

Using Bayesian networks to discover relationships between bibliometric indices. A case study of computer science and artificial intelligence journals

Alfonso Ibáñez - Pedro Larrañaga - Concha Bielza

Abstract As they are used to evaluate the importance of research at different levels by funding agencies and promotion committees, bibliometric indices have received a lot of attention from the scientific community over the last few years. Many bibliometric indices have been developed in order to take into account aspects not previously covered. The result is that, nowadays, the scientific community faces the challenge of selecting which of this pool of indices meets the required quality standards. In view of the vast number of bibliometric indices, it is necessary to analyze how they relate to each other (irrelevant, dependent and so on). Our main purpose is to learn a Bayesian network model from data to analyze the relationships among bibliometric indices. The induced Bayesian network is then used to discover probabilistic conditional (in)dependencies among the indices and, also for probabilistic reasoning. We also run a case study of 14 well-known bibliometric indices on computer science and artificial intelligence journals.

Keywords Bibliometric indices · Bayesian networks · Conditional dependencies and conditional independencies · Computer science and artificial intelligence

Introduction

Nowadays, many funding agencies and promotion committees use bibliometric indices to evaluate the impact of a researcher's work. These indices essentially involve counting the number of times scientific papers are cited. They are based on the assumption that influential researchers and important studies will be cited more frequently than others.

Bibliometric indices are used as a tool for journal evaluation (Garfield 1972a). The first bibliometric index to be calculated for journal assessment was the *impact factor* (Garfield 1972b). More recently, Braun et al. (2006) suggested that the well-known *h-index* could be usefully applied to evaluate the scientific impact of journals. Other indices like *eigenfactor* (Bergstrom et al. 2008), *article influence* (Bergstrom et al. 2008) and *Scimago's journal rank index* (SCImago 2007) have also been developed for the same purpose.

In view of the vast number of bibliometric indices, it is necessary to analyze how they relate to each other. We know that the degree of correlation among journal citation indices has been investigated in the past.

Over the last few years, some studies (Davis 2008; Bollen et al. 2009; Leydesdorff 2009) have examined correlations between a list of journal citation indices using the Pearson ρ coefficient. Davis (2008) found a statistically significant correlation between the number of *citations* and *eigenfactor* ($\rho = 0.95$). Bollen et al. (2009) also found statistically significant correlations between 39 measures of scholarly impact, although the exact values were not reported. Finally, Leydesdorff (2009) showed high correlations between indices, specially the *5 year impact factor* and *article influence* ($\rho = 0.956$).

Similarly, three recent studies (Elkins et al. 2010; Franceschet 2010; Saad in press) have also examined the degree of correlation between some typical journal citation indices, tested using Spearman's ρ . Spearman's ρ can assess how well two variables are related by any monotonic, not necessarily linear, function. Elkins et al. (2010) mentioned that all analyzed pairs of indices showed moderate to strong correlations. The strongest correlation was between *Scimago's journal rank index* and the *2year impact factor* ($\rho = 0.89$). On the other hand, Franceschet (2010) also found strong correlations between the examined indices. He found that the strongest correlation was between the *2 year impact factor* and the *5 year impact factor* ($\rho = 0.96$). Finally, Saad (in press) noticed that the *2 year impact factor* was more correlated with *article influence* than with *eigenfactor*.

To date there have not been many publications analyzing citations in computer science and artificial intelligence. Goodrum et al. (2001), though, did publish an article analyzing citations in computer science literature. The studies objective was to identify additional research areas dealing with information dissemination and citation practices in computer science. On the other hand, a recent article published by Serenko (2010) analyzed journals in the field of artificial intelligence, and calculated some bibliometric index' values (*h-index*, *g-index* and *hc-index*) which correlated almost perfectly with each other (ranging from 0.97 to 0.99).

It should be mentioned that all analyzed indices are obviously correlated since they are all derived from the number of documents and citations, and these are highly correlated.

The interest and originality of our study is that it introduces a new Bayesian network-based approach for analyzing the conditional (in)dependencies between journal citation indices. In this paper, we build some Bayesian networks (yearly and global models) to discover the relationships between 14 journal citation indices. These models are learned using journal publication and citation data (all the journals in the JCR Computer Science and Artificial Intelligence category) during the period 2000–2009, inclusive. Yearly and global models are developed to analyze index relationships within a 1-year publication and citation window. Finally, we measure how some indices influence others in probabilistic terms. Also, the network is able to perform all kinds of probabilistic reasoning, computing, say, the probability of a journal obtaining certain fixed index values given other known values.

The main advantage of our work over earlier studies, which analyze only bivariate correlations between indices, is that we calculate and analyze the joint probability

distribution over all analyzed indices, discovering probabilistic conditional (in)dependencies among triplets of indices. Journal citation indices have never been analyzed like this before.

Using our models, computer science and artificial intelligence journal editorial boards could answer some of the questions related to their journal citation indices, like, for example,

- What would happen to our journal's *impact factor* if articles, which are published by our journal, received more *citations*?
- What would happen to our journal's *h-index* if our journal published a lot of new *documents*?
- What would happen to our journal's *g-index* if our most cited studies, were the only ones to receive new citations?
- What would happen to our journal's *immediacy-index* if our journal accepted more *documents*, but received no new *citations*?

Obviously, this is a general-purpose methodology and can be used for other research areas, and, the editorial boards of any journals could find answers to the above questions.

The remainder of the study is organized as follows. The “Methods” section reviews some basic concepts about Bayesian networks and bibliometric indices on which our work is based. The “Results” section presents the dataset used, the Bayesian networks learned, the discovered probabilistic conditional (in)dependencies among the analyzed indices and examples of probabilistic reasoning. Finally, “Conclusions” contains some conclusions emphasizing the original contribution of the study and future research on the topic.

Methods

Bayesian networks

Bayesian networks are a kind of probabilistic graphical model with two main elements: the graphical component and the probabilistic component. The graphical component is a directed acyclic graph (DAG), which is used to capture the structure of the problem. The probabilistic component is the conditional probability distributions associated with the random variables (nodes in the graph) of the problem. Figure 1 illustrates an example of a Bayesian network where all the variables are binary.

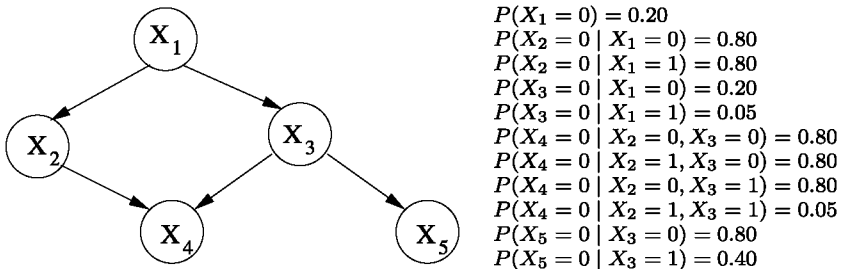


Fig. 1 Example of a Bayesian network: graphical and probabilistic components

Definition of a Bayesian network

Formally, a Bayesian network (Pearl 1988) is defined as a pair (G, P) . The first element, G , is a DAG, $G = (V(G), A(G))$, with a set of nodes given by the random variables of the problem, i.e., $V(G) = \{X_1, \dots, X_n\}$, and a set of arcs $A(G) \subseteq V(G) \times V(G)$ representing the probabilistic conditional (in)dependencies among the nodes. A variable X_i is conditionally independent of variable X_j given variable X_k iff for all x_i, x_j, x_k , $P(X_i = x_i \mid X_j = x_j, X_k = x_k) = P(X_i = x_i \mid X_k = x_k)$. Let $I(X_i, X_j \mid X_k)$ denote this relationship. The second element, P , is the joint probability distribution over (X_1, \dots, X_n) associated with G , defined as:

$$P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i \mid \Pi(X_i)), \quad (1)$$

where $\Pi(X_i)$ represents the set of parents of X_i . A node X_i is a parent of another node X_j if there is an arc from X_i to X_j .

Learning a Bayesian network from data

One of the most popular approaches for learning Bayesian networks from data is based on the score and search methodology. This approach states the learning task as an optimization problem. This involves a search for the best network structure that maximizes a scoring function defined to represent how well a structure fits a given set of data. The K2 scoring metric (Cooper and Herskovits 1992) computes the marginal likelihood of the dataset given the structure, subject to a uniform prior assumption on each variable data distribution. This scoring metric is decomposable, which facilitates the search process.

From a dataset of n variables, $\{X_1, \dots, X_n\}$, and N records, the K2 algorithm uses the marginal likelihood as score to greedily learn a Bayesian network. Starting from the empty graph and a fixed total order of variables, this algorithm adds a variable as a parent to a given variable (from the subset of variables that are before this variable in the ordering only) whenever its inclusion improves the marginal likelihood score. The algorithm stops the addition process when the marginal likelihood score decreases or the algorithm reaches the maximum admissible number of parents for each variable. This number is fixed beforehand. Given the decomposability of the score, the marginal likelihood is maximized by maximizing, for each variable X_i , the expression:

$$g(X_i, \Pi(X_i)) = \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}!, \quad (2)$$

where r_i is the number of possible values of X_i ; q_i is the number of possible values of $\Pi(X_i)$; N_{ijk} is the number of cases in the database in which variable X_i takes its k -th value and $\Pi(X_i)$ its j -th value; and N_{ij} is defined as $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$. Notice that the dataset contains all quantities in Eq. 2.

