Spam: It’s Not Just for Inboxes and Search Engines!
Making Hirsch h-index Robust to Scientospam

Dimitrios Katsaros1,2  Leonidas Akritidis1  Panayiotis Bozanis1
1Dept. of Computer & Communication Engineering, University of Thessaly, Volos, Greece
2Dept. of Informatics, Aristotle University, Thessaloniki, Greece
dimitris@delab.csd.auth.gr, {leoakr,pbozanis}@inf.uth.gr

Abstract
What is the ‘level of excellence’ of a scientist and the real impact of his/her work upon the scientific
taking and practising? How can we design a fair, an unbiased metric – and most importantly – a
metric robust to manipulation?

Quantifying an individual’s scientific merit
The evaluation of the scientific work of a scientist has long attracted significant interest, due
to the benefits by obtaining an unbiased and fair criterion. A few years ago such metrics
were yet another topic of investigation for the scientometric community with only theoretical
importance, without any practical extensions.

Very recently though the situation has dramatically changed; an increasing number of
academic institutions are using such scientometric indicators to decide faculty promotions.
Automated methodologies have been developed to calculate such indicators [6]. Also, funding
agencies use them to allocate funds, and recently some governments are considering the
consistent use of such metrics for funding distribution. For instance, the Australian govern-
ment has established the Research Quality Framework (RQF) as an important feature in the
fabric of research in Australia1; the UK government has established the Research Assessment
Exercise (RAE) to produce quality profiles for each submission of research activity made by
institution2.

The use of such indicators to characterize a scientist’s merit is controversial, and a plethora
of arguments can be stated against their use. In his recent article, David Parnas [5] described
the negative consequences to the scientific progress caused by the “publish or perish” marathon
run by all scientists.

Following the reasoning of the phrase attributed to A. Einstein that “Not everything that
can be counted counts, and not everything that counts can be counted.”, we stress that the
assessment of a scientist is a complex social and scientific process that is difficult to narrow it
into a single scientometric indicator. Most of the times, the verbal descriptions of a scholar’s
quality is probably the best indicator. Though, the expressive and descriptive power of num-
bers (i.e., scientometric indicators) can not unthinkingly be ignored; instead of devaluing them,
we should strive to develop the “correct set” of indicators and, most importantly, to use them
in the right way.

No matter how skeptical is someone against the use of such indicators, the impact of a
scholar can quite safely be described in terms of the acceptance of his/her ideas by the wider
scientific community that s/he belongs to. Traditionally, this acceptance is measured by the
number of authored papers and/or the number of citations. The early metrics are based on

1http://www.uts.edu.au/research/policies/resdata/RQF.html
2http://www.rae.ac.uk
some form of (arithmetics upon) the total number of authored papers, the average number of authored papers per year, the total number of citations, the average number of citations per paper, the mean number of citations per year, the median citations per paper (per year) and so on. Due to the power-law distribution followed by these metrics, they present one or more of the following drawbacks (see also [4]):

- They do not measure the impact of papers.
- They are affected by a small number of “big hits” articles.
- They have difficulty to set administrative parameters.

J. E. Hirsch attempted to collectively overcome all these disadvantages and proposed a pioneering metric, the now famous $h$-index [4]. $h$-index was a really path-breaking idea, and inspired several research efforts to cure various deficiencies of it, e.g., its aging-ignorant behaviour [9].

Nevertheless, there is a latent weakness in all scientometric indicators developed so far, either those for ranking individuals or those for ranking publication fora, and the $h$-index is yet another victim of this complication. The inadequacy of the indicators stems from the existence of what we term here — for the first time in the literature — the scientospam.

The notion of scientospam

With a retrospective look, we see that one of the main technical motivations for the introduction of the $h$-index, was that the metrics used until then (i.e., total, average, max, min, median citation count) were very vulnerable to self-citations, which in general are conceived as a form of “manipulation”. In his original article, Hirsch made specific mention about the robustness of the $h$-index with respect to self-citations and indirectly argued that $h$-index can hardly be manipulated. Indeed $h$-index is more robust than traditional metrics, but it is not immune to them [7]. Actually, none of the existing indicators is robust to self-citations. In general, the issue of self-citations is examined in many studies, e.g., [3], and the usual practise is to ignore them when performing scientometric evaluations, since in many cases it may account for a significant part of a scientist’s reputation [1].

At this point, we argue that there is nothing wrong with self-citations; they can effectively describe the “authoritativeness” of an article, e.g., in the cases that the self-cited author is a pioneer in his/her field and s/he keeps steadily advancing his/her field in an step-by-step publishing fashion, until gradually other scientists discover and follow his/her ideas.

