1-1-2016

The Mathematics of Scientific Research: Scientometrics, Citation Metrics, and Impact Factors

Clayton Hayes
Wayne State University, as6348@wayne.edu

Recommended Citation
http://digitalcommons.wayne.edu/libsp/109

This Presentation is brought to you for free and open access by the Wayne State University Libraries at DigitalCommons@WayneState. It has been accepted for inclusion in Library Scholarly Publications by an authorized administrator of DigitalCommons@WayneState.
The Mathematics of Scientific Research

Scientometrics, Citation Metrics, and Impact Factors

C. Hayes

April 1, 2016
What We Will Cover

In this talk, I will cover:

1. The definition of and foundation for Scientometrics
2. Price and his observations on citation networks
3. Garfield and his development of journal impact factors
   3.1 Bergstrom & West and the Eigenfactor
4. Hirsch and the development of the *h*-index
   4.1 Further developments in author metrics
A Brief Note on Citations

BIMEASURE ALGEBRAS ON LCA GROUPS
Graham & Schreiber
1984
A Brief Note on Citations

- **ON LINEAR TRANSFORMATIONS**
  - Phillips
  - 1940

- **BIMEASURE ALGEBRAS ON LCA GROUPS**
  - Graham & Schreiber
  - 1984

- **ADVANCES IN QUANTIZED FUNCTIONAL ANALYSIS**
  - Effros
  - 1986
A Brief Note on Citations

- **ON LINEAR TRANSFORMATIONS**
  - Phillips
  - 1940

- **BIMEASURE ALGEBRAS ON LCA GROUPS**
  - Graham & Schreiber
  - 1984

- **ADVANCES IN QUANTIZED FUNCTIONAL ANALYSIS**
  - Effros
  - 1986

- **Cites**
  - C. Hayes
  - The Mathematics of Scientific Research
  - April 1, 2016

- **Cited by**
  - C. Hayes
  - The Mathematics of Scientific Research
  - April 1, 2016
A Brief Note on Citations

- **ON LINEAR TRANSFORMATIONS**
  - Phillips
  - 1940

- **BIMEASURE ALGEBRAS ON LCA GROUPS**
  - Graham & Schreiber
  - 1984

- **ADVANCES IN QUANTIZED FUNCTIONAL ANALYSIS**
  - Effros
  - 1986
Scientometrics

**Scientometrics** is best described as the quantitative study of scientific communication. Its foundations were laid out in a paper published by S. C. Bradford in 1934 [1].
**Scientometrics** is best described as the quantitative study of scientific communication. Its foundations were laid out in a paper published by S. C. Bradford in 1934 [1].
**Scientometrics** is best described as the quantitative study of scientific communication. Its foundations were laid out in a paper published by S. C. Bradford in 1934 [1].

\[ J_0 \]

\[ J_{11} \quad J_{12} \quad J_{13} \quad J_{14} \quad J_{15} \]
Scientometrics is best described as the quantitative study of scientific communication. Its foundations were laid out in a paper published by S. C. Bradford in 1934 [1].
The Foundations of Scientometrics

Scientometrics had its true germination in the Cold War era, in the 1960s and 1970s. There are a three main reasons for this:

- Computers were more accessible and could handle much of the necessary work
- The onset of "Big Science" made the meta-analysis of scientific research more relevant
- The founding of the Institute for Scientific Information in 1960 and of the Science Citation Index in 1964

This situation laid the groundwork for Price and Garfield, the two most influential figures in the development of Scientometrics.
The Foundations of Scientometrics

Scientometrics had its true germination in the Cold War era, in the 1960s and 1970s. There are three main reasons for this:

- Computers were more accessible and could handle much of the necessary work

  - The onset of "Big Science" made the meta-analysis of scientific research more relevant
  - The founding of the Institute for Scientific Information in 1960 and of the Science Citation Index in 1964

This situation laid the groundwork for Price and Garfield, the two most influential figures in the development of Scientometrics.
Scientometrics had its true germination in the Cold War era, in the 1960s and 1970s. There are three main reasons for this:

- Computers were more accessible and could handle much of the necessary work
- The onset of “Big Science” made the meta-analysis of scientific research more relevant
The Foundations of Scientometrics

Scientometrics had its true germination in the Cold War era, in the 1960s and 1970s. There are a three main reasons for this:

- Computers were more accessible and could handle much of the necessary work
- The onset of “Big Science” made the meta-analysis of scientific research more relevant
- The founding of the Institute for Scientific Information in 1960 and of the Science Citation Index in 1964
Scientometrics had its true germination in the Cold War era, in the 1960s and 1970s. There are three main reasons for this:

