliographic coupling network) by matrix multiplication using the aforementioned matrix and performing a matrix multiplication with the transposed matrix. It should be noted that matrix multiplication, in contrast to its scalar relative, is sensitive to the order of multiplicand and multiplier. $A \cdot A^T$ is not the same as $A^T \cdot A$. When confusing the order the result is, interestingly, the co-citation network.

Limitations

Bibliographic coupling is over-indexing on citation behavior

Bibliographic coupling is limited to the analysis of reference lists. Hence, it only reflects the similarities based on the chosen references, which might miss broader connections such as language use.

Bibliographic coupling paints a static picture

As mentioned above, reference lists are static. In contrast to co-citation, which can change over time as new publications citing others are being produced, bibliographic coupling is static, i.e. it will not change as the reference lists have already been compiled prior to final publication. BC therefore offers a snapshot based on the time of publication, not accounting for subsequent developments.

Bibliographic Coupling is biased toward older contributions

More recent papers have less opportunity to be bibliographically coupled due to their shorter presence in the literature.

Bibliographic Coupling is dependent on referencing and citation culture

Different academic fields have different referencing practices, which can affect the outcome of the analysis.

Further Reading

Glänzel, W., & Czerwon, H. J. (1996). A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level. *Scientometrics, 37*(2), 195–221. https://doi.org/10.1007/BF02093621

Jarneving, B. (2005). A comparison of two bibliometric methods for mapping of the research front. *Scientometrics, 65*(2), 245–63. https://doi.org/10.1007/s11192-005-0270-7

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation, 14*(1), 10–25. https://doi.org/10.1002/asi.5090140103

Kessler, M. M. (1963). Bibliographic coupling extended in time: Ten case histories. *Information Storage and Retrieval, 1*(4), 169–87. https://doi.org/10.1016/0020-0271(63)90016-0

Weinberg, B. H. (1974). Bibliographic coupling: A review. *Information Storage and Retrieval, 10*(5–6), 189–96. https://doi.org/10.1016/0020-0271(74)90058-8

Zhao, D., & Strotmann, A. (2008). Evolution of research activities and intellectual influences in information science 1996–2005: Introducing author bibliographic-coupling analysis. *Journal of the American Society for Information Science and Technology, 59*(13), 2070–86. https://doi.org/10.1002/asi.20910

Co-Citation Analysis

What is co-citation analysis?

Co-citation analysis is a method used in bibliometrics to study the relationship between scientific publications and, in some of the derivatives, such as author-co-citation analysis (ACA), between other entities (nodes) that can be attributed to scientific publications. Two documents are co-cited, when they both receive (!) a citation from the same later (!) document. The strength of co-citation equals the number of citing documents for which this is the case, i.e. a co-citation of strength 3 means that 3 documents cite both scientific publications under analysis. This implies that an analysis of co-citation can be performed for each combination of citable items in a bibliographic dataset. Essentially, if two papers are often cited together, it implies a topical or methodological relationship between them that in turn implies intellectual heritage. This concept was introduced by Henry Small in the 1970s and is based on his notion that citations are *concept symbols* bringing together ideas from bibliometrics and semiotics, the latter being quite fashionable at that time. It shares a number of characteristics with bibliographic coupling (shared entries in reference lists between different papers), a concept that also relates to notions of *intellectual heritage*.

Why is co-citation analysis important?

Co-citation analysis is important because it helps to map the structure, evolution and dynamics of scientific fields as they can change over time, with new scientific publications being produced that can contribute to co-citation structures. By *being attributed together*, co-citation reveals how knowledge areas are interconnected and can identify emerging trends and, with sufficient data and computational power, even the emergence of (sub-)disciplines. The analysis is also crucial for understanding the influence of a particular work in a field, as frequent co-citations indicate a shared relevance or foundational and conceptual status among the cited works.

How Does it Work?

See Bibliographic Coupling but reverse the order of multiplicand and multiplier when performing matrix multiplication.

