Introduction to Bibliometrics and Scientometrics

Prof. Tom Kwanya

What is Bibliometrics?

- A field of study which uses mathematical and statistical techniques to study publishing and communication patterns in the distribution of information (Diodato, 1994).
- Mathematical and statistical analysis of patterns that arise in the publication and use of documents.
- Publications as used in bibliometrics refer to scientific information sources/resources.

What is Bibliometrics...Ctd?

- The use of published scientific literature (articles, books, conference proceedings, etc.) for measuring research activity e.g. output volume, science 'quality', interdisciplinarity, networking (Grant, 2015)
- New knowledge created by scientists is embedded in the scientific literature; by measuring scientific literature, we measure knowledge and the ways it is produced.

Bibliometrics vs. Informetrics

- Sometimes used interchangeably but there is a distinction between the two concepts.
- Informetrics deal with analysing the patterns in the production and use of any form of information. Bibliometrics deal with published information.
- Thus, bibliometrics can be considered as a subset of informetrics.

What is Scientometrics?

- Scientometrics applies bibliometric approaches to science (physical and natural sciences excluding social sciences).
- Scientometrics also goes beyond the usual realms of bibliometrics e.g. when it considers the politics or development of science.
- Policies on scientific research at all levels also fall within the ambit of scientometrics.

The three Metrics Compared

- Informetrics deals with all forms of information – scientific and non-scientific.
- Bibliometrics deals with all scientific publications in all fields of study.
- Scientometrics deals with scientific publications in the disciplines in natural and physical sciences.
- Scientometrics is confined; bibliometrics is wide while informetrics is overall.

Descriptive vs. Evaluative Metrics

 Descriptive metrics deal with measuring the production of publications in a given field for the purpose of comparing the amount of research in different countries; the amount of research produced during different periods of time; the amount of research output produced in different subdivisions of the field; or a combination of all the above.

Descriptive vs. Evaluative Metrics...Ctd

- Evaluative metrics assesses the quality and impact of scientific publications.
- Evaluative metrics focuses particularly on the evaluation of scientific activity, and more, in particular, on quality aspects of scientific performance (Narin, 1976).
- Such a study is often made by counting the references cited by a large number of research workers in their papers.

Historical Background

- The term bibliometrics was coined by Allan Pritchard in 1969. He defined it as "the application of mathematics and statistical methods to books and other media of communication".
- However, the concept is much older as it can be traced back to 1896 when Campbell used statistical methods to study subject scattering in publications.

Historical Background...ctd

- This was followed by a 1917 study by Cole and Eales who statistically studied the growth of literature in comparative anatomy between 1550-1860 through bibliographical citations.
- In 1923 Hulme proposed the use of counting of publications to understand the development of science and technology.
- Ranganathan proposed librametry in 1948.

Foundations

- Bibliometrics is based on the assumption that the majority of research findings are published as articles in academic journals and are read in that context by other researchers who then go on to cite these articles in their subsequent articles.
- The more citations an article receives, the greater the impact this can be taken to have had.

Foundations...ctd

- Bibliometrics involves measuring the number of published academic articles by a certain group of authors and the number of times these articles are cited, as well as studying the statistical connections between different articles, authors and subjects.
- Although there is no absolute consensus about bibliometrics, it is gaining prominence on the research scene.

Sources of Bibliometrics Data

- Thomson Reuters Citation Indices:
 - Science Citation Index Expanded
 - Social Sciences Citation Index
 - Arts and Humanities Citation Index
- Pubmed/Medline
- Scopus
- Google Scholar
- Institutional repositories

Application

- Ranking of universities and departments based on scientific production.
- Evaluation of publications and other information resources.
- Assessment of factors which influence information production, flow and use.
- Analysis of trends in the themes and consumption of information and knowledge.

Application...ctd

- Evaluation of the effectiveness or impact of information services.
- Conducting a quantitative analysis of academic literature.
- Identifying areas of research strengths and weaknesses.
- Identifying top researchers or scientific journals.