Variables in the Bayesian networks

Our Bayesian networks represent relationships between journal citation indices. In particular, each node, X_i , in the network represents a specific index, while the arcs between indices represent the conditional (in)dependencies among these indices. By learning these

Bayesian networks from data, our aim is to discover probabilistic conditional (in)dependencies among the set of 14 bibliometric indices. The indices analyzed in this study are: *documents*, *citations*, *h-index* (Hirsch 2005), *g-index* (Egghe 2006), *hg-index* (Alonso et al. 2010), *a-index* (Jin 2006), *m-index* (Bornmann et al. 2008a), *q²-index* (Cabrerizo et al. 2010), *r-index* (Jin et al. 2007), *e-index* (Zhang 2009), *w-index* (Woeginger 2008), *rational h-index* (Ruane and Tol 2008), *impact factor* (Garfield 1972b) and *immediacy-index*. All these indices were obtained from the information provided by the Institute for Scientific Information (<http://isiwebofknowledge.com/>).

Bibliometric indices

Bibliometric indices are quantitative metrics for evaluating and comparing the research activity of individual scientists according to their output. The main advantage of these indices is that they can summarize the scientific production of an author as a single number. At the same time, this advantage can be a limitation, because it removes many details of citation records.

We have evaluated the research activity of journals using bibliometric indices such as the ones listed in “[Variables in the Bayesian networks](#)”. These indices, which were originally developed to evaluate the quality of a researcher’s work, have been adapted to assess journals. In the following, we show how the indices were adapted.

Documents

Documents is an index associated with the number of articles published by each analyzed journal in the entire period. This index represents the productivity of each specific journal.

Citations

Citations is an index associated with the number of citations received by each journal analyzed in the entire period. This index represents the visibility of each specific journal.

The h-index

Over the last few years, the *h-index*, proposed by Hirsch (2005), has received a lot of attention from the scientific community because it combines the productivity and visibility of a scientist in a single indicator. The *h-index* is defined as a number such that, for the all the articles published in a journal, *h* papers received at least *h* citations whereas the other studies received no more than *h* citations. Braun et al. (2006) applied the *h-index* to journals:

A journal has index *h* if *h* of its published articles have at least *h* citations each, and the other papers have no more than *h* citations each.

The g-index

Since the *h-index* tends to underestimate the achievement of journals that have a “selective publication strategy”, that is, journals that do not publish a lot of documents but have a major international impact (Egghe 2006). The *g-index* is defined as the highest rank such

that the cumulative sum of the number of citations received is larger than or equal to the square of this rank.

The highest number g of papers that together received g^2 or more citations.

Unlike the h -index, the g -index takes into account the exact number of citations received by highly cited articles, favoring journals with a selective publication strategy.

The hg-index

Alonso et al. (2010) presented a new index called the hg -index, which is based on the h -index and the g -index. It fuses both measures to obtain a more balanced view of the scientific production of journals to minimize some of their weaknesses. The hg -index of a journal is computed as the geometric mean of its h -index and g -index, that is,

$$hg\text{-index} = \sqrt{h \cdot g},$$

where h is the value of the h -index, and g is the value of the g -index.

The a-index

The a -index was proposed by Jin (2006). This index is calculated for papers that are in the Hirsch core (h -core) only, that is, the first h papers. It is defined as the average number of citations received by the articles included in the h -core. This index measures the citation intensity in the h -core. The a -index can be very sensitive to just a very few studies receiving extremely high citation counts. Mathematically, this is:

$$a\text{-index} = \frac{1}{h} \sum_{i=1}^h \text{Cit}(i),$$

where h is the value of the h -index; and $\text{Cit}(i)$ is the number of citations received by article i belonging to the h -core.

The m-index

As the distribution of citation counts is usually skewed, the median and not the arithmetic mean should be used as the measure of central tendency. Therefore, Bornmann et al. (2008a) proposed a new index, called m -index, as a variation of the a -index. This index, which was designed to illustrate the impact of the papers in the h -core, is the median number of citations received by papers in the h -core.

The q²-index

Cabrerizo et al. (2010) developed a new index, called q^2 -index, to provide a more global view of scientific production. This index is based on the geometric mean of the h -index, describing the number of the papers (quantitative dimension), and the m -index, depicting the impact of the papers (qualitative dimension), that is,

$$q^2\text{-index} = \sqrt{h \cdot m},$$

where h is the value of the h -index, and m is the value of the m -index.

The *r-index*

Some problems related to the *a-index* were overcome in Jin et al. (2007) by another index. Unlike the *a-index*, which involves a division by the *h-index*, this index does not punish journals for having a higher *h-index* value. Instead of dividing by the *h-index*, this index takes the square root of the sum of citations in the *h-core* to calculate the final value. As a mathematical formula the *r-index* is defined as

$$r\text{-index} = \sqrt{\sum_{i=1}^h \text{Cit}(i)},$$

where h is the value of the *h-index*, and $\text{Cit}(i)$ is the number of citations received by article i belonging to the *h-core*.

The *e-index*

The *e-index* was developed by Zhang (2009) to solve a problem with the *h-index*: excess citations, which are not taken into account for calculating the *h-index*, are completely ignored. The *e-index* is a complement to the *h-index* and it represents the excess citations received by all studies in the *h-core*. As a mathematical formula the *e-index* is defined as:

$$e\text{-index} = \sqrt{\sum_{i=1}^h (\text{Cit}(i) - h)},$$

where h is the value of the *h-index*, and $\text{Cit}(i)$ is the number of citations received by article i belonging to the *h-core*.

The *w-index*

The *w-index* was developed by Woeginger (2008) and is defined as follows:

A *w-index* of at least k means that there are k distinct publications that have at least 1, 2, 3, 4, ..., k citations, respectively.

Like the *h-index*, the *w-index* tends to cluster many journals into the same index value. However, its range could be up to twice the range of the *h-index*. Therefore, the *w-index* should lead to a somewhat finer ranking than the *h-index*.

The rational *h-index*

Ruane and Tol (2008) proposed the *rational h-index*, which is an extension of the original *h-index*. This index takes into account the number of citations needed to increase the *h-index* by one unit. It measures the distance to the next value of the *h-index*. Mathematically, this is

$$\text{rational } h\text{-index} = (h + 1) - \frac{\text{Cit}(h + 1)}{2h + 1},$$

where h is the value of the *h-index*, and $\text{Cit}(h+1)$ is the number of citations received by article $h+1$.

The impact factor

The *impact factor (IF)* for a given journal in the year y is the average number of times the articles that it published in the past 2 years were cited in year y . The *impact factor* is calculated by dividing the number of citations during year y by the total number of articles published by the journal in the previous 2 years. Mathematically, this is

$$IF(y) = \frac{\text{Cites in year } (y) \text{ to items published in years } (y-1) \text{ and } (y-2)}{\text{Number of items published in years } (y-1) \text{ and } (y-2)}$$

The immediacy-index

The *immediacy-index* is the average number of times an article is cited in the year that it is published. This index indicates how quickly articles in a journal are cited. It is calculated by dividing the number of citations to articles published in a given year by the number of articles published in that year. Mathematically, this is

$$\text{immediacy-index } (y) = \frac{\text{Cites in year } (y) \text{ to items published in year } (y)}{\text{Number of items published in year } (y)}$$

Results

Data collection

In this study, we have selected the Computer Science and Artificial Intelligence field as a case study. We have used Thomson Reuters' Web of Knowledge platform to download publication and citation data. In the following, we illustrate the different phases of dataset construction.

In the first step, we collected journal data from the JCR's Computer Science and Artificial Intelligence category. There are 94 journals in this category of the 2008 JCR Science Edition. In view of the objective of this study, we only took into account the 70 journals (Table 1) that published papers from January 1, 2000 to December 31, 2009. We stored this information in a database designed for this purpose.

The next step was to obtain the publication list and citation data for these journals. This information was downloaded from the Web of Science (WoS), hosted by the Web of Knowledge platform.

Finally, the last step was to use all the information stored in our database to calculate some scientific impact indices associated with the selected journals. These indices are: *documents*, *citations*, the *h-index*, the *g-index*, the *hg-index*, the *a-index*, the *m-index*, the *q²-index*, the *r-index*, the *e-index*, the *w-index*, the *rational h-index*, the *impact factor* and the *immediacy-index* (all described in "Bibliometric indices" section). These index values have been calculated yearly for each of the 70 journals in the ten-year period from 2000 to 2009.

To illustrate some of the calculated index values, Table 2 lists some of the journals ranked top according to six selected indices: *documents*, *citations*, the *h-index*, the *g-index*, the *impact factor* and the *immediacy-index*. Table 2 shows some rankings obtained using data for a one-year publication window, specifically for 2009. These rankings reveal that some journals are always positioned near to the top. Taking *IEEE Transactions on Pattern*

Table 1 List of journals in JCR Computer Science and Artificial Intelligence category that have papers every year throughout the 2000–2009 period

Journals
Adaptive Behavior
AI Communications
AI Edam-Artificial Intelligence for Engineering Design Analysis and Manufacturing
AI Magazine
Annals of Mathematics and Artificial Intelligence
Applied Artificial Intelligence
Applied Intelligence
Artificial Intelligence
Artificial Intelligence in Medicine
Artificial Intelligence Review
Artificial Life
Autonomous Agents and Multi-Agent Systems
Autonomous Robots
Chemometrics and Intelligent Laboratory Systems
Computational Intelligence
Computer Speech and Language
Computer Vision and Image Understanding
Connection Science
Data and Knowledge Engineering
Data Mining and Knowledge Discovery
Decision Support Systems
Engineering Applications of Artificial Intelligence
Engineering Intelligent Systems for Electrical Engineering and Communications
Expert Systems
Expert Systems with Applications
IEEE Transactions on Evolutionary Computation
IEEE Transactions on Fuzzy Systems
IEEE Transactions on Image Processing
IEEE Transactions on Knowledge and Data Engineering
IEEE Transactions on Neural Networks
IEEE Transactions on Pattern Analysis and Machine Intelligence
IEEE Transactions on Systems, Man and Cybernetics Part B-Cybernetics
IEEE Transactions on Systems, Man and Cybernetics Part C-Applications and Reviews
Image and Vision Computing
International Journal of Approximate Reasoning
International Journal of Computer Vision
International Journal of Intelligent Systems
International Journal of Patter Recognition and Artificial Intelligence
International Journal of Software Engineering and Knowledge Engineering
International Journal of Uncertainty Fuzziness and Knowledge-Based Systems
Integrated Computer-Aided Engineering
Intelligent Automation and Soft Computing
Journal of Artificial Intelligent Research