- Computers were more accessible and could handle much of the necessary work
- The onset of “Big Science” made the meta-analysis of scientific research more relevant
- The founding of the Institute for Scientific Information in 1960 and of the Science Citation Index in 1964

This situation laid the groundwork for Price and Garfield, the two most influential figures in the development of Scientometrics.
Derek de Solla Price was a physicist and mathematician who developed an interest in the history of science in his late 20s. He developed a theory that the study of science was growing exponentially…
Derek de Solla Price was a physicist and mathematician who developed an interest in the history of science in his late 20s. He developed a theory that the study of science was growing exponentially...

Philosophical Transactions of the Royal Society
Citation Networks

Based on his background in math and physics, and because of the availability of the Science Citation Index, he introduced the idea of a Citation Network in 1965 [2].
Citation Networks
Based on his background in math and physics, and because of the availability of the Science Citation Index, he introduced the idea of a Citation Network in 1965 [2].
Citation Networks

Based on his background in math and physics, and because of the availability of the Science Citation Index, he introduced the idea of a Citation Network in 1965 [2].
Citation Networks

Based on his background in math and physics, and because of the availability of the Science Citation Index, he introduced the idea of a Citation Network in 1965 [2].

\[ A_{12} \rightarrow A \rightarrow A_{21} \]
\[ A_{11} \rightarrow A \rightarrow A_{22} \]
\[ A_{13} \rightarrow A \rightarrow A_{23} \]
Citation Networks

Based on his background in math and physics, and because of the availability of the Science Citation Index, he introduced the idea of a Citation Network in 1965 [2].
Citation Networks
Based on his background in math and physics, and because of the availability of the Science Citation Index, he introduced the idea of a Citation Network in 1965 [2].
Citation Networks

Based on his background in math and physics, and because of the availability of the Science Citation Index, he introduced the idea of a Citation Network in 1965 [2].
Citation Networks

Based on his background in math and physics, and because of the availability of the Science Citation Index, he introduced the idea of a Citation Network in 1965 [2].

A12 → A31 → A22

A11 → A → A21

A13 → A23 → A32
Price’s Observations

Price’s primary concern was to develop a reasonable picture of the nature of scholarship in the sciences, and made some conclusions based on his observations of the SCI data:

▶ Each year, about 10% of papers “die”, i.e. are never cited again
▶ For the “live” papers, odds of being cited in a particular year are 60%
▶ About 1% of papers are completely isolated
▶ The age of a paper has a significant impact on the number of times it is cited
▶ Around 70% of citations are randomly distributed over all published papers
▶ The other 30% are to a more selective collection of recent literature

Price developed this further by referring to that small, selective collection of recent literature as a research front.
Price’s Observations

Price’s primary concern was to develop a reasonable picture of the nature of scholarship in the sciences, and made some conclusions based on his observations of the SCI data:

- Each year, about 10% of papers “die”, i.e. are never cited again
Price’s Observations

Price’s primary concern was to develop a reasonable picture of the nature of scholarship in the sciences, and made some conclusions based on his observations of the SCI data:

- Each year, about 10% of papers “die”, i.e. are never cited again
- For the “live” papers, odds of being cited in a particular year are 60%
Price’s primary concern was to develop a reasonable picture of the nature of scholarship in the sciences, and made some conclusions based on his observations of the SCI data:

- Each year, about 10% of papers “die”, i.e. are never cited again
- For the “live” papers, odds of being cited in a particular year are 60%
- About 1% of papers are completely isolated
Price’s primary concern was to develop a reasonable picture of the nature of scholarship in the sciences, and made some conclusions based on his observations of the SCI data:

- Each year, about 10% of papers “die”, i.e. are never cited again
- For the “live” papers, odds of being cited in a particular year are 60%
- About 1% of papers are completely isolated
- The age of a paper has a significant impact on the number of times it is cited
  - Around 70% of citations are randomly distributed over all published papers
  - The other 30% are to a more selective collection of recent literature
Price’s primary concern was to develop a reasonable picture of the nature of scholarship in the sciences, and made some conclusions based on his observations of the SCI data:

- Each year, about 10% of papers “die”, i.e. are never cited again
- For the “live” papers, odds of being cited in a particular year are 60%
- About 1% of papers are completely isolated
- The age of a paper has a significant impact on the number of times it is cited
  - Around 70% of citations are randomly distributed over all published papers
  - The other 30% are to a more selective collection of recent literature