Limitations

Calculating Co-Citation networks can be computationally intensive

Citation networks can grow large very quickly, which is mainly due to the dominant nature of most phenomena in bibliometrics being *log-normal* distributed. The probably most efficient one-shot computational method to calculate co-citation is matrix multiplication. Yet, computing matrix multiplication is currently considered *P-hard*, which means that even though it actually can be solved in polynomial time, the computational time required for matrix multiplication grows rapidly, making it an expensive operation for large matrices, which in turn means that even if the number of papers to observe might be rather small, the dispersedness of citations can produce large matrices... and hence produce long computational times.

Calculating co-citation can require substantial computational resources

As stated above, co-citation can be data intensive, which means that, despite the challenge of computation it also faces the challenge of very high quality data. In short: While the point about computational time is a *CPU (or GPU) problem*, the issue here is a *RAM (memory) problem*. The problems are not the same. Parallelizing the problem to decrease computation time, might come to a grinding halt because RAM is obliterated by the parallel computing threads and the large amounts of data.

Co-Citation is less immediate compared to bibliographic coupling

In bibliographic coupling, older publications might be over-represented. In co-citation the opposite is true as there is a time-lag involved (see citation windows). Citation data may not reflect the most current trends due to publication and citation delays. Yet, citation windows usually are not being applied (in contrast to evaluative bibliometrics).

Co-Citation is dependent on citation practices

Variations in citation practices across fields can affect the results. It may therefore over- or under-represent some structures based in inter- or transdisciplinary fields.

Co-citation has a bias towards *disciplinarity*

Even though co-citation can be used to identify new topics and fields, there is a bias involved that relates to the point about citation practices. Highly interdisciplinary work might be underrepresented in co-citation networks if it doesn't fit neatly into established citation patterns.

Further Reading

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science, 24*(4), 265–69. https://doi.org/10.1002/asi.4630240406

White, H. D., & Griffith, B. C. (1981). Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for Information Science, 32*(3), 163–71. https://doi.org/10.1002/asi.4630320302

Bibliography

Aksnes, D. W., Schneider, J. W., & Gunnarsson, M. (2012). Ranking national research systems by citation indicators. A comparative analysis using whole and fractionalised counting methods. *Journal of Informetrics, 6*(1), 36–43. https://doi.org/10.1016/j.joi.2011.08.002

Baeza-Yates, R., & Ribeiro-Neto, B. (2011). *Modern information retrieval: The concepts and technology behind search* (2nd ed.). Addison-Wesley Publishing Company.

Bailón-Moreno, R., Jurado-Alameda, E., Ruiz-Baños, R., & Courtial, J. P. (2005). Bibliometric laws: Empirical flaws of fit. *Scientometrics, 63*(2), 209–229. https://doi.org/10.1007/s11192-005-0211-5

Bar-Ilan, J. (2008). Which h-index? — A comparison of WoS, Scopus and Google Scholar. *Scientometrics, 74*(2), 257–271. https://doi.org/10.1007/s11192-008-0216-y

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 2008*(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Bornmann, L. (2007). How can citation impact in bibliometrics be normalized? A new approach combining citing-side normalization and citation percentiles. *Quantitative Science Studies, 1*(4), 1553–1569. https://doi.org/10.1162/qss_a_00089

Bornmann, L. (2020). An evaluation of percentile measures of citation impact, and a proposal for making them better. *Scientometrics, 124*(2), 1457–1478. https://doi.org/10.1007/s11192-020-03512-7

Bornmann, L., & Daniel, H.-D. (2007). What do we know about the h index? *Journal of the American Society for Information Science and Technology, 58*(9), 1381–1385. https://doi.org/10.1002/asi.20609

Bornmann, L., & Mutz, R. (2011). Further steps towards an ideal method of measuring citation performance: The avoidance of citation (ratio) averages in field-normalization. *Journal of Informetrics, 5*(1), 228–230. https://doi.org/10.1016/j.joi.2010.10.009

Bornmann, L., & Williams, R. (2020). An evaluation of percentile measures of citation impact, and a proposal for making them better. *Scientometrics, 124*(2), 1457–1478. https://doi.org/10.1007/s11192-020-03512-7

Bornmann, L., Haunschild, R., & Mutz, R. (2020). Should citations be field-normalized in evaluative bibliometrics? An empirical analysis based on propensity score matching. *Journal of Informetrics, 14*(4), 101098. https://doi.org/10.1016/j.joi.2020.101098

Bradford, S. C. (1934). Sources of information on specific subjects. *Engineering, 26*(4), 85–86.