Table 1 continued

Journals
Journal of Automated Reasoning
Journal of Chemometrics
Journal of Computer and Systems Sciences International
Journal of Experimental and Theoretical Artificial Intelligence
Journal of Heuristics
Journal of Intelligent and Fuzzy Systems
Journal of Intelligent Information Systems
Journal of Intelligent Manufacturing
Journal of Intelligent and Robotic Systems
Journal of Mathematical Imaging and Vision
Knowledge Engineering Review
Knowledge-Based Systems
Machine Learning
Machine Vision and Applications
Mechatronics
Medical Image Analysis
Minds and Machines
Network-Computation in Neural Systems
Neural Computation
Neural Computing and Applications
Neural Networks
Neural Processing Letters
Neurocomputing
Pattern Analysis and Applications
Pattern Recognition
Pattern Recognition Letters
Robotics and Autonomous Systems

Analysis and Machine Intelligence as an example, we find that this journal published 188 articles in 2009, which received 54 citations in the same year. Furthermore, its *h-index*' value and *g-index*' value were 3 and 4, respectively. Finally, it had an *impact factor* value of 5.960 and an *immediacy-index* value of 0.669. Remember that all these values were obtained using a 2009-year publication window.

We also list some index values calculated in other years. Table 3 illustrates the range of index values in each analyzed year. Taking the *documents* value for the year 2000 as an example, we find that the minimum number of documents published by a specific journal in the year 2000 was 10 and the maximum was 219.

Analyzing Table 3, we find that index values are higher for recent than older years. Specifically, most of the highest values for each index are obtained between 2007 and 2009. Although we note that the values of all indices tend to increase, there is no index whose value increases year by year.

We notice that the highest values differ greatly depending on the selected index and year. On the one hand, the values of indices like *documents* or *citations* have undergone a more significant increase than other indices over the analyzed years. The number of

Table 2 Top five positions of the journal rankings using a 2009-year publication and citation window according to six bibliometric indices

Position	Journal	<i>documents</i>
1	Expert Systems with Applications	1399
2	Neurocomputing	352
3	Pattern Recognition	312
4	IEEE Transactions on Image Processing	217
5	IEEE Transactions on Pattern Analysis and Machine Intelligence	188
Position	Journal	<i>citations</i>
1	Expert Systems with Applications	402
2	Neurocomputing	68
3	IEEE Transactions on Pattern Analysis and Machine Intelligence	54
4	Neural Networks	49
5	Pattern Recognition	45
Position	Journal	<i>h-index</i>
1	Expert Systems with Applications	5
2	Neurocomputing	4
3	Neural Networks	4
4	IEEE Transactions on Pattern Analysis and Machine Intelligence	3
5	Image and Vision Computing	2
Position	Journal	<i>g-index</i>
1	Expert Systems with Applications	6
2	Neurocomputing	5
3	Neural Networks	4
4	IEEE Transactions on Pattern Analysis and Machine Intelligence	4
5	IEEE Transactions on Knowledge and Data Engineering	3
Position	Journal	<i>impact factor</i>
1	IEEE Transactions on Pattern Analysis and Machine Intelligence	5.960
2	International Journal of Computer Vision	5.358
3	IEEE Transactions on Evolutionary Computation	3.736
4	IEEE Transactions on Neural Networks	3.726
5	IEEE Transactions on Fuzzy Systems	3.624
Position	Journal	<i>immediacy-index</i>
1	Computational Intelligence	1.091
2	IEEE Transactions on Pattern Analysis and Machine Intelligence	0.669
3	Artificial Intelligence	0.667
4	International Journal of Computer Vision	0.659
5	Journal of Automated Reasoning	0.600

Table 3 Range of index values in each analyzed year. *Numbers in boldface* represent the maximum value for each index in the 2000–2009 period

Index	2000		2001		2002		2003		2004	
	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max
<i>documents</i>	10	219	12	257	8	308	15	297	12	295
<i>citations</i>	0	45	0	47	0	39	0	78	0	48
<i>h-index</i>	0	3	0	2	0	3	0	4	0	3
<i>g-index</i>	0	3	0	4	0	3	0	4	0	4
<i>hg-index</i>	0.0	3.0	0.0	2.8	0.0	3.0	0.0	4.0	0.0	3.5
<i>a-index</i>	0.0	5.0	0.0	6.5	0.0	4.0	0.0	5.0	0.0	6.0
<i>m-index</i>	0	3	0	4	0	3	0	4	0	3
<i>q²-index</i>	0.0	3.0	0.0	2.8	0.0	3.0	0.0	4.0	0.0	3.0
<i>r-index</i>	0.0	3.5	0.0	3.6	0.0	3.5	0.0	4.5	0.0	3.9
<i>e-index</i>	0.0	2.0	0.0	3.0	0.0	1.7	0.0	2.4	0.0	2.6
<i>w-index</i>	0	4	0	4	0	5	0	6	0	5
<i>rational h-index</i>	0.0	3.6	0.0	2.8	0.0	3.7	0.0	4.8	0.0	3.6
<i>impact factor</i>	0.0	2.8	0.0	2.7	0.0	2.7	0.0	2.9	0.0	4.4
<i>immediacy-index</i>	0.0	0.7	0.0	1.0	0.0	0.7	0.0	0.9	0.0	0.7

Index	2005		2006		2007		2008		2009	
	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max
<i>documents</i>	15	255	18	335	16	342	3	523	3	1399
<i>citations</i>	0	86	0	81	0	73	0	164	0	402
<i>h-index</i>	0	4	0	4	0	3	0	4	0	5
<i>g-index</i>	0	5	0	4	0	5	0	5	0	6
<i>hg-index</i>	0.0	4.5	0.0	4.0	0.0	3.9	0.0	4.5	0.0	5.0
<i>a-index</i>	0.0	6.0	0.0	5.0	0.0	6.7	0.0	6.5	0.0	6.6
<i>m-index</i>	0	5ePara>	0	4	0	4	0	5	0	6
<i>q²-index</i>	0.0	4.5	0.0	4.0	0.0	3.5	0.0	4.5	0.0	5.5
<i>r-index</i>	0.0	4.9	0.0	4.2	0.0	4.5	0.0	4.8	0.0	5.7
<i>e-index</i>	0.0	2.8	0.0	2.4	0.0	3.3	0.0	3.0	0.0	2.8
<i>w-index</i>	0	7	0	5	0	6	0	7	0	8
<i>rational h-index</i>	0.0	4.9	0.0	4.4	0.0	3.9	0.0	4.9	0.0	5.8
<i>impact factor</i>	0.1	4.3	0.0	3.8	0.1	6.1	0.0	3.6	0.0	6.0
<i>immediacy-index</i>	0.0	1.0	0.0	1.0	0.0	1.2	0.0	0.8	0.0	1.1

documents and *citations* incremented sharply in the time period. In fact, they are 6 and 9 times greater, respectively, in the last year than in the first year. On the other hand, the values of other indices (*h-index*, *g-index*, *m-index*, *w-index*, *rational h-index*, *impact factor*) have not increased as significantly as *documents* and *citations*. In the most recent years we find them to be approximately 2 times greater than in the early years. For example, the *g-index* has a value of 3 in the year 2000 and a value of 6 in 2009. In the same way, the *impact factor* has a value of 2.8 in the year 2000 and a value of 6.0 in 2009. Finally, the values of the other indices (*hg-index*, *a-index*, *q²-index*, *r-index*, *e-index*, *immediacy-index*) have also increased within the time period, but to a lesser extent than the other indices. For

example, the *a-index* has a value of 5.0 in the year 2000 and a value of 6.6 in 2009. Similarly, the *immediacy-index* has a value of 0.7 in the year 2000 and a value of 1.1 in 2009.

Yearly models

Ten yearly Bayesian networks models were learned to analyze the relationships among indices within the same 1-year publication window. For this reason, each yearly model is associated with one of the ten analyzed years and with one of the ten datasets. Each dataset contains 70 cases (journals), each one with its 14 index values.

Before running the K2 algorithm to learn a Bayesian network from each dataset, we had to make some decisions about three of K2's requirements. Since K2 needs the variables to be ordered, the first decision was to specify an order. Taking into account the index definitions, we placed indices that could be parents of the other indices first. The established order was: *documents*, *citations*, *h-index*, *g-index*, *hg-index*, *a-index*, *m-index*, *q²-index*, *r-index*, *e-index*, *w-index*, *rational h-index*, *impact factor* and *immediacy-index*. We obtained high values of marginal likelihood using this order. The second requirement was to assign a value to the maximum number of parents. This was set at two due to the dataset characteristics. There was a third requirement: index values had to be discretized into intervals. Due to the number of dataset cases, we discretized the values into three intervals with equal frequency. In this way, the index values were assigned to one of the three possible values (*low*, *medium* and *high*).

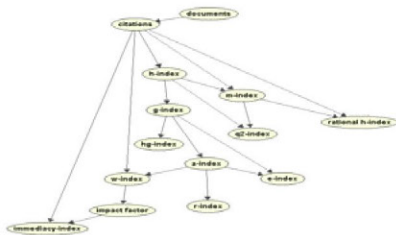
In the following, we represent the structures of the learned yearly Bayesian networks and then explain some probabilistic conditional (in)dependencies between the indices that they discovered.