Limitations of Bibliometrics

- Citation patterns can differ greatly between disciplines.
- Some disciplines such as the arts, humanities and social sciences rely less on publishing in journals yet bibliometrics commonly focuses on journal article citations.
- A paper may be cited in a negative rather than a positive way yet the citation would still be counted.

Limitations of Bibliometrics...ctd

- The tools used to gather bibliometric data do not cover all research areas and do not index all publications. Results will vary depending on the tool you use.
- Manipulation of the system by researchers inappropriately self-citing, citing colleagues, splitting outputs into many articles etc can distort the data.

Tips for Success

- Always compare like with like
- Don't rely on a single bibliometric tool
- Be aware that some disciplines rely less on publishing in journals than others and will therefore fare less favourably.
- Put the data in context using a combination of metrics and other qualitative information where appropriate.

Bibliometrics Indicators

Prof. Tom Kwanya

Definitions of Indicators

- Bibliometric indicators are quantitative measures used in bibliometrics.
- Indicators which compare the quantity, quality and visibility of research implemented by various individuals or institutions.
- Bibliometrics indicators are metrics used to gauge bibliometrics measures – focusing on quantity, quality and visibility of research.

Purpose of Indicators

- Bibliometrics indicators are used to quantify diverse aspects of production, use and impact of scientific publications.
- Bibliometrics indicators can be used to evaluate research excellence for individual researchers, research teams and institutions.
- Bibliometrics indicators can also be used to measure research contribution of countries.

Types of Indicators

- Direct indicators these are indicators which are objective and are found directly on the documents (e.g. number of publications).
- Derived indicators these are calculated based on an analysis of documents (e.g. citations).
- Assigned indicators based on external assessments e.g. peer judgment, reputation.

Publication Count

- Total number of publications
- Number of indexed publications
- Number of publications in high impact journals
- Number of works per channel of publication
- Number of papers in local, regional or international channels of publication
- Publications per field

Publication Count...Ctd

- Relative activity index (unit's world share)
 - The analysed unit's world share of publications in a given field divided by the unit's world share of publications overall.
- Relative specialisation index (how active a unit is in a specified field)
 - A value of -1 indicates a completely idle research field and a value of 1 if all publications from the unit are in one field.

Citations

- Number of citations
- Number of citations per publication
- Normalised citation score (the relative number of citations to publications from a specific unit, compared to the world average)
- Hirsch (h) index
- Self citations
- Uncitedness

Citations...Ctd

- Average percentile (The average of shares of publications within the same types, ages and subject areas that have fewer citations than the analyzed publications.)
- Top 5% shows the share of publications attributed to a unit that belong to the 5% most highly cited publications in the world from the same year, in the same subject and of the same document type.

Citations...Ctd

- Co-citation analysis
 - monitors the number of times (two) papers are cited together in single articles or patents in a particular field
- Co-word analysis
 - monitors the number of times key words are mentioned together in publications or patents in a particular field

Collaboration

- Co-authorship
- Number of authors per publication
- Number of affiliations per publication
- Number of countries per publication
- Number of fields (disciplines) per publication
- Co-delivery of programmes
- Collaborative research

Journal Factors

- Impact factor the average number of citations a publication in a specific journal has received during the two years following its publication.
- Normalized journal impact
 - the relative number of citations to publications in one specific journal, compared to the world average of citations to publications of the same document type, age and subject area.

Bibliometrics Laws

Prof. Tom Kwanya

Lotka's Law

- Also known as the Law of Scientific Productivity.
- It deals with predicting the number of authors publishing a given number of papers.
- Developed by Alfred Lotka, a mathematician, in 1926.
- He says the number of authors making x contributions in a given period is a fraction of the number making a single (1) contribution.

- The number of authors publishing a certain number of articles is a fixed ratio to the number of authors publishing a single article.
- As the number of articles published increases, authors producing that many publications become less frequent.
- He saw that the number of persons making 2 contributions is about one fourth of those making one (1) contribution.