Campanario, J. M. (2011). Empirical study of journal impact factors obtained using the classical two-year citation window versus a five-year citation window. *Scientometrics, 87*(1), 189–204. https://doi.org/10.1007/s11192-010-0334-1

Clauset, A., Newman, M. E. J., & Moore, C. (2004). Finding community structure in very large networks. *Physical Review E, 70*(6), 066111. https://doi.org/10.1103/PhysRevE.70.066111

Cleverdon, C. W. (1972). On the inverse relationship of recall and precision. *Journal of Documentation, 28*(3), 195–201. https://doi.org/10.1108/eb026538

Cronin, B. (2001). Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices? *Journal of the American Society for Information Science and Technology, 52*(7), 558–569. https://doi.org/10.1002/asi.1097

Daniel, B. Klein, & Chiang, E. (2004). The social science citation index: A black box—with an ideological bias? *Econ Journal Watch, 1*(1), 134–165.

Daraio, C., Lenzerini, M., Leporelli, C., Naggar, P., Bonaccorsi, A., & Bartolucci, A. (2016). The advantages of an ontology-based data management approach: Openness, interoperability and data quality. *Scientometrics, 108*(1), 441–455. https://doi.org/10.1007/s11192-016-1913-6

Donner, P. (2018). Effect of publication month on citation impact. *Journal of Informetrics, 12*(1), 330–343. https://doi.org/10.1016/j.joi.2018.01.012

Donner, P. (2024). Remarks on modified fractional counting. *Journal of Informetrics, 18*(4), 101585. https://doi.org/10.1016/j.joi.2024.101585

Donner, P., Rimmert, C., & Van Eck, N. J. (2020). Comparing institutional-level bibliometric research performance indicator values based on different affiliation disambiguation systems. *Quantitative Science Studies, 1*(1), 150–170. https://doi.org/10.1162/qss_a_00013

Egghe, L. (2010). The Hirsch index and related impact measures. *Annual Review of Information Science and Technology, 44*(1), 65–114. https://doi.org/10.1002/aris.2010.1440440109

Egghe, L. (2012). Averages of ratios compared to ratios of averages: Mathematical results. *Journal of Informetrics, 6*(2), 307–317. https://doi.org/10.1016/j.joi.2011.12.007

Egghe, L., & Rousseau, R. (1996). Average and global impact of a set of journals. *Scientometrics, 36*(1), 97–107. https://doi.org/10.1007/BF02126648

El Gibari, S., Gómez, T., & Ruiz, F. (2022). Combining reference point based composite indicators with data envelopment analysis: Application to the assessment of universities. *Scientometrics, 127*(8), 4363–4395. https://doi.org/10.1007/s11192-022-04436-0

Gauffriau, M., Larsen, P. O., Maye, I., Roulin-Perriard, A., & Von Ins, M. (2008). Comparisons of results of publication counting using different methods. *Scientometrics, 77*(1), 147–176. https://doi.org/10.1007/s11192-007-1934-2

Glänzel, W. (2004). Towards a model for diachronous and synchronous citation analyses. *Scientometrics, 60*(3), 511–522. https://doi.org/10.1023/B:SCIE.0000034391.06240.2a

Glänzel, W., & Czerwon, H. J. (1996). A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level. *Scientometrics, 37*(2), 195–221. https://doi.org/10.1007/BF02093621

Glänzel, W., Schlemmer, B., & Thijs, B. (2003). Better late than never? On the chance to become highly cited only beyond the standard bibliometric time horizon. *Scientometrics, 58*(3), 571–586. https://doi.org/10.1023/B:SCIE.0000006881.30700.ea

Glänzel, W., & Thijs, B. (2012). Using 'core documents' for detecting and labelling new emerging topics. *Scientometrics, 91*(2), 399–416. https://doi.org/10.1007/s11192-011-0591-7