Bayesian networks structure

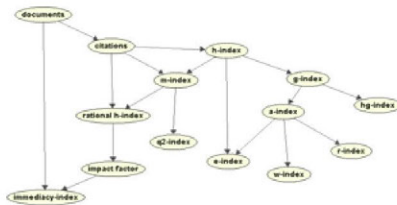
In this section we present the structure of our yearly Bayesian networks, see Fig. 2. We observe that there are a lot of coincident arcs in our Bayesian networks as shown in Table 4. Taking the value of the first-row and second-column as an example, the value displayed is 14. This value indicates that the Bayesian network for the year 2000 and the Bayesian network for the year 2001 have 14 identical arcs. In other words, the Bayesian network of the year 2000 has 19 arcs, and 14 of these arcs are also represented in the Bayesian network of the year 2001. Examining Table 4, we find that the relationships between indices are very similar in each year of the analyzed period. They are described in further detail in the following section.

Analyzing the networks in Fig. 2, we notice that there are some specific arcs that are represented in most of the networks. For example, the arcs *h-index* \rightarrow *g-index*, *documents* \rightarrow *citations*, *citations* \rightarrow *h-index*, *a-index* \rightarrow *e-index*, among others, always appear in our 10 Bayesian networks. The number of times that an arc is shown in our Bayesian networks is reported in Table 5. Taking the value of the *citations* \rightarrow *m-index* arc as an example, the value displayed is 9. This value means that the relationship between *citations* and *m-index* is present in 9 out of our 10 yearly Bayesian networks.

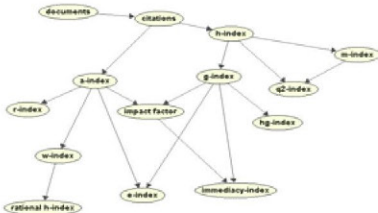
With the intention of representing the main relationships between indices in a single year-independent Bayesian network, we built an aggregated Bayesian network using only those arcs that had appeared at least 3 times. After applying the above filter, we obtained the aggregated Bayesian network shown in Fig. 3. The values above the arcs represent the number of times that the arc appeared in our 10 yearly Bayesian networks.



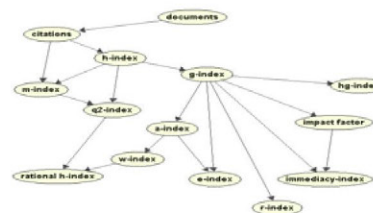
(a) Year 2000



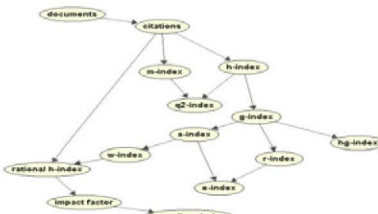
(b) Year 2001



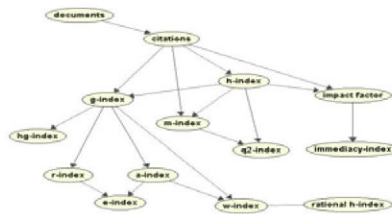
(c) Year 2002



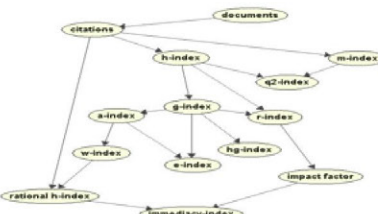
(d) Year 2003



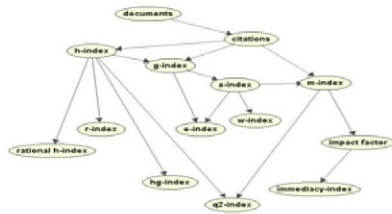
(e) Year 2004



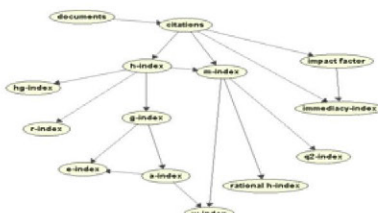
(f) Year 2005



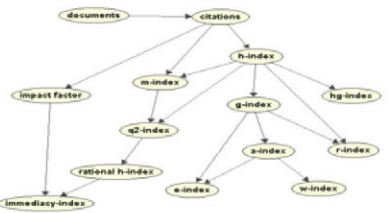
(g) Year 2006



(h) Year 2007



(i) Year 2008



(j) Year 2009

Fig. 2 Bayesian network structures learned for each analyzed year

Table 4 Number of coincident arcs in the 10 different networks

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
2000 (19 arcs)	–	14	12	13	12	12	13	11	13	12
2001 (18 arcs)	–	–	10	11	12	11	11	9	11	10
2002 (17 arcs)	–	–	–	14	10	11	11	9	9	10
2003 (18 arcs)	–	–	–	–	13	14	14	11	11	14
2004 (16 arcs)	–	–	–	–	–	14	14	10	9	11
2005 (19 arcs)	–	–	–	–	–	–	13	11	11	13
2006 (18 arcs)	–	–	–	–	–	–	–	12	11	14
2007 (17 arcs)	–	–	–	–	–	–	–	–	12	13
2008 (17 arcs)	–	–	–	–	–	–	–	–	–	14
2009 (18 arcs)	–	–	–	–	–	–	–	–	–	–

Dependencies among indices in the aggregated Bayesian network

Examining the index definitions, we note that some of them can be defined according to the values of other indices. For example, on the one hand, the *hg-index* can be expressed in terms of *h-* and *g-index*’s values ($hg\text{-index} = \sqrt{h \cdot g}$) and, on the other hand, the $q^2\text{-index}$ can be defined according to *h-* and *m-index*’s values ($q^2\text{-index} = \sqrt{h \cdot m}$).

Although the deterministic aspect between index definitions is reduced after discretizing the index values into three intervals (*low*, *medium* and *high*), our aggregated Bayesian network discovers such dependencies between indices.

Looking at Fig. 3, we find that the dependencies that can be defined according to the values of other indices are represented in the aggregated Bayesian network. For example, the *hg-index* (see “The *hg-index*” section) is represented in the network since the *h-index* and the *g-index* are parents of the *hg-index* in our network. Similarly, the $q^2\text{-index}$ as a function of the *h-index* and the *m-index* (see “The $q^2\text{-index}$ ” section) is also represented in the network.

Some dependencies, like *documents* \rightarrow *citations* and *h-index* \rightarrow *g-index*, among others, are not derived from the index definition, but were expected because many works showed their correlations (Costas and Bordons 2008; Schreiber 2008). Other dependencies, e.g., *citations* \rightarrow *h-index* and *citations* \rightarrow *impact factor* are also represented in the aggregated Bayesian network. We noticed that although the *h-index* and the *impact factor* cannot be defined in terms of *citations* values only, they do exhibit a high value correlation coefficient (Bornmann et al. 2008b).

Other dependencies, e.g., the arc between *a-index* and *e-index*, is an example of a dependency that was not initially expected. Remember that the *a-index* represents the average number of citations received by the articles included in the *h-core*, whereas the *e-index* represents the excess citations received by the articles in the *h-core*. Thus, both refer to citations of articles in the *h-core*. More examples of such dependencies are: *m-index* \rightarrow *rational h-index*, *a-index* \rightarrow *w-index*, and *w-index* \rightarrow *rational h-index* among others.

Conditional independencies among indices in the aggregated Bayesian network

Bayesian networks are powerful tools not only for capturing dependencies but also for encoding conditional independencies among our indices. In the following, we explain some

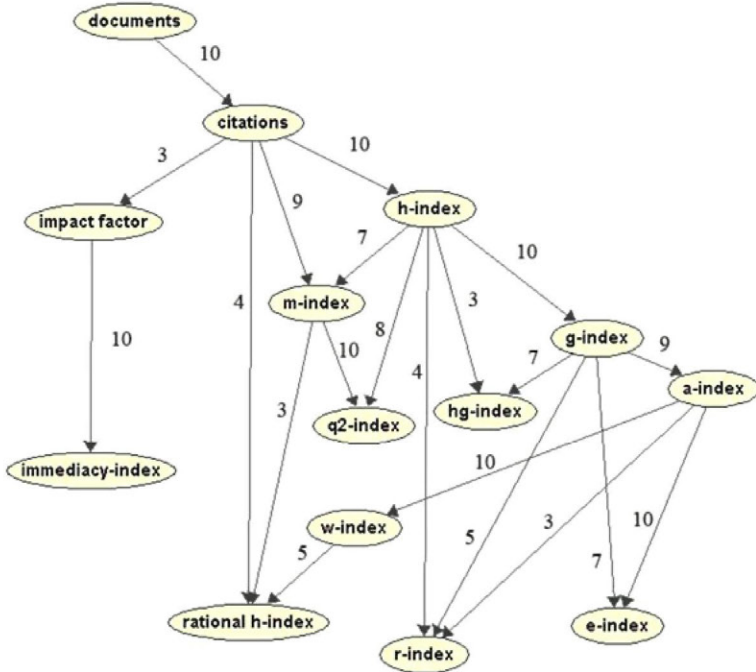


Fig. 3 Aggregated Bayesian network structure

properties for discovering conditional independencies among the analyzed indices. We choose Markov properties as the criteria for this purpose.

The local Markov property states that any node X in any Bayesian network is conditionally independent of its *non-descendants* given its *parents*, that is, $I(X, \text{non-descendants}(X) \mid \Pi(X))$. The global Markov property is also used for discovering conditional independencies among the analyzed indices. This property states that any node X is conditionally independent of any other node given its Markov blanket (MB). The MB of a node includes its parents, its children, and its children's parents, that is, $I(X, \text{non-MB}(X) \mid \text{MB}(X))$.