- The number making 3 contributions is about one ninth (1/9).
- Thus, the number of authors making n contributions is about 1/n² of those making one (1) contribution.
- The proportion of all contributors making a single (1) contribution is about 60% of all the contributors.

- Generally, Lotka's Law is an inverse square law that for every 100 authors contributing one article, 25 will contribute 2, 11 will contribute 3, and 6 will contribute 4 each.
- There is a general decrease in performance among a body of authors following 1:n²
- The pattern is reflected in the table in the following slide.

Portion of articles written	Number of authors writing that number of articles
10	$100/10^2 = 1$
9	$100/9^2 \approx 1 (1.23)$
8	$100/8^2 \approx 2 (1.56)$
7	$100/7^2 \approx 2 (2.04)$
6	$100/6^2 \approx 3 (2.77)$
5	$100/5^2 = 4$
4	$100/4^2 \approx 6 (6.25)$
3	$100/3^2 \approx 11 (11.111)$
2	$100/2^2 = 25$
1	100

Bradford's Law

- Also known as the Law of Scattering.
- Developed by Samuel Clement Bradford, a British librarian/mathematician, in 1934.
- Dr. Bradford was concerned about the scattering of articles on specific subjects in various journals leading to some of them being left out by indexing agencies.

- Three types of scattering have been identified by scholars who support Bradford's Law:
 - Lexical scattering: this involves scattering of words in collections of text;
 - Semantic scattering: involves scattering of concepts in collections of text;
 - Subject scattering: the scattering of items useful in solving specific problems.

- Bradford argued that every scientific field is related somehow – however remotely – to all other fields.
- He stated that in any bibliography on a subject, there is always a small group of core journals which account for a substantial (1/3) proportion of the articles on that subject or discipline.

- Then there is a larger group which account for another 1/3 and even a much larger group accounting for the other 1/3.
- He stated that if all journals in a field are sorted according to the number of articles into three groups with each group having a third of the articles, the number of journals will be proportional to 1:n:n² where n is Bradford's multiplier (a constant).

- A small core of, for example, journals have as many papers on a given subject as a much larger number of journals, n, which again has as many papers on the subject as n² journals.
- A growth in the number of articles in a subject requires a growth in the number of journals or information sources in the subject.

- Bradford's Law has been used as an argument about:
 - how to build collections,
 - how to select journals to be indexed in bibliographies,
 - how to measure the coverage of bibliographies,
 - how to solve practical problems related to information seeking and retrieval

Zipf's Law

- Also known as the Law of Word Occurrence developed by linguist George Kingsley Zipf in 1935. The law counts the frequency with which specific words occur in scientific publications.
- He analysed the words (29,899) used in a novel known as *Ulysses* written by James Joyce.
- He ranked all the words used in the book in order of the frequency of their occurrence with the highest ranked 1 (r=1) and the least ranked last (r=29,899).

Zipf's Law...Ctd

- He noticed that by multiplying the numerical value of the rank (r) with the frequency of occurrence (f), he got a product, C, which was constant throughout the entire list of words.
- Based on the above observation he developed a formula rf=C which became known as Zipf's Law.
- The frequency of any word is inversely proportional to its rank in the frequency table.

Zipf's Law...Ctd

- Thus, the most frequent word will occur approximately twice as often as the second most frequent word, three times as often as the third most frequent word, etc.
- Therefore, the word in the position n appears
 1/n times as often as the most frequent word.
- In this type of distribution, frequency declines sharply as the rank number increases.

Zipf's Law...Ctd

- So, a small number of items appear very often, and a large number rarely occur.
- The most common word in English is "the," which appears about one-tenth of the time in a typical text; the next most common word (rank 2) is "of," which appears about one-twentieth of the time.
- Studies show that this phenomenon also applies in nearly every language.

Garfield's Law

- Also known as Garfield's Law of Concentration.
- It points out that for any field of science, articles are concentrated within the same highly cited or multidisciplinary journals.
- He also opined that the core literature for all scientific literature involves a group of not more than 1,000 journals and could be as few as 500 journals.