Price developed this further by referring to that small, selective collection of recent literature as a research front.
Price’s Observations

New Papers

- 
- 
- 
- 
- 

C. Hayes  The Mathematics of Scientific Research  April 1, 2016 9 / 29
Price’s Observations

Half of published papers (one each)

New Papers

•

•

•

•

•
Price’s Observations

“Research Front”

Half of published papers (one each)

New Papers
Price’s Observations

“Research Front”

Half of published papers (one each)

New Papers

C. Hayes
The Mathematics of Scientific Research
April 1, 2016 9 / 29
Price and “Cumulative Advantage”

Price developed these theories further, and outlined what he called cumulative advantage in a 1976 paper [3]. You are probably familiar with the usual example of drawing items from an urn:

\[
\begin{align*}
\text{\( n \) red} \\
\text{\( m \) black}
\end{align*}
\]
Price and “Cumulative Advantage”

Price developed these theories further, and outlined what he called cumulative advantage in a 1976 paper [3]. You are probably familiar with the usual example of drawing items from an urn:

\[
P(\text{red}) = \frac{n}{n + m}
\]

with replacement

\[
P(\text{red}) = \frac{n-1}{n+m-1}
\]

without replacement
Price and “Cumulative Advantage”

Price developed these theories further, and outlined what he called cumulative advantage in a 1976 paper [3]. You are probably familiar with the usual example of drawing items from an urn:

\[ P(\text{red}) = \frac{n}{n+m} \]
Price and “Cumulative Advantage”

Price developed these theories further, and outlined what he called cumulative advantage in a 1976 paper [3]. You are probably familiar with the usual example of drawing items from an urn:

\[
P(\text{red}) = \frac{n}{n+m}
\]
Price and “Cumulative Advantage”

Price developed these theories further, and outlined what he called **cumulative advantage** in a 1976 paper [3]. You are probably familiar with the usual example of drawing items from an urn:

\[
P(\text{red}) = \frac{n}{n+m}
\]

\[
P(\text{red}) = \frac{n-1}{n+m-1}
\]
“Cumulative Advantage”

Consider the following instead:

\[ P(\text{red}) = \frac{n}{n+m} \]

If red is drawn

If black is drawn
“Cumulative Advantage”

Consider the following instead:

\[
P(\text{red}) = \frac{n}{n+m}
\]

if red is drawn

\[
P(\text{red}) = \frac{n+c}{n+c+m}
\]

if black is drawn
Consider the following instead:

\[
P(\text{red}) = \frac{n}{n+m} \\
P(\text{red}) = \frac{n+c}{n+c+m}
\]

if red is drawn

\[
P(\text{red}) = \frac{n}{n+m}
\]

if black is drawn
Price extrapolated this to develop what he termed the **Cumulative Advantage Distribution**. Consider the base case, with $n = m = c = 1$, and let $k$ denote number of successes:

1 red
1 black

$k = 0$
The Cumulative Advantage Distribution

Price extrapolated this to develop what he termed the **Cumulative Advantage Distribution**. Consider the base case, with \( n = m = c = 1 \), and let \( k \) denote number of successes:

\[
P(red) = \frac{1}{2}
\]

1 red
1 black

\( k = 0 \)
The Cumulative Advantage Distribution

Price extrapolated this to develop what he termed the **Cumulative Advantage Distribution**. Consider the base case, with $n = m = c = 1$, and let $k$ denote number of successes:

- $k = 0$
  - 1 red
  - 1 black
  - $P(\text{red}) = \frac{1}{2}$

- $k = 1$
  - 2 red
  - 1 black
The Cumulative Advantage Distribution

Price extrapolated this to develop what he termed the **Cumulative Advantage Distribution**. Consider the base case, with \( n = m = c = 1 \), and let \( k \) denote number of successes:

1 red 1 black

\[ k = 0 \]

\[ P(\text{red}) = \frac{1}{2} \]

2 red 1 black

\[ k = 1 \]

\[ P(\text{red}) = \frac{2}{3} \]
The Cumulative Advantage Distribution

Price extrapolated this to develop what he termed the **Cumulative Advantage Distribution**. Consider the base case, with $n = m = c = 1$, and let $k$ denote number of successes:

- **Case 1**: $k = 0$
  - 1 red
  - 1 black
  - $P(red) = \frac{1}{2}$

- **Case 2**: $k = 1$
  - 2 red
  - 1 black
  - $P(red) = \frac{2}{3}$

- **Case 3**: $k = 2$
  - 3 red
  - 1 black
  - $P(red) = \frac{3}{4}$

- **Case 4**: $k = 3$
  - 4 red
  - 1 black
  - $P(red) = \frac{4}{5}$

So the probability of success after $k$ successes is then...
The Cumulative Advantage Distribution