In the sequel we will exhibit that the problem is much more complex and goes beyond self-citations; it involves the ground meaning of a citation. Consider for instance the citing patterns appearing in Figure 1.

![Citing patterns](image)

Figure 1. Citing extremes: (Left) No overlap at all. (Right) Full overlap.

Article-1 is cited by three other papers (the ovals) and these citing articles have been authored by (strictly) discrete sets of authors, i.e., $\{a_1, a_2\}$, $\{a_3, a_4\}$ and $\{a_5, a_6\}$, respectively. On the other hand, Article-2 is cited by three other papers which all have been authored by the same author $\{a_1\}$. Notice that we make no specific mention about the identity of the authors of Article-1 or Article-2 with respect to the identity of the authors $a_i$; some of the authors of
the citing papers may coincide with those of the cited articles. Our problem treatment is more
generic than self-citations.

While we have no problem to accept that Article-1 has received three citations, we feel
that Article-2 has received no more than one citation. Reasons to have this feeling include for
instance the heavy influence of Article-2 to author $a_1$ combined with the large productivity
of this author. Nevertheless, considering that all authors $a_1$ to $a_6$ have read (have they?)
Article-1 and only one author has read Article-2, it seems that the former article has a larger
impact upon the scientific thinking. On the one hand, we could argue that the contents of
Article-2 are so sophisticated and advanced that only a few scholars, if any, could even grasp
some of the article’s ideas. On the other hand, for how long could such situation persist? If
Article-2 is a significant contribution, then it would get, after some time, its right position in
the citation network, even if the scientific subcommunity to which it belongs is substantially
smaller that the subcommunity of Article-1.

The situation is even more complicated if we consider the citation pattern appearing in
Figure 2, where there exist overlapping sets of authors in the citing papers. For instance,
author $a_3$ is a coauthor in all three citing papers.

![Figure 2. Citing articles with author overlap.](image)

This pattern of citation, where some author has coauthored multiple papers citing another
paper is the spirit of what is termed in this article the *scientometric spam* or *scientospam*.
The term spam is used in another two cases; it defines malicious emails (*e-mail spam*) and also
Web links (*link spam*) that attempt to mislead the search engines when the engines exploit
some form of link analysis ranking. Whereas the word spam has received a negative reputation
representing malicious behaviour, we use it here as a means to describe misinformation.

Apparently, there exists no prior work on combating scientospam; the closest relevant
works include techniques to filter self-citations or weigh multi-author self-citations [7, 8]. Our
target is to develop a metric of scientific excellence for individuals that will be really robust to
scientospam. We firmly believe that the exclusion of self-citations is not a fair action; neither
is any form of ad hoc normalization. Each and every citation has its value, the problem is to
quantify this value.

The notion of scientospam leads naturally to the process of the discovery of *spamming
patterns* and their “controlled discount”. If we look more carefully at the citation data, we can
gain a deeper knowledge and thus produce a fairer and more robust evaluation. A more careful
look implies that we have to pay some more computational cost than that for simple indicators,
like $h$-index, but in general we are willing to pay it, since the evaluation is an offline process.
On the other hand, we have to avoid time-consuming and doubtful clustering procedures and
special treatment of self-citations, so as to maintain the indicators’ simplicity and beauty.

The $f$-index

We consider the citing example shown in Figure 2 where an article, say $A$, is cited by three other
articles and let us define the quantity $nca^A$ to be equal to the number of articles citing article
$A$. We define the series of sets $F_i^A = \{a_j: \text{author } a_j \text{ appears in exactly } i \text{ articles citing } A\}$. 
For the case of article ART-3, we have that \(F_1^{ART} = \{a_5, a_6, a_7\}, \ F_2^{ART} = \{a_1, a_2, a_4\}, \ F_3^{ART} = \{a_3\}.

Then, we define \(f_i^A\) to be equal to the ratio of the cardinality of \(F_i^A\) to the total number of distinct authors citing article \(A\), i.e., \(f_i^A = \frac{|F_i^A|}{\text{total number of distinct authors}}\). These quantities constitute the coordinates of a \(nca\)-dimensional vector \(f^A\), which is equal to \(f^A = \{f_1^A, f_2^A, f_3^A, \ldots, f_{nca}^A\}\). The coordinates of this vector define a probability mass, since \(\sum_{i=1}^{nca} f_i^A = 1\). For the above example of the cited article ART-3, we have that \(f^{ART-3} = \{\frac{1}{3}, \frac{2}{3}, \frac{1}{3}\}\). Similarly, for the cited article ART-1, we have that \(f^{ART-1} = \{0, 0, 0\}\) and for ART-2, we have that \(f^{ART-2} = \{0, 0, 0\}\).