Garfield's Law

- This observation implies that a good general science library need not have more journals than a special library holding a specialised collection covering a specific discipline.
- He also argued that an index does not have to include all journals in a discipline as long as it covers the core journals.

OPEN ACCESS SCHOLARLY PUBLISHING

Prof. Tom Kwanya

Definitions

- Open access is a mechanism by which research outputs are distributed online, free of cost or other access barriers.
- This mechanism reduces or completely eliminates access barriers related to copyrights or licenses.

Types of Open Access

- Gold provides full access to publications.
 Such access is provided through creative commons licenses.
- Green allows authors to self-archive prepublication copies of research output on institutional repositories.
- <u>Hybrid</u> publications which have both open and closed access articles side by side.

Types of Open Access...Ctd

- Bronze publishes articles for subscription but open them after an embargo period ranging from months to years.
- Black describes systems which enable subscribers to share restricted publications with people who have not paid for the same.
- <u>Platinum</u> open access publishers who do not charge authors to publish articles.

Open Access Licenses

- Attribution (CC BY)
 - lets others distribute, remix, tweak, and build upon one's work, even commercially, as long as they credit you for the original creation.
- Attribution ShareAlike (CC BY-SA)
 - lets others remix, tweak, and build upon one's work even for commercial purposes, as long as they credit the original author and license their new creations under identical terms.

Open Access Licenses...Ctd

- Attribution NonCommercial-ShareAlike (CC BY-NC-SA)
 - lets others remix, tweak, and build upon one's work non-commercially, as long as they credit the original wuthor and license their new creations under the identical terms.
- Attribution NonCommercial NoDerivs (CC BY-NC-ND)
 - allows others to download works and share them with others as long as they credit the originator; they can't change them in any way or use them commercially.

Merits of Open Access

- Enhanced access
- Immediacy
- Enhanced impact (citations)
- Enhanced visibility (author and institutional)
- Stimulates knowledge creation and diffusion
- Liberalises research sphere (less control)
- Value for money (increases public good)

Demerits of Open Access

- Quality (credibility)
- Article handling costs (APC)
- Predatory publishers
- Non-standardised access licensing regime
- Less sustainable (no direct revenue)
- Depends highly on technology
- Low impact factor publications

Quality Assurance in Open Access

- The main quality control mechanism is peer review
- Two main types of peer review exist:
 - Blind (double blind): parties (authors and peer reviewers) do not know each other.
 - Open: parties in the peer review process know each other.
- Double blind peer review is commonly used.

Quality Assurance in Open Access

- Limitations of peer review:
 - Bias
 - Veracity
 - Lack of transparency
 - Time constraints
 - Lack of incentives to reviewers
 - Lack of competent peer reviewers
 - Contradictions

Introduction to Social Network Analysis

Prof. Tom Kwanya

Definitions

- Social network analysis (SNA) is the process of investigating social structures through the use of networks (Otte and Rousseau, 2002).
- The mapping and measuring of relationships and flows between people, groups, organizations, computers, URLs, and other connected information/knowledge entities. The nodes in the network are the people and groups.

Key Concepts...Ctd

- Sociogram This is a visualization of a social network with defined boundaries of connections in the network.
- Size A measure of the number of actors (nodes) in a complete or egocentric network.
- Distance The number of "steps" between any two actors in a network.
- Node Entities in a social network.

Key Concepts

- Density The number of connections a participant has, divided by the total possible connections a participant could have.
- Centrality Measures the extent to which an individual interacts with other individuals in the network. The more an individual connects to others in a network, the greater their centrality in the network.

Types of Social Networks

- **Egocentric** Focuses on the individual; studies an individual's personal network and its affects on that individual.
- Sociocentric Focuses on large groups of people; quantifies relationships between people in a group; studies patterns of interactions and how these patterns affect the group as a whole.

Metrics – Connections

- Homophily The extent to which actors form ties with similar versus dissimilar others.
- Multiplexity The number of content-forms contained in a tie. For example, two people who are friends and also work together would have a multiplexity of 2.
- Mutuality/Reciprocity The extent to which two actors reciprocate each other's friendship or other interaction.