Harzing, A.-W., & Alakangas, S. (2016). Google Scholar, Scopus and the Web of Science: A longitudinal and cross-disciplinary comparison. *Scientometrics, 106*(2), 787–804. https://doi.org/10.1007/s11192-015-1798-9

Heintz, B. (2010). Numerische Differenz. Überlegungen zu einer Soziologie des (quantitativen) Vergleichs / Numerical difference. Toward a sociology of (quantitative) comparisons. *Zeitschrift für Soziologie, 39*(3), 162–181. https://doi.org/10.1515/zfsoz-2010-0301

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for Research Metrics. *Nature, 520*(7548), 429–431. https://doi.org/10.1038/520429a

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences, 102*(46), 16569–16572. https://doi.org/10.1073/pnas.0507655102

Jarneving, B. (2005). A comparison of two bibliometric methods for mapping of the research front. *Scientometrics, 65*(2), 245–263. https://doi.org/10.1007/s11192-005-0270-7

Johnes, J. (2018). University rankings: What do they really show? *Scientometrics, 115*(1), 585–606. https://doi.org/10.1007/s11192-018-2666-1

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation, 14*(1), 10–25. https://doi.org/10.1002/asi.5090140103

Kessler, M. M. (1963). Bibliographic coupling extended in time: Ten case histories. *Information Storage and Retrieval, 1*(4), 169–187. https://doi.org/10.1016/0020-0271(63)90016-0

Larivière, V., & Gingras, Y. (2011). Averages of ratios vs. ratios of averages: An empirical analysis of four levels of aggregation. *Journal of Informetrics, 5*(3), 392–399. https://doi.org/10.1016/j.joi.2011.02.001

Larsen, P. O., & Von Ins, M. (2010). The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics, 84*(3), 575–603. https://doi.org/10.1007/s11192-010-0202-z

Leydesdorff, L., & Opthof, T. (2010). Normalization at the field level: Fractional counting of citations. https://doi.org/10.48550/ARXIV.1006.2896

Luke, D. A. (2015). *A user's guide to network analysis in R*. Use R! Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-23883-8

Makkonen, T., & Van Der Have, R. P. (2013). Benchmarking regional innovative performance: Composite measures and direct innovation counts. *Scientometrics, 94*(1), 247–262. https://doi.org/10.1007/s11192-012-0753-2

Manning, C. D., Raghavan, P., & Schütze, H. (2009). *Introduction to information retrieval*. Cambridge University Press. https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf

Merton, R. K. (1968). The Matthew effect in science: The reward and communication systems of science are considered. *Science, 159*(3810), 56–63. https://doi.org/10.1126/science.159.3810.56

Moed, H. F. (2010). Measuring contextual citation impact of scientific journals. *Journal of Informetrics, 4*(3), 265–277. https://doi.org/10.1016/j.joi.2010.01.002

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: A comparative analysis. *Scientometrics, 106*(1), 213–228. https://doi.org/10.1007/s11192-015-1765-5

Moon, H. S., & Lee, J. D. (2005). A fuzzy set theory approach to national composite S&T indices. *Scientometrics, 64*(1), 67–83. https://doi.org/10.1007/s11192-005-0238-7

Müller, M. C., Reitz, F., & Roy, N. (2017). Data sets for author name disambiguation: An empirical analysis and a new resource. *Scientometrics, 111*(3), 1467–1500. https://doi.org/10.1007/s11192-017-2363-5

Mutschke, P., & Mayr, P. (2015). Science models for search: A study on combining scholarly information retrieval and scientometrics. *Scientometrics, 102*(3), 2323–2345. https://doi.org/10.1007/s11192-014-1485-2

Nasir, A., Ali, T. M., Shahdin, S., & Rahman, T. U. (2011). Technology achievement index 2009: Ranking and comparative study of nations. *Scientometrics, 87*(1), 41–62. https://doi.org/10.1007/s11192-010-0285-6

Nicolaisen, J., & Hjørland, B. (2007). Practical potentials of Bradford's law: A critical examination of the received view. *Journal of Documentation, 63*(3), 359–377. https://doi.org/10.1108/00220410710743298