Price extrapolated this to develop what he termed the Cumulative Advantage Distribution. Consider the base case, with \( n = m = c = 1 \), and let \( k \) denote number of successes:

- \( k = 0 \)
  - 1 red, 1 black
  - \( P(\text{red}) = \frac{1}{2} \)
  - 3 red, 1 black
  - \( P(\text{red}) = \frac{3}{4} \)

- \( k = 1 \)
  - 2 red, 1 black
  - \( P(\text{red}) = \frac{2}{3} \)
  - 4 red, 1 black
  - \( P(\text{red}) = \frac{4}{5} \)

So the probability of success after \( k \) successes is then \( \ldots \)
The Cumulative Advantage Distribution

Price extrapolated this to develop what he termed the Cumulative Advantage Distribution. Consider the base case, with $n = m = c = 1$, and let $k$ denote number of successes:

1 red 
1 black

$k = 0$

$P(red) = \frac{1}{2}$

3 red
1 black

$k = 2$

$P(red) = \frac{3}{4}$

2 red
1 black

$k = 1$

$P(red) = \frac{2}{3}$

4 red
1 black

$k = 3$

$P(red) = \frac{4}{5}$

So the probability of success after $k$ successes is then $\frac{1+k}{2+k}$. 
The Cumulative Advantage Distribution

Do this for $K$ urns, stopping each whenever you draw your first “failure”. One would then expect to have...

$k = 1 \quad K/2$
successes

$k = 2 \quad K/3$
successes

$k = 3 \quad K/4$
...
...

The expected number of urns with at least $k$ successes is then $K/k + 1$
while the number of urns with exactly $k$ successes is $K/k + 1 - K/k + 2 = K(k+1)(k+2)$
The Cumulative Advantage Distribution

Do this for \( K \) urns, stopping each whenever you draw your first “failure”. One would then expect to have...

\[
\begin{align*}
  k = 1 & \quad \frac{K}{2} \text{ successes} \\
  k = 2 & \quad \frac{K}{3} \text{ successes} \\
  k = 3 & \quad \frac{K}{4} \\
  \vdots & \quad \vdots
\end{align*}
\]
The Cumulative Advantage Distribution

Do this for $K$ urns, stopping each whenever you draw your first “failure”. One would then expect to have...

\[
\begin{align*}
  k = 1 & \quad K/2 \text{ successes} \\
  k = 2 & \quad K/3 \text{ successes} \\
  k = 3 & \quad K/4 \\
  \vdots & \quad \vdots
\end{align*}
\]

The expected number of urns with \textit{at least} $k$ successes is then
The Cumulative Advantage Distribution

Do this for \( K \) urns, stopping each whenever you draw your first “failure”. One would then expect to have...

\[
\begin{align*}
  k &= 1 & K/2 & \text{successes} \\
  k &= 2 & K/3 & \text{successes} \\
  k &= 3 & K/4 & \\
  & \vdots & & \vdots
\end{align*}
\]

The expected number of urns with \textit{at least} \( k \) successes is then

\[
K/k + 1
\]
The Cumulative Advantage Distribution

Do this for $K$ urns, stopping each whenever you draw your first “failure”. One would then expect to have...

\[
\begin{align*}
  k = 1 & \quad \frac{K}{2} \text{ successes} \\
  k = 2 & \quad \frac{K}{3} \text{ successes} \\
  k = 3 & \quad \frac{K}{4} \\
  & \quad \vdots \\
\end{align*}
\]

The expected number of urns with \textit{at least} $k$ successes is then

\[
\frac{K}{k} + 1
\]

while the number of urns with \textit{exactly} $k$ successes is
The Cumulative Advantage Distribution

Do this for $K$ urns, stopping each whenever you draw your first “failure”. One would then expect to have...

\[
\begin{align*}
  k = 1 & \quad K/2 \text{ successes} \\
  k = 2 & \quad K/3 \text{ successes} \\
  k = 3 & \quad K/4 \\
  \vdots & \quad \vdots
\end{align*}
\]

The expected number of urns with \textit{at least} $k$ successes is then

\[K/k + 1\]

while the number of urns with \textit{exactly} $k$ successes is

\[
\frac{K}{k+1} - \frac{K}{k+2} = \frac{K}{(k+1)(k+2)}
\]
The Cumulative Advantage Distribution

What you will get is something like the following:

- **Urn1**: 6 red, 1 black, $k = 5$
- **Urn2**: 1 red, 1 black, $k = 0$
- **Urn3**: 2 red, 1 black, $k = 1$
- **Urn4**: 4 red, 1 black, $k = 3$
- **Urn5**: 1 red, 1 black, $k = 0$
- **Urn6**: 2 red, 1 black, $k = 1$
- **...**
The Cumulative Advantage Distribution

If you’re looking to actually model a citation network, taking into account new articles and new citations, it’s much more complex.