Metrics – Distributions

- Bridge An individual whose weak ties fill a structural hole, providing the only link between two individuals or clusters
- Centrality The "importance" or "influence" of a node in a network.
- Density The proportion of direct ties in a network relative to the total number possible.
- Distance The minimum number of ties required to connect two particular actors.

Metrics – Distributions...Ctd

- Structural holes The absence of ties between two parts of a network.
- Tie strength Defined by the linear combination of time, emotional intensity, intimacy and reciprocity. Strong ties are associated with homophily, propinquity (closeness) and transitivity (cross-relations), while weak ties are associated with bridges.

Metrics – Segmentation

- Groups are identified as 'cliques' if every individual is directly tied to every other individual; 'social circles' if there is less direct.
- Clustering coefficient A measure of the likelihood that two associates of a node are associates; a higher clustering coefficient indicates a greater 'cliquishness'.
- Cohesion The degree to which actors are connected directly to each other

Application

- Friendship and acquaintance networks
- Collaboration networks
- Kinship and relationships
- Disease transmission
- Customer interaction analysis
- Marketing
- Business intelligence
- Individual and group engagement

Application...Ctd

- In Information and knowledge management, social network analysis can be used to study the following:
 - information needs;
 - information exposure;
 - information legitimation;
 - information routes; and
 - information opportunities.

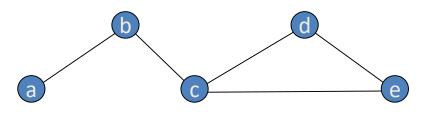
Social Network Analysis Tools

- AllegroGraph
- Automat
- Gephi
- GraphStream
- Graphviz
- Mathematica
- NodeXL

- NetworkX
- UNISoN
- UCINET
- Tulip
- Wolfram Alpha
- Commetrix
- EgoNet

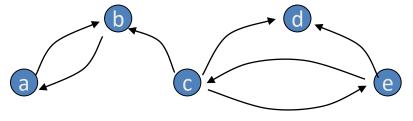
Data Structures

From pictures to matrices



Undirected, binary

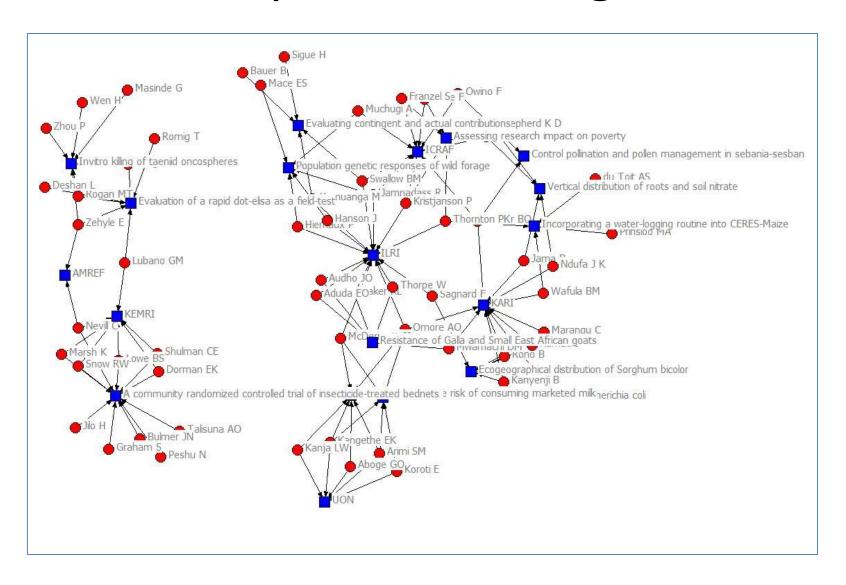
•	а	b	С	d	е
a		1			
b	1		1		
С		1		1	1
d			1		1
е			1	1	



Directed, binary

	a	b	С	d	е
a		1			
b	1				
С		1		1	1
d					
e			1	1	

Example of a Sociogram



Significance

"To speak of social life is to speak of the association between people – their associating in work and in play, in love and in war, to trade or to worship, to help or to hinder. It is in the social relations men establish that their interests find expression and their desires become realized."