Table 6 illustrates such relationships. Although this table shows a specific list of relationships between indices, new relationships can be derived using some conditional independencies properties (Castillo et al. 1997):

- Symmetry: $I(X, Y \mid Z) \Leftrightarrow I(Y, X \mid Z)$
- Decomposition: $I(X, (Y \cup W) \mid Z) \Rightarrow I(X, Y \mid Z) \text{ and } I(X, W \mid Z)$
- Strong joint: $I(X, Y \mid Z) \Rightarrow I(X, Y \mid (Z \cup W))$

Taking the q^2 -index as an example, we find in Table 6 that, given the h -index and the m -index together, the q^2 -index is independent of most of the indices (*documents*, *citations*, the hg -index, the a -index, the r -index, the e -index, the w -index, the *rational h-index*, the *impact factor* and the *immediacy-index*). On the other hand, we observe that the g -index is independent of the q^2 -index given the h -index. Taking into account the above independency relationships and the conditional independency properties, we state that given the h -index and the m -index together, the q^2 -index is independent of any of the other indices.

Table 6 Conditional independencies among indices, derived using Markov properties in the aggregated Bayesian network

Index	Is conditionally independent of	Given
<i>documents</i>	<i>h-index, g-index, hg-index, a-index, m-index, q^2-index, r-index, e-index, w-index, rational h-index, impact factor, immediacy-index</i>	<i>citations</i>
<i>citations</i>	<i>hg-index, a-index, r-index, e-index, immediacy-index</i>	<i>documents, h-index, m-index, w-index, rational h-index, impact factor</i>
<i>h-index</i>	<i>immediacy-index</i>	<i>citations</i>
<i>h-index</i>	<i>rational h-index</i>	<i>citations, g-index, hg-index, a-index, m-index, q^2-index, r-index</i>
<i>g-index</i>	<i>documents, citations, m-index, q^2-index, impact factor, immediacy-index</i>	<i>h-index</i>
<i>g-index</i>	<i>rational h-index</i>	<i>h-index, hg-index, e-index, a-index, r-index</i>
<i>hg-index</i>	<i>documents, citations, m-index, q^2-index, r-index, e-index, w-index, rational h-index, impact factor, immediacy-index</i>	<i>h-index, g-index</i>
<i>a-index</i>	<i>documents, citations, h-index, hg-index, m-index, q^2-index, impact factor, immediacy-index</i>	<i>g-index</i>
<i>a-index</i>	<i>rational h-index</i>	<i>h-index, g-index, r-index, e-index, w-index</i>
<i>m-index</i>	<i>hg-index, a-index, r-index, e-index, w-index, immediacy-index</i>	<i>citations, h-index</i>
<i>q^2-index</i>	<i>documents, citations, hg-index, a-index, r-index, e-index, w-index, rational h-index, impact factor, immediacy-index</i>	<i>h-index, m-index</i>
<i>r-index</i>	<i>documents, citations, m-index, q^2-index, rational h-index, impact factor, immediacy-index</i>	<i>h-index, g-index, a-index</i>
<i>e-index</i>	<i>documents, citations, h-index, hg-index, m-index, q^2-index, r-index, rational h-index, impact factor, immediacy-index</i>	<i>g-index, a-index</i>
<i>w-index</i>	<i>documents, citations, h-index, g-index, hg-index, m-index, q^2-index, r-index, e-index, impact factor, immediacy-index</i>	<i>a-index</i>
<i>rational h-index</i>	<i>h-index, g-index, hg-index, a-index, q^2-index, r-index, e-index, immediacy-index</i>	<i>citations, m-index, w-index</i>
<i>impact factor</i>	<i>h-index, g-index, hg-index, a-index, m-index, q^2-index, r-index, e-index, w-index, rational h-index</i>	<i>citations</i>

Table 6 continued

Index	Is conditionally independent of	Given
<i>immediacy-index</i>	<i>documents, citations, h-index, g-index, hg-index, a-index, m-index, q²-index, r-index, e-index, w-index, rational h-index</i>	<i>impact factor</i>

This means that when we know the *h-index* and the *m-index* values, knowledge of the *e-index*, for example, provides no information on the occurrence of the *q²-index*. Similarly, we note that given the *h-index* and the *g-index* together, the *hg-index* is independent of any of the other indices.

The above conditional independencies between indices can be checked by analyzing the index definitions. Our aggregated Bayesian network is able to encode the conditional independencies that are derived from the index definition, but also discovers other new conditional independencies that are not strictly derived from definitions.

On the one hand, we expected the *m-index*, which is the median number of citations received by papers in the *h-core*, to be conditionally independent of some variables given *citations* and *h-index*. Table 6 shows that given *citations* and *h-index*, the *m-index* is conditionally independent of *hg-index*, *a-index*, *r-index*, *e-index*, *w-index* and *immediacy-index*.

On the other hand, some other conditional independencies were not so obvious. For example, the *immediacy-index*, which is defined by means of *documents*' values and *citations*' values, is independent of *documents* and *citations* given the *impact factor*. The presence of *impact factors* values has a significant influence on *immediacy-index* values. This means that when we know the *impact factor*, knowledge of *documents* and *citations* does not provide any information on the occurrence of the *immediacy-index*.

Remember that the conditional independencies between indices encoded in our Bayesian networks do not represent a causality relationship, but refer to a probability relationship between indices.

Global model

The global model's objective is to analyze the relationships between indices within a one-year target window. It has the same as the yearly models, but, unlike them, this global model is built using a different dataset.

We select and merge the previous datasets into a single dataset to build the dataset for the global model. In this way, the global model dataset contains 700 cases. Each case in the dataset was referred to index values calculated within a 1-year target window, and, consequently, they all cases can be easily merged into a single dataset.

We have also used the K2 scoring metric to learn our Bayesian network. Although we have established the same variables order, we have slightly modified some decisions about K2's requirements. As there are more cases we can increase the maximum number of parents for any node up to 3 and the number of intervals for the index discrete domain to 4 (*low*, *medium-low*, *medium-high* and *high*).

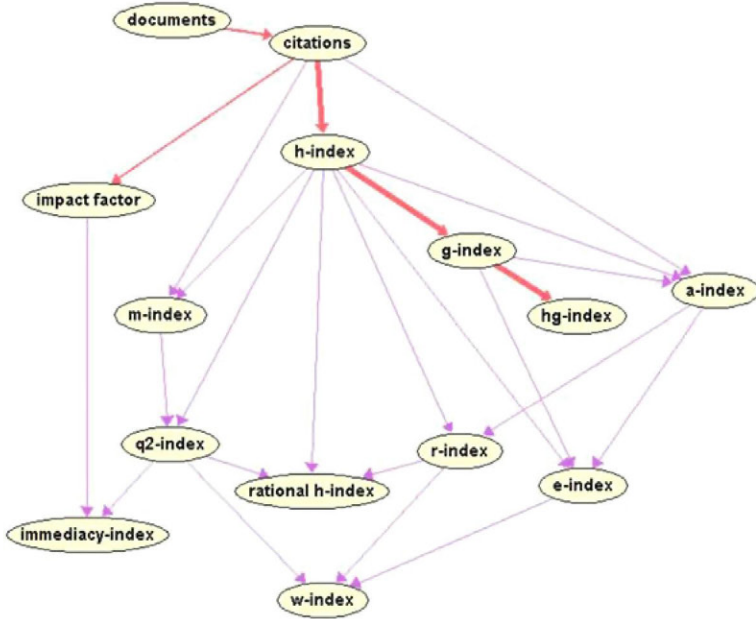


Fig. 4 Global Bayesian network structure

Bayesian network structure

Figure 4 shows our global Bayesian network structure.

Comparing the arcs in Fig. 3 (aggregated Bayesian network structure) and Fig. 4 (global Bayesian network structure), we find that the network structures are similar. There are many arcs that appear in both networks, although this new model includes some specific arcs not represented before: *citations* \rightarrow *a-index*, *h-index* \rightarrow *a-index*, *h-index* \rightarrow *e-index*, *h-index* \rightarrow *rational h-index*, among others. For this reason, we believe that the global model more accurately represents the index definition than the aggregated model. We explain these new dependencies in the following.

We examined some centrality measures in order to analyze some of the index characteristics in our global Bayesian network. These values are shown in Table 7.

Degree centrality is defined as the number of arcs incident upon an index. Degree is often interpreted in terms of the opportunity for influencing any other index. We define two separate measures of degree centrality (indegree and outdegree). A node's indegree is the number of arcs directed to the node, and outdegree is the number of arcs that the node directs to others. Therefore, indegree is the number of parents, whereas outdegree is the number of children.

Examining Table 7, we note that the *h-index* has a lot of influence on other indices. It has the highest degree centrality value ($1 + 7 = 8$). On the other hand, indices like *documents* or *hg-index*, which have a degree centrality of 1, do not influence the other indices so much.