What Price showed is that, if $f(k)$ represents the fraction of the population in state $k$ (e.g. having exactly $k$ successes),

$$f(k) = CP_j B(k, j+2)$$

where $C$ is Euler’s constant, $P$ is the size of the population, $j$ is a constant parameter, and $B(\cdot, \cdot)$ is the Beta Function:

$$B(a, b) = B(b, a) = \int_0^1 x^{a-1} (1-x)^{b-1} \, dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \frac{(a-1)! (b-1)!}{(a+b-1)!}$$

He defined the Cumulative Advantage Distribution as having density function $f^*(k) = (j+1) B(k, j+2)$.
The Cumulative Advantage Distribution

If you’re looking to actually model a citation network, taking into account new articles and new citations, its much more complex.

What Price showed is that, if $f(k)$ represents the fraction of the population in state $k$ (e.g. having exactly $k$ successes),

$$f(k) = CP^j B(k, j + 2)$$

where $C$ is Euler’s constant, $P$ is the size of the population, $j$ is a constant parameter, and $B(\cdot, \cdot)$ is the Beta Function:

$$B(a, b) = B(b, a) = \int_0^1 x^{a-1}(1 - x)^{b-1} dx$$

$$= \frac{\Gamma(a)\Gamma(b)}{\Gamma(a + b)} = \frac{(a - 1)!(b - 1)!}{(a + b - 1)!}$$
The Cumulative Advantage Distribution

If you’re looking to actually model a citation network, taking into account new articles and new citations, its much more complex.

What Price showed is that, if \( f(k) \) represents the fraction of the population in state \( k \) (e.g. having exactly \( k \) successes),

\[
f(k) = CP^j B(k, j + 2)
\]

where \( C \) is Euler’s constant, \( P \) is the size of the population, \( j \) is a constant parameter, and \( B(\cdot, \cdot) \) is the Beta Function:

\[
B(a, b) = B(b, a) = \int_0^1 x^{a-1}(1 - x)^{b-1} dx
\]

\[
= \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \frac{(a - 1)!(b - 1)!}{(a + b - 1)!}
\]

He defined the Cumulative Advantage Distribution as having density function

\[
f^*(k) = (j + 1) B(k, j + 2)
\]
The Cumulative Advantage Distribution

\[ f^*(k) = (j + 1) B(k, j + 2) \]

The parameter \( j \) mentioned previously more or less entirely defines the distribution. For example, \( j = 0 \) gives the simplified Urn model shown previously. Generally, for given \( j > 1 \),

- \( B(1, j) \) gives the total number of successes
- \( B(1, j + 1) \) gives the population size
- \( B(k, j + 1) \) gives the number of members with at least \( k \) successes
- \( B(k, j + 2) \) gives the number of members with exactly \( k \) successes

One can adjust for the population size using constants, while \( j \) (very roughly speaking) helps to model the average number of successes per item.
“Cumulative Advantage” was later given the name “Preferential Attachment”, which is still used today. Preferential attachment processes can give rise to Power Law Distributions, where the proportion $P(k)$ of a population in state $k$ is given by

$$P(k) \sim k^{-\gamma}$$

where $\gamma$ is some parameter and for sufficiently large $k$.

Considering citation networks where members are nodes and “success” is a connection from one node to another, then this power law distribution means that a citation network forms what is called a *scale-free* network.
Notes on Citation Spaces

A citation network can be considered to have an inherent geometric structure based on the fact that they form a Directed Acyclic Graph having causal connections that are constrained by time.

Consider our little citation network model from before:
Notes on Citation Spaces

A citation network can be considered to have an inherent geometric structure based on the fact that they form a Directed Acyclic Graph having causal connections that are constrained by time.