Peter M. Blau

Exchange and Power in Social Life, 1964

Significance...Ctd

"If we ever get to the point of charting a whole city or a whole nation, we would have ... a picture of a vast solar system of intangible structures, powerfully influencing conduct, as gravitation does in space. Such an invisible structure underlies society and has its influence in determining the conduct of society as a whole."

J.L. Moreno

New York Times, April 13, 1933

Significance...Ctd

"For the last thirty years, empirical social research has been dominated by the sample survey. But as usually practiced, ..., the survey is a sociological meat grinder, tearing the individual from his social context and guaranteeing that nobody in the study interacts with anyone else in it."

Allen Barton, 1968

(Quoted in Freeman 2004)

Significance...Ctd

MOTHER GOOSE & GRIMM By Mike Peters.





SALLY FORTH By Greg Howard



CITATION ANALYSIS

Prof. Tom Kwanya

Definitions

- Citation analysis is a way of measuring the relative importance or impact of an author, an article or a publication by counting the number of times that author, article, or publication has been cited by other works.
- Citation analysis may be conducted to establish the impact that a particular work has had; learn more about a field or a topic; and to determine what impact a particular author has had.



Library Management 41 (6/7), 339-353

Prof. Tom Kwanya 🗸



Professor of Information and Knowledge Management, The Technical University of Kenya

Verified email at tukenya.ac.ke - Homepage

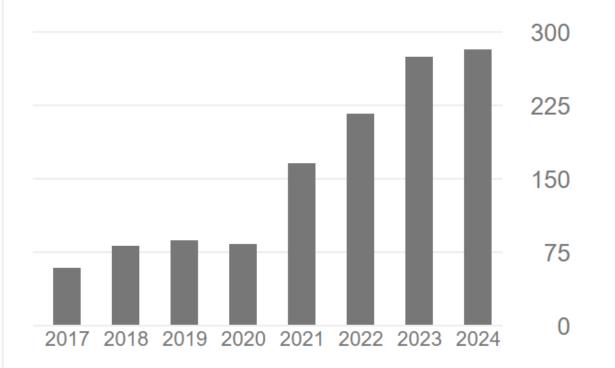
knowledge management information science communication public relations web publishing

TITLE	1	‡	CITED BY	YEAR
T Kwany	n, C Stily	rries and apomediators: Distinguishing between Library 3.0 and Library 2.0 well, PG Underwood and Information Science 45 (3), 187-197	98	2013
Library 3.0: intelligent libraries and apomediation T Kwanya, C Stilwell, P Underwood Elsevier			81	2014
T Kwany	a, C Stily	rsus other library service models: A critical analysis well, PG Underwood unship and Information Science 44 (3), 145-162	54	2012
T Kwany	a, C Stilv	n of Web 2.0 tools by libraries in Kenya: a reality check vell, P Underwood Conference, Nairobi, Kenya: June	51	2012
T Kwans	/B	d perishing? Publishing patterns of information science academics in Kenya opment 36 (1), 5-15	43	2020
	K Wakur	I in Africa: challenges and opportunities	36	2023
		roles of academic librarians in Kenya; apomediaries or infomediaries?	31	2020

Cited	by		VIEW ALL
		All	Since 2019
Citation	ns	1438	1121
h-index	t.	21	18
i10-ind	ex	46	39
			300
			225
		ш	150
- 1		Ш	150 75
2017 20	18 2019 2020	2021 2022 20	75
2017 20 Co-au		2021 2022 20	75
		vell	75 123 2024 0 EDIT

Cited by	VIEW ALL
----------	----------

	All	Since 2019	
Citations	1438	1121	
h-index	21	18	
i10-index	46	39	



Citation Metrics

- Total number of citations
- Average number of citations per paper
- Number of citations per author (co-authored)
- Number of citations per author per year
- Number of authors per paper
- Average number of authors per paper (sum of the author counts across all papers, divided by the total number of papers)

H-Index

- Author-level metric which measures both productivity and citation impact of scientific publications of a scholar.
- It is based on the set of the scientist's most cited papers and the number of citations that they have received.
- The h-index is the maximum value of h such that the given author/journal has published h papers that have each been cited at least h times.