Finally, Table 8 displays the range of index values and the values assigned to each interval after the discretization to illustrate the global model dataset. We observe that the interval width is not the same for the four categories since the dataset has been discretized

Table 7 Measures of centrality (indegree and outdegree) for the 14 analyzed indices in the global Bayesian network

Index	Indegree	Outdegree
<i>documents</i>	0	1
<i>citations</i>	1	4
<i>h-index</i>	1	7
<i>g-index</i>	1	3
<i>hg-index</i>	1	0
<i>a-index</i>	3	2
<i>m-index</i>	2	1
<i>q²-index</i>	2	3
<i>r-index</i>	2	2
<i>e-index</i>	3	1
<i>w-index</i>	3	0
<i>rational h-index</i>	3	0
<i>impact factor</i>	1	1
<i>immediacy-index</i>	2	0

Table 8 Range of index values related to each interval of the global model dataset

Indices	Values' range	<i>low</i>	<i>medium-low</i>	<i>medium-high</i>	<i>high</i>
<i>documents</i>	[3, 1399]	[3, 29)	[29, 47)	[47, 79)	[79, 1399]
<i>citations</i>	[0, 402]	[0, 1)	[1, 3)	[3, 12)	[12, 402]
<i>h-index</i>	[0, 5]	[0, 1)	[1, 2)	[2, 3)	[3, 5]
<i>g-index</i>	[0, 6]	[0, 1)	[1, 2)	[2, 3)	[3, 6]
<i>hg-index</i>	[0.0, 5.5]	[0.0, 1.0)	[1.0, 1.4)	[1.4, 2.4)	[2.4, 5.5]
<i>a-index</i>	[0.0, 6.7]	[0.0, 1.0)	[1.0, 2.0)	[2.0, 3.0)	[3.0, 6.7]
<i>m-index</i>	[0, 6]	[0.0, 1.0)	[1.0, 2.0)	[2.0, 3.0)	[3.0, 6]
<i>q²-index</i>	[0.0, 5.5]	[0.0, 1.0)	[1.0, 1.4)	[1.4, 2.0)	[2.0, 5.5]
<i>r-index</i>	[0.0, 5.7]	[0.0, 1.0)	[1.0, 1.4)	[1.4, 2.4)	[2.4, 5.7]
<i>e-index</i>	[0.0, 3.3]	[0.0, 1.0)	[1.0, 1.4)	[1.4, 1.7)	[1.7, 3.3]
<i>w-index</i>	[0, 8]	[0.0, 1.0)	[1, 2)	[2, 3)	[3, 8]
<i>rational h-index</i>	[0.0, 5.8]	[0.0, 1.0)	[1.0, 1.3)	[1.3, 2.0)	[2.0, 5.8]
<i>impact factor</i>	[0.0, 6.1]	[0.0, 0.4)	[0.4, 0.8)	[0.8, 1.5)	[1.5, 6.1]
<i>immediacy-index</i>	[0.0, 1.2]	[0.0, 0.0]	(0.0, 0.1]	(0.1, 0.2]	(0.2, 1.2]

in intervals of equal frequency. A journal publishes a number of between 3 and 1399 *documents* per year, and these journals can then be categorized according to the index values. For example, a journal that has published 43 documents per year is placed in the *medium-low* category.

Dependencies among indices in the global Bayesian network

After examining the dependency relationships between indices shown in Fig. 4, we find that the aggregated model identified some, but not all, the index dependencies.

In the following, we analyze some index dependencies encoded in our global Bayesian network. We find that the *citations* and the *h-index* are the parent nodes of the *a-index* in the Bayesian network. These nodes are represented in this index’s definition (see “The *a-index*” section). Similarly, the *m-index* definition is also represented in the network, since its parent nodes are the *citations* and the *h-index*. On the other hand, the *q²-index*, which is dependent on the *h-index* and the *m-index* (see “The *q²-index*” section), has these nodes as parents in the network. The parents of the *r-index* are the *h-index* and the *a-index*. Initially, these indices do not appear in the *r-index* definition (see “The *r-index*” section), but, after a transformation of the original definition, we obtain that the *r-index* is also obviously defined as $\sqrt{a \cdot h}$. Finally, the *e-index* is dependent on the *h-index* and the *a-index*. Like the *r-index*, the above indices are not part of the *e-index* definition (see “The *e-index*” section), but, after few transformations of this definition, we conclude that the

e-index is defined as $\sqrt{a \cdot h - h^2}$.

Besides discovering dependencies between indices, which can be checked against the index definitions, the Bayesian network is also able to discover other kinds of probabilistic dependencies. One example is *h-index* \rightarrow *rational h-index*, not directly derived from, but related to index definitions. In this case, we note that the information about *h-index* influences the probability of the *rational h-index*. Another example (see Fig. 4) is the probabilistic dependency between *r-index* and *rational h-index*. The objective of the *r-index* is to measure the citation intensity in the *h-core*, whereas the *rational h-index* measures the distance to the next value of the *h-index*. We note that these indices measure different things, but they are probabilistically dependent. More examples of such dependencies are: *q²-index* \rightarrow *w-index*, *r-index* \rightarrow *w-index* and *q²-index* \rightarrow *immediacy-index*.

Both the aggregated and the global models have been learned using Elvira software (Elvira-Consortium 2002). One of the most useful features of Elvira is the automatic coloring of arcs, which offers qualitative insight about the conditional probability tables attached to each node. This coloring is based on the sign of influence (Wellman 1990) and the magnitude of influence (Lacave 2003).

In order to understand the dependencies between the indices represented in Fig. 4, we explain some concepts about the influence and color of the arcs. For example, an arc from *X* to *Y* is said to have a positive influence if higher values of *X* lead to higher probabilities of *Y* taking higher values for any configuration of its other parents. The definition of negative influence and null influence are analogous. When the influence is neither positive nor negative nor null, then it is said to be undefined. Positive, negative, undefined, and null influence is colored in red, blue, purple, and black, respectively.

Taking into account the above concepts, we find that *documents* has a positive influence on *citations*. Likewise, *citations* influences the *h-index* and *impact factor* positively. High values of *citations* are associated with high values of *h-index* and *impact factor*. Furthermore, the *h-index* has a positive influence on the *g-index*, which also has a positive influence on the *hg-index*. On the other hand, the other arcs in Fig. 4 represent an undefined influence between the parent and child nodes. Finally, the thickness of the arc is proportional to the magnitude of the influence. Thus, we find that the relationships that have a grater influence are: *citations* \rightarrow *h-index*, *h-index* \rightarrow *g-index* and *g-index* \rightarrow *hg-index*.

Conditional independencies among indices in the global Bayesian network

Table 9 lists conditional independencies between indices, derived using the Markov properties in the global Bayesian network.

Remember that new conditional independencies can be derived using some conditional independency properties, such as, symmetry, factorization or strong joint.

Analyzing Table 9, we note that some conditional independency relationships are represented in both the aggregated Bayesian network and the global Bayesian network. Some of these conditional independencies are:

$$\begin{array}{ll} -I(\text{documents}, g\text{-index} \mid \text{citations}) & -I(g\text{-index}, \text{immediacy-index} \mid h\text{-index}) \\ -I(\text{impact factor}, h\text{-index} \mid \text{citations}) & -I(q^2\text{-index}, e\text{-index} \mid h\text{-index}, m\text{-index}) \end{array}$$

We also observe that other conditional independencies were not shown before because they had a slightly different Bayesian network structure. We note that the *a-index*, the *w-index* and the *rational h-index* are indices whose parents have undergone an important change. Taking the *a-index* independency relationships as an example, we note that *citations* and *h-index* are new parents of *a-index* in the global model, and this determines new *a-index* conditional independencies, such as, $I(a\text{-index}, \text{documents} \mid \text{citations}, h\text{-index}, g\text{-index})$.

Our global Bayesian network finds conditional independencies of which some are justified by index definitions. On the other hand, though, it discovers other conditional independencies that were not derived from such definitions.

Looking at the *e-index*, it represents the excess citations received by all papers in the *h-core*. According to this definition, it is reasonable to expect that *e-index* and *citations* would be dependent, but our global model shows that the above indices are independent given *h-index*, *g-index* and *a-index*.

Similarly, we analyze the relationships between *a-index* and *m-index*. The *a-index* is the average number of citations received by the articles included in the *h-core*, whereas the *m-index* is the median number of citations received by papers in the *h-core*. Initially, one might expect there to be a dependency relationship between *a-index* and *m-index*, but our global model suggests that the relationship is of conditional independency, given *citations* and *h-index*.

Finally, a *w-index* of at least *k* means that there are *k* distinct publications that have at least 1, 2, 3, 4, ..., *k* citations, respectively. According to its definition, the *w-index* should depend on *documents* and *citations*, but the global model shows that they are conditionally independent given *q²-index*, *r-index* and *e-index*. Other conditional independencies between indices that are not derived from index definitions are:

- $I(w\text{-index}, h\text{-index} \mid q^2\text{-index}, r\text{-index}, e\text{-index})$
- $I(\text{rational } h\text{-index}, g\text{-index} \mid h\text{-index}, q^2\text{-index}, r\text{-index})$
- $I(e\text{-index}, \text{impact factor} \mid h\text{-index}, g\text{-index}, a\text{-index})$

Exploiting the global Bayesian network model

We believe that our best model is the global Bayesian network because its structure reflects more index definitions than the aggregated model and discovers new interesting conditional independencies between indices. For this reason, we apply evidence propagation and abduction to our global model.

So far, we have used the graphical component of global Bayesian network to discover conditional (in)dependencies only. In this section, however, we also use the probabilistic component of the global Bayesian network to precisely quantify, the effect of knowing some fixed variables on the occurrence of other variables.