Consider our little citation network model from before:

```
A12 ← A11 → A13
      ↖       ↗
      A        A
      ↘       ↘
A31 → A21 → A22
      ↖       ↗
      A        A
      ↘       ↘
A23 → A32
```
Notes on Citation Spaces

Clough and Evans [4] have proposed modeling these “citation spaces” with a Lorentzian manifold, wherein one dimension (time) is considered separately from the other (spacial) dimensions.

The simplest manifold of this type is Minkowski space, which is also used for special relativity.
Notes on Citation Spaces

Clough and Evans [4] have proposed modeling these “citation spaces” with a Lorentzian manifold, wherein one dimension (time) is considered separately from the other (spacial) dimensions.

The simplest manifold of this type is Minkowski space, which is also used for special relativity.

By leveraging existing methods related to modeling quantum gravity, Clough and Evans attempted to estimate the dimension of several sections of the Physics arXiv network by considering them to be embedded in a Minkowski space:

- High-Energy Theory (hep-th) 2
- High-Energy Phenomenology (hep-ph) 3
- Astrophysics (astro-ph) 3.5
- Quantum Physics (quant-ph) 3

This may be used to help differentiate between the research practices of closely-related fields, or to provide a way to define the “distance” between papers in a particular discipline.
For better or for worse, Eugene Garfield is largely responsible for how we decide which journals in the sciences are “top” and which are not. This is all possible because of the factors mentioned previously. Once you are able to quickly determine how many citations are coming in to a particular journal, you have some idea of how “influential” that journal is. Garfield introduced the notion of **impact factor** in a 1972 paper [5] as the number of times a journal has been cited divided by the number of articles it has produced.
Garfield and Impact Factor

For better or for worse, Eugene Garfield is largely responsible for how we decide which journals in the sciences are “top” and which are not. This is all possible because of the factors mentioned previously.

Once you are able to quickly determine how many citations are coming in to a particular journal, you have some idea of how “influential” that journal is.

Garfield introduced the notion of **impact factor** in a 1972 paper [5] as the number of times a journal has been cited divided by the number of articles it has produced.

\[
IF_{y,t} = \frac{\sum_{i=y}^{y+t} c_i}{\sum_{i=y}^{y+t} a_i}
\]

where \( y \) is a particular year, \( t \) is the number of years, \( c_i \) are the number of citations received, and \( a_i \) are the number of articles.
Impact Factor

When used today, the Impact Factor for a given year $y$ is

$$IF_y = \frac{c_{y-1} + c_{y-2}}{a_{y-1} + a_{y-2}}$$

where the $c_i$ are the number of citations from year $y$ to articles *in year* $i$ and the $a_i$ are the number of *citable* articles in year $i$. 
Impact Factor

When used today, the Impact Factor for a given year $y$ is

$$IF_y = \frac{c_{y-1} + c_{y-2}}{a_{y-1} + a_{y-2}}$$

where the $c_i$ are the number of citations from year $y$ to articles in year $i$ and the $a_i$ are the number of citable articles in year $i$.

The 5-year Impact Factor in a year $y$ is defined similarly:

$$IF_{5y} = \frac{c_{y-1} + c_{y-2} + \cdots + c_{y-5}}{a_{y-1} + a_{y-2} + \cdots + a_{y-5}}$$
Impact Factor

When used today, the Impact Factor for a given year $y$ is

$$IF_y = \frac{c_{y-1} + c_{y-2}}{a_{y-1} + a_{y-2}}$$

where the $c_i$ are the number of citations from year $y$ to articles in year $i$ and the $a_i$ are the number of citable articles in year $i$.

The 5-year Impact Factor in a year $y$ is defined similarly:

$$IF_{5y} = \frac{c_{y-1} + c_{y-2} + \cdots + c_{y-5}}{a_{y-1} + a_{y-2} + \cdots + a_{y-5}}$$

There is also a kind of zero case called the **Immediacy Index**:

$$II_y = \frac{c_y}{a_y}$$
## Sample Impact Factors

<table>
<thead>
<tr>
<th>Journal Title</th>
<th>IF</th>
<th>5-yr IF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annals of Mathematics</td>
<td>3.263</td>
<td>3.654</td>
</tr>
<tr>
<td>Duke Mathematical Journal</td>
<td>1.578</td>
<td>2.009</td>
</tr>
<tr>
<td>Advances in Applied Probability</td>
<td>0.709</td>
<td>0.831</td>
</tr>
<tr>
<td>Proceedings of the AMS</td>
<td>0.681</td>
<td>0.680</td>
</tr>
<tr>
<td>Algebraic &amp; Geometric Topology</td>
<td>0.445</td>
<td>0.581</td>
</tr>
<tr>
<td>Nature</td>
<td>41.456</td>
<td>41.296</td>
</tr>
<tr>
<td>JAMA</td>
<td>35.289</td>
<td>31.026</td>
</tr>
<tr>
<td>PLoS One</td>
<td>3.234</td>
<td>3.702</td>
</tr>
</tbody>
</table>