Individual H-Index

- Divides the standard h-index by the average number of authors in the articles that contribute to the h-index, in order to reduce the effects of co-authorship.
- It provides a framework for apportioning the popularity of co-authored works to individual contributing authors.
- Normalises influence of co-authored works.

i10-Index

- Created by Google Scholar
- i10-Index = the number of publications of a scholar which have at least 10 citations.
- It is a way of assessing the impact of a scholar's work based on the number of citations their most popular works have attracted.
- It is only used by Google Scholar.

Author Influence - ResearchGate



Tom Kwanya D Edit

PhD in Information Studies - Professor at Technical University of Kenya

Website Kenya

Current activity

Research Interest Score 808.2 692 Citations h-index 14

Profile

Research (127)

Stats

Following

Saved list

Add research

View your latest weekly report >

Citations over time

Overall publications stats

808.2

Research Interest Score

+2.7 last week

57,531

Reads (1)

→ +158 last week

692

Citations

→ +2 last week

307

Recommendations

> ---

Journal Impact Factor

- A measure of the frequency with which the average article in a journal has been cited in a particular year.
- It is used to measure the importance or rank of a journal by calculating the times its articles are cited.
- Often, journal impact is pegged on the number of citations attracted in the first two years after publication.

Journal Impact Factor

Top publications

Top cited publications over the last five years Learn more

	Publication	h5-index	h5-median
1.	Nature	488	745
2.	IEEE/CVF Conference on Computer Vision and Pattern Recognition	440	689
3.	The New England Journal of Medicine	434	897
4.	Science	409	633
5.	Nature Communications	375	492
6.	The Lancet	368	678
7.	Neural Information Processing Systems	337	614
8.	Advanced Materials	327	420
9.	Cell	320	482
10.	International Conference on Learning Representations	304	584

Journal h5-index

- The h5-index for a journal is the largest number h such that h articles in the journal have been cited at least h times each.
- To calculate the h5-index of a journal, you would typically consider the h-index of the articles published in that journal.
- The specific h5-index of a journal or conference can vary and may change over time as new articles are published and cited.

Journal h5-median

- This is a metric used in academic and scientific research to evaluate the impact of scholarly journals, much like the h5-index.
- While the h5-index emphasizes the most highly cited articles, the h5-median considers the middle point in the distribution of citations.
- Journals with a high h5-median value are more likely to have a consistent impact across their articles, as opposed to a few highly cited ones skewing the results.

Age-Weighted Citation Rate

- The AWCR measures the number of citations to an entire body of work, adjusted for the age of each individual paper.
- It is calculated by dividing the number of citations to a given paper by the age of that paper.
- It helps to normalise the influence of the age of publication of papers on their citation.

Immediacy Index

- This is the average number of times an article is cited in the year of its publication.
- The Immediacy Index is calculated by dividing the number of citations to articles published in a given year by the number of articles published in that year.
- Immediacy index gives an indication of the uptake of latest research.

• Group 1:

- What are the primary advantages of using the h5index as a metric for evaluating the impact of scholarly journals?
- Are there any limitations or demerits associated with relying solely on the h5-index? If so, what are they?

• Group 2:

- How does the h5-median differ from the h5-index in terms of what it measures in academic publications?
- What are the strengths and weaknesses of the h5median as a metric for assessing the impact of journals?

• Group 3:

- In what situations would it be beneficial to consider both the h5-index and the h5-median when evaluating the influence of a journal?
- How might combining these two metrics provide a more comprehensive view of the impact of a journal?

• Group 4:

- What are the contextual factors that researchers and scholars should consider when using the h5index and h5-median to assess journals in specific fields or disciplines?
- How might the choice between these metrics depend on the goals and objectives of the assessment?