Table 9 Conditional independencies among indices, derived using Markov properties in the global Bayesian network

Index	Is conditionally independent of	Given
<i>documents</i>	<i>h-index, g-index, hg-index, a-index, m-index, q²-index, r-index, e-index, w-index, rational h-index, impact factor, immediacy-index</i>	<i>citations</i>
<i>citations</i>	<i>hg-index, e-index, w-index, rational h-index, immediacy-index</i>	<i>documents, h-index, a-index, m-index, impact factor</i>
<i>h-index</i>	<i>immediacy-index</i>	<i>citations, g-index, a-index, m-index, q²-index, r-index, e-index, rational h-index</i>
<i>g-index</i>	<i>w-index, rational h-index</i>	<i>h-index, hg-index, a-index, e-index</i>
<i>g-index</i>	<i>documents, citations, m-index, q²-index, impact factor, immediacy-index</i>	<i>h-index</i>
<i>hg-index</i>	<i>documents, citations, h-index, a-index, m-index, q²-index, r-index, e-index, w-index, rational h-index, impact factor, immediacy-index</i>	<i>g-index</i>
<i>a-index</i>	<i>w-index, rational h-index</i>	<i>citations, h-index, g-index, r-index, e-index</i>
<i>a-index</i>	<i>q²-index, immediacy-index</i>	<i>citations, h-index, g-index</i>
<i>m-index</i>	<i>w-index, rational h-index, immediacy-index</i>	<i>citations, h-index, q²-index</i>
<i>m-index</i>	<i>hg-index, a-index, r-index, e-index</i>	<i>citations, h-index</i>
<i>q²-index</i>	<i>documents, citations, hg-index, a-index, r-index, e-index, impact factor</i>	<i>h-index, m-index</i>
<i>r-index</i>	<i>documents, citations, g-index, hg-index, m-index, q²-index, e-index, impact factor, immediacy-index</i>	<i>h-index, a-index</i>
<i>e-index</i>	<i>documents, citations, m-index, q²-index, rational h-index, impact factor, immediacy-index</i>	<i>h-index, g-index, a-index</i>
<i>w-index</i>	<i>documents, citations, h-index, g-index, hg-index, a-index, m-index, rational h-index, impact factor, immediacy-index</i>	<i>q²-index, r-index, e-index</i>
<i>rational h-index</i>	<i>documents, citations, g-index, hg-index, a-index, m-index, e-index, w-index, impact factor, immediacy-index</i>	<i>h-index, q²-index, r-index</i>
<i>impact factor</i>	<i>h-index, g-index, hg-index, a-index, m-index, q²-index, r-index, e-index, w-index, rational h-index</i>	<i>citations</i>
<i>immediacy-index</i>	<i>documents, citations, h-index, g-index, hg-index, a-index, m-index, q²-index, r-index, e-index, w-index, rational h-index</i>	<i>impact factor</i>

In the context of Bayesian networks, evidence propagation usually refers to computing the posterior probability of each single variable given the available evidence (i.e., some fixed variables), while abduction consists of finding the most probable configuration of a set of variables of interest given the evidence.

As regards evidence propagation, we would like to know the effect on index probabilities of introducing some specific values for other indices as evidence.

Our first inference is to fix the *citations* value to *medium-low*. After setting this evidence level, we calculate the posterior probabilities of each index in Fig. 5, top. Note that the mode of all indices is *low* or *medium-low*. Taking the *impact factor* as an example, we observe the following probabilities:

- $P(\text{impact factor}=\text{low} \mid \text{citations}=\text{medium-low})=0.37$
- $P(\text{impact factor}=\text{medium-low} \mid \text{citations}=\text{medium-low})=0.30$
- $P(\text{impact factor}=\text{medium-high} \mid \text{citations}=\text{medium-low})=0.22$
- $P(\text{impact factor}=\text{high} \mid \text{citations}=\text{medium-low})=0.11$

These conditional probabilities are reasonable since fixing *citations=medium-low* as evidence, *impact factor*, which depends on *citations* (positive influence), the value of the mode should be *low* or *medium-low*. Analyzing the above conditional probabilities, we observe that *low* and *medium-low* are the most probable values, at 0.37 and 0.30, respectively.

On the other hand, the second inference is to assign a *high* value to *citations*, see Fig. 5, bottom. In this case, the value of the mode of most of the indices is *high*. Now, the different probabilities of the *impact factor* are:

- $P(\text{impact factor}=\text{low} \mid \text{citations}=\text{high})=0.02$
- $P(\text{impact factor}=\text{medium-low} \mid \text{citations}=\text{high})=0.10$
- $P(\text{impact factor}=\text{medium-high} \mid \text{citations}=\text{high})=0.31$
- $P(\text{impact factor}=\text{high} \mid \text{citations}=\text{high})=0.56$

These conditional probabilities are also reasonable.

The probabilities of the above inferences answer a question raised in the introduction, namely, What would happen to a specific journal impact factor if the papers, that it published received more citations?. The answer lies in the total distribution of the different impact factor values. Similarly, setting *citations=low*, we answer the question, What would happen to a specific journal impact factor if the papers that it published received fewer citations?.

Analyzing Fig. 5, we notice that our Bayesian network model rules out many situations with the above index values. These situations are indicated by a probability equal to zero, like, e.g.,

- $P(h\text{-index}=\text{high} \mid \text{citations}=\text{medium-low})=0.00$
- $P(g\text{-index}=\text{high} \mid \text{citations}=\text{medium-low})=0.00$
- $P(a\text{-index}=\text{low} \mid \text{citations}=\text{high})=0.00$
- $P(m\text{-index}=\text{low} \mid \text{citations}=\text{high})=0.00$.

To get the most likely plausible explanation $P(\text{configuration} \mid \text{evidence})$, we should search the configuration of values of the non-observed indices (called explanation set) that maximizes the above probability. This is possible using abductive inference (Pearl 1988).

Table 10 shows three examples of abductive inference. We set three different evidence levels at *h-index=medium-low* (like, e.g., International Journal of Pattern Recognition and

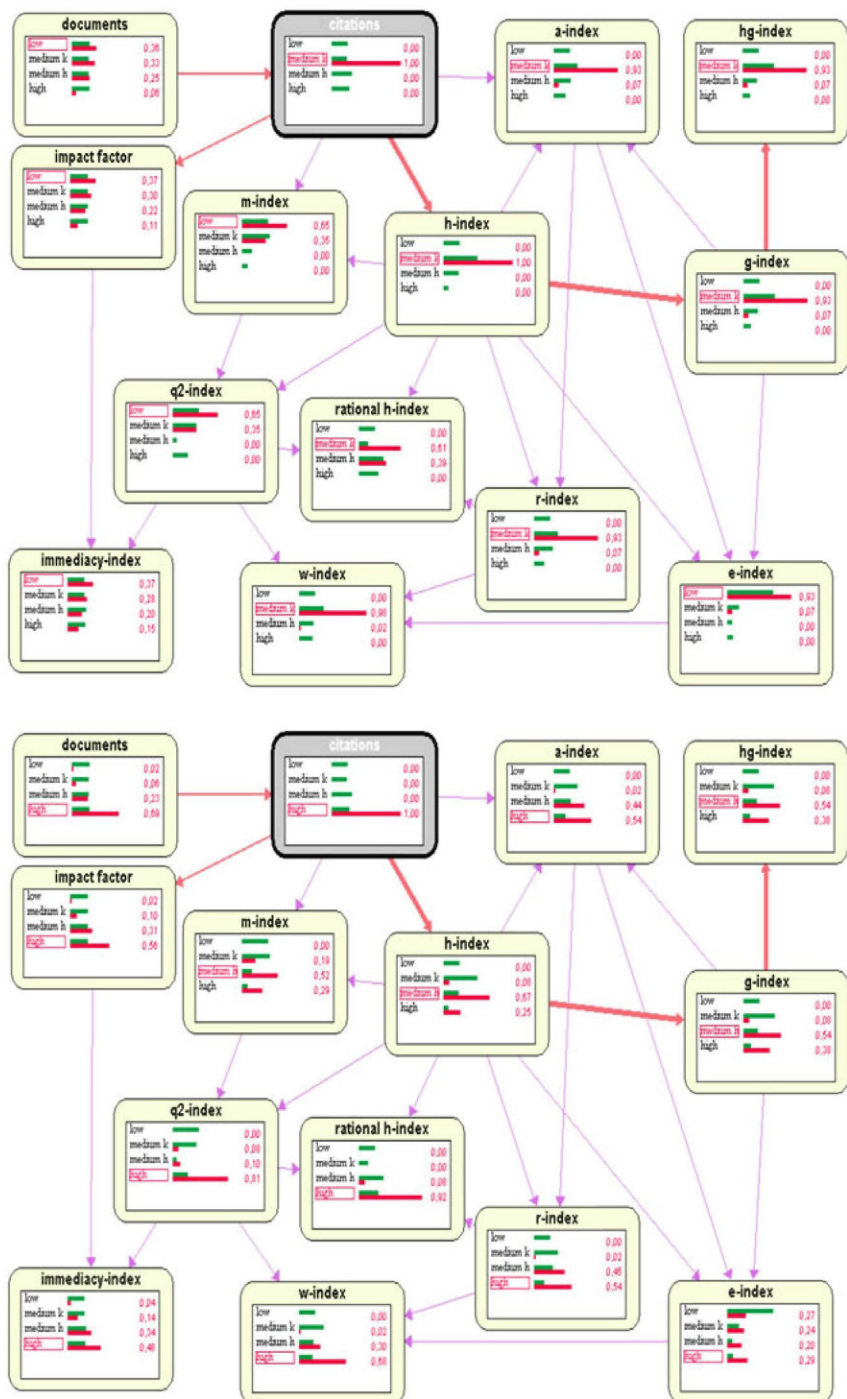


Fig. 5 Index probabilities after setting the citations value at medium-low (top) and high (bottom)

Table 10 Most likely configurations of indices for a given evidence level

Explanation set Index (X_i)	Evidence 1 <i>h-index</i> = <i>medium-low</i>	Evidence 2 <i>impact factor</i> = <i>medium-high</i>	Evidence 3 <i>immediacy-index</i> = <i>high</i>
<i>documents</i>	<i>low</i>	<i>medium-high</i>	<i>high</i>
<i>citations</i>	<i>medium-low</i>	<i>medium-high</i>	<i>high</i>
<i>h-index</i>	—	<i>medium-low</i>	<i>medium-high</i>
<i>g-index</i>	<i>medium-low</i>	<i>medium-low</i>	<i>medium-high</i>
<i>hg-index</i>	<i>medium-low</i>	<i>medium-low</i>	<i>medium-high</i>
<i>a-index</i>	<i>medium-low</i>	<i>medium-low</i>	<i>medium-high</i>
<i>m-index</i>	<i>low</i>	<i>medium-low</i>	<i>medium-high</i>
<i>q²-index</i>	<i>low</i>	<i>medium-low</i>	<i>high</i>
<i>r-index</i>	<i>medium-low</i>	<i>medium-low</i>	<i>medium-high</i>
<i>e-index</i>	<i>low</i>	<i>low</i>	<i>low</i>
<i>w-index</i>	<i>medium-low</i>	<i>medium-low</i>	<i>medium-high</i>
<i>rational h-index</i>	<i>medium-low</i>	<i>medium-high</i>	<i>high</i>
<i>impact factor</i>	<i>low</i>	—	<i>high</i>
<i>immediacy-index</i>	<i>low</i>	<i>medium-high</i>	—
<i>P</i> (Explanation set evidence)	0.000245	0.000147	0.000839

Artificial Intelligence), *impact factor*=*medium-high* (like, e.g., International Journal of Intelligent Systems) and *immediacy-index*=*high* (like, e.g., Machine Learning) in Table 10, columns 2, 3, and 4, respectively.