This data comes from the Journal Citation Reports (JCR) database, owned by Thomson-Reuters, who also bought the ISI and SCI in the 90s.
The Eigenfactor

Introduced by Bergstrom and West in 2007 [6], the general theory behind the **Eigenfactor score** is tracking a researcher on a “random walk” through the body of literature in a particular year.
The Eigenfactor

Introduced by Bergstrom and West in 2007 [6], the general theory behind the **Eigenfactor score** is tracking a researcher on a “random walk” through the body of literature in a particular year.
The Eigenfactor

Introduced by Bergstrom and West in 2007 [6], the general theory behind the **Eigenfactor score** is tracking a researcher on a “random walk” through the body of literature in a particular year.
The Eigenfactor

Introduced by Bergstrom and West in 2007 [6], the general theory behind the **Eigenfactor score** is tracking a researcher on a “random walk” through the body of literature in a particular year.
The Eigenfactor

Introduced by Bergstrom and West in 2007 [6], the general theory behind the **Eigenfactor score** is tracking a researcher on a “random walk” through the body of literature in a particular year.
Introduced by Bergstrom and West in 2007 [6], the general theory behind the **Eigenfactor score** is tracking a researcher on a “random walk” through the body of literature in a particular year.
Introduced by Bergstrom and West in 2007 [6], the general theory behind the **Eigenfactor score** is tracking a researcher on a “random walk” through the body of literature in a particular year.

Unfortunately, the actual algorithm is patented, so there isn’t much more to say about it!
Sample Impact Factors & Eigenfactors

<table>
<thead>
<tr>
<th>Journal Title</th>
<th>Eigenfactor*</th>
<th>IF</th>
<th>5-yr IF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annals of Mathematics</td>
<td>0.03843</td>
<td>3.263</td>
<td>3.654</td>
</tr>
<tr>
<td>Duke Mathematical Journal</td>
<td>0.02013</td>
<td>1.578</td>
<td>2.009</td>
</tr>
<tr>
<td>Advances in Applied Probability</td>
<td>0.00380</td>
<td>0.709</td>
<td>0.831</td>
</tr>
<tr>
<td>Proceedings of the AMS</td>
<td>0.03015</td>
<td>0.681</td>
<td>0.680</td>
</tr>
<tr>
<td>Algebraic &amp; Geometric Topology</td>
<td>0.00713</td>
<td>0.445</td>
<td>0.581</td>
</tr>
<tr>
<td>Nature</td>
<td>1.49869</td>
<td>41.456</td>
<td>41.296</td>
</tr>
<tr>
<td>JAMA</td>
<td>0.26099</td>
<td>35.289</td>
<td>31.026</td>
</tr>
<tr>
<td>PLoS One</td>
<td>1.53341</td>
<td>3.234</td>
<td>3.702</td>
</tr>
</tbody>
</table>

* Note that Eigenfactors are scaled so that the sum of all Eigenfactors over the Journal Citation Reports database is 100.
Hirsch and the $h$-Index

If you can rank journals based on citation data, why not rank researchers, too?

Author A's Papers

<table>
<thead>
<tr>
<th>Papers</th>
<th>Citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>105</td>
<td>16</td>
</tr>
</tbody>
</table>

$h_A = 6$

Author B's Papers

<table>
<thead>
<tr>
<th>Papers</th>
<th>Citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>83</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>13</td>
</tr>
</tbody>
</table>

$h_B = 7$
Hirsch and the $h$-Index

If you can rank journals based on citation data, why not rank researchers, too?

Hirsch introduced the $h$-index in 2005 [7], and it’s still the most common “author metric”. An author $A$ has an $h$-index of $h_A$ if:

$$h_A \text{ is the largest number such that } h_A \text{ of their articles have been cited at least } h_A \text{ times.}$$
Hirsch and the $h$-Index

If you can rank journals based on citation data, why not rank researchers, too?