OECD/European Union/EC-JRC. (2008). *Handbook on constructing composite indicators: Methodology and user guide*. Paris: OECD Publishing. https://doi.org/10.1787/9789264043466-en

Powers, D. M. W. (2008). Evaluation: From precision, recall and F-Measure to ROC, informedness, markedness and correlation. *Mach. Learn. Technol., 2*. https://doi.org/10.48550/ARXIV.2010.16061

Pudovkin, A. I., & Garfield, E. (2009). Percentile rank and author superiority indexes for evaluating individual journal articles and the author's overall citation performance. *Collnet Journal of Scientometrics and Information Management, 3*(2), 3–10. https://doi.org/10.1080/09737766.2009.10700871

Rimmert, C., Schwechheimer, H., & Winterhager, M. (2017). Disambiguation of author addresses in bibliometric databases - technical report. Bielefeld: Universität Bielefeld, Institute for Interdisciplinary Studies of Science (I²SoS).

Shenton, A. K., & Hay-Gibson, N. V. (2011). Bradford's Law and its relevance to researchers. *Education for Information, 27*(4), 217–230. https://doi.org/10.3233/EFI-2009-0882

Sivertsen, G., Rousseau, R., & Zhang, L. (2019). Measuring scientific contributions with modified fractional counting. *Journal of Informetrics, 13*(2), 679–694. https://doi.org/10.1016/j.joi.2019.03.010

Small, H. (1973). Co‐citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science, 24*(4), 265–269. https://doi.org/10.1002/asi.4630240406

Stahlschmidt, S., & Stephen, D. (2020). Comparison of Web of Science, Scopus and Dimensions databases. Berlin: Deutsches Zentrum für Hochschul- und Wissenschaftsforschung. https://bibliometrie.info/downloads/DZHW-Comparison-DIM-SCP-WOS.PDF

Tang, L., & Walsh, J. P. (2010). Bibliometric fingerprints: Name disambiguation based on approximate structure equivalence of cognitive maps. *Scientometrics, 84*(3), 763–784. https://doi.org/10.1007/s11192-010-0196-6

Tijssen, R. J. W., Visser, M. S., & Van Leeuwen, T. N. (2002). Benchmarking international scientific excellence: Are highly cited research papers an appropriate frame of reference? *Scientometrics, 54*(3), 381–397. https://doi.org/10.1023/A:1016082432660

Traag, V. A., Waltman, L., & Van Eck, N. J. (2019). From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports, 9*(1), 5233. https://doi.org/10.1038/s41598-019-41695-z

Traag, V. A., & Šubelj, L. (2023). Large network community detection by fast label propagation. *Scientific Reports, 13*(1), 2701. https://doi.org/10.1038/s41598-023-29610-z

Van Raan, A. F. J. (2006). Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. *Scientometrics, 67*(3), 491–502. https://doi.org/10.1556/Scient.67.2006.3.10

Vinkler, P. (2006). Composite scientometric indicators for evaluating publications of research institutes. *Scientometrics, 68*(3), 629–642. https://doi.org/10.1007/s11192-006-0123-z

Waltman, L., & Van Eck, N. J. (2015). Field-normalized citation impact indicators and the choice of an appropriate counting method. *Journal of Informetrics, 9*(4), 872–894. https://doi.org/10.1016/j.joi.2015.08.001

Wang, J. (2013). Citation time window choice for research impact evaluation. *Scientometrics, 94*(3), 851–872. https://doi.org/10.1007/s11192-012-0775-9

Weinberg, B. H. (1974). Bibliographic coupling: A review. *Information Storage and Retrieval, 10*(5–6), 189–196. https://doi.org/10.1016/0020-0271(74)90058-8

White, H. D., & Griffith, B. C. (1981). Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for Information Science, 32*(3), 163–171. https://doi.org/10.1002/asi.4630320302

Zhao, D., & Strotmann, A. (2008). Evolution of research activities and intellectual influences in information science 1996–2005: Introducing author bibliographic coupling analysis. *Journal of the American Society for Information Science and Technology, 59*(13), 2070–2086. https://doi.org/10.1002/asi.20910