Taking the third inference as an example, we show that the most probable configuration of index values when *immediacy-index*=*high* is *documents*=*high*, *citations*=*high*, *h-index*=*medium-high*, *g-index*=*medium-high*, *hg-index*=*medium-high*, *a-index*=*medium-high*, *m-index*=*medium-high*, *q²-index*=*high*, *r-index*=*medium-high*, *e-index*=*low*, *w-index*=*medium-high*, *rational h-index*=*high* and *impact factor*=*high*.

The above configuration of index values answers the question, What kind of actions should journal editorial boards take to get a higher *immediacy-index* value in a specific year?. The answer is publish a lot of *documents* (≥ 79), receive a lot of *citations* (≥ 12), get high *q²-index* (≥ 2.0), *rational h-index* (≥ 2.0) and *impact factor* (≥ 1.5) index values. Moreover, journal editorial boards should aspire to the following index values: *h-index* = [2,3), *g-index* = [2,3), *hg-index* = [1.4,2.4), *a-index* = [2.0,3.0), *m-index* = [2.0,3.0), *r-index* = [1.4,2.4), *e-index* = [0.0,3.3) and *w-index* = [2,3).

Let us examine the joint probability ($P = 0.000839$) of these index values in the last row of Table 10. Although it seems very low, the number of different configurations of index values is 4^{13} . Therefore, the joint probability obtained via abduction is considerably greater than would be expected purely by chance ($\frac{1}{4^{13}} = 1.5 \cdot 10^{-8}$).

Conclusions

The advantages of Bayesian networks justify their choice as a tool for building graphical models and representing relationships among bibliometric indices. On the one hand, Bayesian networks are a graph-based model of joint multivariate probability distributions

that capture properties of conditional independency among variables. On the other hand, the statistical foundations and computational algorithms for learning Bayesian networks from observations, are well understood and have been used successfully in many applications. These models have been widely used to solve different kinds of problems (classification, regression, simulation...), and they can perform different types of reasoning (predictive, diagnostic, abductive...)

We believe that Bayesian networks are a promising tool for modeling and analyzing the dependencies and independencies among bibliometric indices. This approach places no restriction on the number of indices analyzed, and many other indices besides the 14 well-known indices covered in this study could be added.

In this study, several graphical models (yearly and global models) were developed to discover the relationships among bibliometric indices. The aim of both models is to represent relationships between index values using a within 1-year publication and citation window.

Analyzing our best Bayesian network (global model), we notice that its structure matches many index definitions. In addition, this model learns new knowledge derived from index definitions and discovers new interesting conditional (in)dependencies between analyzed indices. These conditional (in)dependency relationships have been analyzed using Markov properties.

Using our models, editorial boards of journals could find the answer to questions related to their journal citation indices. Evidence propagation and abduction inference in Bayesian networks are very useful for answering bibliometric questions.

In the future, our target will be to build new models that incorporate other journal citation indices like *eigenfactor*, *article influence* and *Scimago's journal rank index*, among others. These models could also be induced using different Bayesian network learning algorithms. The way index values are handled influences the results. They could be modeled as continuous variables instead of discretizing the values. Finally, the number of citations could vary depending on the consulted source (Google Scholar, Scopus, WoS, etc.), see Bar-Ilan (2008), which is a point to be taken into account.

Acknowledgments Research supported by Spanish Ministry of Science and Innovation, Project TIN2008-04528-E. The study has also been partially supported by Spanish Ministry of Science and Innovation, grants TIN2010-20900-C04-04, Cajal Blue Brain and Consolider Ingenio 2010-CSD2007-00018.

References

- Alonso, S., Cabrerizo, F., Herrera-Viedma, E., & Herrera, F. (2010). hg-index: A new index to characterize the scientific output of researchers based on the h- and g-indices. *Scientometrics*, 82(2), 391–400.
- Bar-Ilan, J. (2008). Which h-index? A comparison of WoS, Scopus and Google scholar. *Scientometrics*, 74(2), 257–271.
- Bergstrom, C., West, J., & Wiseman, M. (2008). The eigenfactor metrics. *Journal of Neuroscience*, 28(45), 11,433–11,434.
- Bollen, J., de Sompel, H. V., Hagberg, A., & Chute, R. (2009). A principal component analysis of 39 scientific impact measures. *Plos One*, 4(6), e6022.
- Bornmann, L., Mutz, R., & Daniel, H. (2008a). Are there better indices for evaluation purposes than the h index? A comparison of nine different variants of the h index using data from biomedicine. *Journal of the American Society for Information Science and Technology* 59(5), 830–837.
- Bornmann, L., Wallon, G., & Ledin, A. (2008b). Is the h-index related to (standard) measures and to the assessments by peers? An investigation of the h-index by using molecular life sciences data. *Research Evaluation*, 17(2), 149–156.

- Braun, T., Glanzel, W., & Schubert, A. (2006). A Hirsch-type index for journals. *Scientometrics*, 69(1): 169–173.
- Cabrerizo, F., Alonso, S., Herrera-Viedma, E., & Herrera, F. (2010). q^2 -index: Quantitative and qualitative evaluation based on the number and impact of papers in the Hirsch core. *Journal of Informetrics*, 4(1), 23–28.
- Castillo, E., Gutierrez, J., & Hadi, A. (1997). *Expert systems and probabilistic network models*. New York: Springer-Verlag.
- Cooper, G., & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4), 309–347.
- Costas, R., & Bordons, M. (2008). Is g-index better than h-index? An exploratory study at the individual level. *Scientometrics*, 77(2):267–288.
- Davis, P. (2008). Eigenfactor: Does the principle of repeated improvement result in better journal impact estimates than raw citation counts? *Journal of the American Society for Information Science and Technology*, 59(13), 2186–2188.
- Egghe, L. (2006). An improvement of the h-index: The g-index. *ISSI Newsletter* 2(1), 8–9.
- Elkins, M., Maher, C., Herbert, R., Moseley, A., & Sherrington, C. (2010). Correlation between the journal impact factor and three other journal citation indices. *Scientometrics*, 85(1), 81–93.
- Elvira-Consortium (2002). Elvira: An environment for probabilistic graphical models. In: Proceedings of the First European Workshop on Probabilistic Graphical Models (PGM'02), pp. 222–230.
- Franceschet, M. (2010). Journal influence factors. *Journal of Informetrics*, 4(3), 239–248.
- Garfield, E. (1972a). Citation analysis as a tool in journal evaluation. *Science*, 178(4060), 471–479.
- Garfield, E. (1972b). 'Citations to' divided by 'items published' gives journal impact factor. *Essays of an Information Scientist. Current Contents*, 1(7):270–273.
- Goodrum, A., McCain, K., Lawrence, S., Giles, C. (2001). Scholarly publishing in the internet age: A citation analysis of computer science literature. *Information Processing and Management*, 37(5), 661–675.
- Hirsch, J. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences* 102(46), 16,569–16,572.
- Jin, B. (2006). h-index: An evaluation indicator proposed by scientist. *Science Focus*, 1(1):8–9.
- Jin, B., Liang, L., Rousseau, R., & Egghe, L. (2007). The r- and ar-indices: Complementing the h-index. *Chinese Science Bulletin*, 52(6), 855–863.
- Lacave, C. (2003). Explanation in causal Bayesian networks. Medical applications. PhD thesis, Department of Inteligencia Artificial. UNED, Madrid, Spain (in Spanish).
- Leydesdorff, L. (2009). How are new citation-based journal indicators adding to the bibliometric toolbox? *Journal of the American Society for Information Science and Technology*, 60(7), 1327–1336.
- Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann, San Francisco.
- Ruane, F., & Tol, R. (2008). Rational (successive) h-indices: An application to economics in the Republic of Ireland. *Scientometrics*, 75(2), 395–405.
- Saad, G. (in press). Convergent validity between metrics of journal prestige: The eigenfactor, article influence, h-index scores, and impact factors. *Journal of the American Society for Information Science and Technology*.
- Schreiber, M. (2008). An empirical investigation of the g-index for 26 physicists in comparison with the h-index, the a-index, and the r-index. *Journal of the American Society for Information Science and Technology* 59(9), 1513–1522.
- SCImago (2007) SJR-SCImago Journal and Country Rank, Retrieved November 12, 2010, from <http://www.scimagojr.co>.
- Serenko, A. (2010). The development of an AI journal ranking based on the revealed preference approach. *Journal of Informetrics*, 4(4):447–459.
- Wellman, M. (1990). Fundamental concepts of qualitative probabilistic networks. *Artificial Intelligence*, 44(3), 257–303.
- Woeginger, G. (2008). An axiomatic characterization of the Hirsch-index. *Mathematical Social Science*, 56(2), 224–232.
- Zhang, C. (2009). The e-index, complementing the h-index for excess citations. *Plos One*, 4(5), e5429.