Hirsch introduced the $h$-index in 2005 [7], and it’s still the most common “author metric”. An author $A$ has an $h$-index of $h_A$ if:

$$h_A \text{ is the largest number such that } h_A \text{ of their articles have been cited at least } h_A \text{ times.}$$

**Author A’s Papers**

| 15 | 3 | 10 | 7 | 105 | 16 | 7 |
Hirsch and the \( h \)-Index

If you can rank journals based on citation data, why not rank researchers, too? Hirsch introduced the \( h \)-index in 2005 [7], and it’s still the most common “author metric”. An author \( A \) has an \( h \)-index of \( h_A \) if:

\[ h_A \text{ is the largest number such that } h_A \text{ of their articles have been cited at least } h_A \text{ times.} \]

Author A’s Papers

\[
\begin{array}{ccccccc}
15 & 3 & 10 & 7 & 105 & 16 & 7
\end{array}
\]

\( h_A = 6 \)
Hirsch and the $h$-Index

If you can rank journals based on citation data, why not rank researchers, too?

Hirsch introduced the **$h$-index** in 2005 [7], and it’s still the most common “author metric”. An author $A$ has an $h$-index of $h_A$ if:

$$h_A \text{ is the largest number such that } h_A \text{ of their articles have been cited at least } h_A \text{ times.}$$

Author A’s Papers

| 15 | 3 | 10 | 7 | 105 | 16 | 7 |

$h_A = 6$

Author B’s Papers

| 7 | 16 | 97 | 17 | 83 | 9 | 13 | 5 |
Hirsch and the $h$-Index

If you can rank journals based on citation data, why not rank researchers, too?

Hirsch introduced the $h$-index in 2005 [7], and it’s still the most common “author metric”. An author $A$ has an $h$-index of $h_A$ if:

\[ h_A \text{ is the largest number such that } h_A \text{ of their articles have been cited at least } h_A \text{ times.} \]

<table>
<thead>
<tr>
<th>Author A’s Papers</th>
<th>15</th>
<th>3</th>
<th>10</th>
<th>7</th>
<th>105</th>
<th>16</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h_A$</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Author B’s Papers</th>
<th>7</th>
<th>16</th>
<th>97</th>
<th>17</th>
<th>83</th>
<th>9</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$h_B$</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Hirsch and the $h$-Index

Generally speaking, it’s easiest to order papers according to number of citations, from largest to smallest:
Hirsch and the $h$-Index

Generally speaking, it’s easiest to order papers according to number of citations, from largest to smallest:

![Graph showing $h$-index concept]

<table>
<thead>
<tr>
<th>Author A’s Papers</th>
<th>105</th>
<th>16</th>
<th>15</th>
<th>10</th>
<th>7</th>
<th>7</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author B’s Papers</td>
<td>97</td>
<td>83</td>
<td>17</td>
<td>16</td>
<td>13</td>
<td>9</td>
<td>7</td>
</tr>
</tbody>
</table>

Hirsch and the $h$-Index

Generally speaking, it’s easiest to order papers according to number of citations, from largest to smallest:

Author A’s Papers

| 105 | 16 | 15 | 10 | 7 | 7 | 3 |

Author B’s Papers

| 97 | 83 | 17 | 16 | 13 | 9 | 7 | 5 |
Variations on the $h$ index

The $h_I$-index seeks to control for differing author interests [8]:

$$h_I = \frac{h^2}{N_a(T)}$$

where $h$ is the usual $h$-index and $N_a(T)$ is the total number of authors in those $h$ papers.
Variations on the $h$ index

The $h_I$-index seeks to control for differing author interests [8]:

$$h_I = \frac{h^2}{N_a^{(T)}}$$

where $h$ is the usual $h$-index and $N_a^{(T)}$ is the total number of authors in those $h$ papers.

The $g$-index seeks to weigh highly-cited papers more heavily [9]. An author $A$ has $g$-index $g_A$ if

$g_A$ is the largest number such that a collection of $g_A$ of their articles has been cited a total of at least $(g_A)^2$ times.
Variations on the $h$ index

The **contemporary $h$-index** seeks to control for the age of an article [10]. It first defines the score $S^c(i)$ of an article $i$ as:

$$S^c(i) = \gamma \times (Y(now) - Y(i) + 1)^{-\delta} \times |C(i)|$$

where $Y(i)$ is the publication year of $i$, $C(i)$ is the set of articles citing $i$, $\delta$ is the degree to which you desire age to be a factor, and $\gamma$ is an offset. The developers of this index used $\delta = 1$ and $\gamma = 4$.

An author has contemporary $h$-index $h^c$ if

$h^c$ is the largest number such that $h^c$ of their articles have a score of $S^c(i) \geq h^c$. 

References


Photo of Eugene Garfield provided by the Chemical Heritage Foundation under a CC-BY-SA license.