Thus, we have converted a scalar quantity, i.e., the number of citations that an article has received, into a vector quantity, i.e., \(f^A\), which represents the penetration of \(A\)’s ideas — and consequently of its author(s) — to the scientific community; the more people know a scholar’s work, the more significant s/he is. In general, these vectors are sparse with a lot of 0’s after the first coordinates. The sparsity of the vector reduces for the cited articles which have only a few citations. Naturally, for successful scholars we would prefer the probability mass to be concentrated to the first coordinates, which would mean that consistently new scientists become aware of and use the article’s ideas. As the probability mass gets concentrated on the coordinates near the end of \(f^A\), the “audience” gets narrower and it implies the existence of cliques, and/or citations due to minimal publishable increment, as they are both described by Parnas [5].

Though, working with vectors is complicated and a single number would be the preferred choice. At this point, we can exploit a “spreading” vector, say \(s\), to convert vector \(f\) into a single number through a dot-product operation, i.e., \(\hat{f} = f \cdot s\). For the moment will use the plainest vector defined as \(s_1 = \{nca, nca - 1, \ldots, 1\}\); other choices will be presented in the sequel. Thus, for the example article ART-3 which we are working with, we compute a new decimal number characterizing its significance, and this number is equal to \(N_f^A \cdot s_1 = \frac{2}{3} * 3 + \frac{2}{3} * 2 + \frac{1}{3} * 1 = \frac{16}{9} \Rightarrow N_f^{ART-3} \approx 2.28\).

The \(f\)-index. Now, we can define the proposed \(f\)-index in a spirit completely analogous to that of \(h\)-index. To compute the \(f\)-index of an author, we calculate the quantities \(N_f^{A_i}\) for each one of his/her authored articles \(A_i\) and rank them in a non-increasing order. The point where the rank becomes larger than the respective \(N_f^{A_i}\) in the sorted sequence, defines the value of \(f\)-index for that author.

The spreading vector. Earlier, we used the most simple spreading vector; different such vectors can disclose different facts about the importance of the cited article. Apart from \(s_1\), we propose also a couple of easy-to-conceive versions of the spreading vector. The vector \(s_2 = \{nca, 0, \ldots, 0\}\) lies at the other extreme of the spectrum with respect to \(s_1\). Finally, if we suppose that the last non-zero coordinate of \(f^A\) is \(f_{k}^A\), then we have a third version of the spreading version defined as \(s_3 = \{nca, nca - \frac{nca}{k}, nca - \frac{2nca}{k}, \ldots, 1\}\). For each one of these spreading vectors, we define the respective \(f\)-index as \(f_s^1, f_s^2,\) and \(f_s^3\). None of these three versions of the spreading vector, and consequently of the respective indexes, can be considered superior to the other two. They present merits and deficiencies in difference cases. For instance, the \(f_s^1\) index does not make any difference for large \(h\)-index values; for scientists with \(h\)-index smaller than 15, the obtained \(f_s^1\) index can be as much as 50% of the respective \(h\)-index.

Validation
As we stressed right from the beginning of the article, when it comes to characterize the entire professional life of a scholar with a single number, things get really complicated. The validation
of the usefulness of the proposed indexes is not an easy task, given our respect to the principle that “not everything that can be counted counts”. This article aims at introducing the notion of scientospam and proposing method to combat it. The comments made in this article should not harm the reputation and will not reduce the contributions of any mentioned scientist. We selected as input data to apply our ideas a number of computer scientists with high $h$-index (http://www.cs.ucla.edu/~paberg/h-number.html), who are beyond any question top-quality researchers.

Since the data provided by the aforementioned URL are not up-to-date and also they are faulty, we cleansed them first, we kept the scientists with $h$-index larger than 30. The ranking in non-increasing $h$-index is illustrated in Table 1.

| Scientist   | $f_1$ | Scientist   | $f_1$ | Scientist   | $f_1$ |
|-------------|-------|-------------|-------|-------------|-------|
| Peter Norvig | 1759  | Debbie Estrin | 982   | H. V. Jagadish | 466   |
| Kevin Kreiger | 1514 | Jose Meseguer | 400   | Richard Karp | 331   |
| John Mitchell | 1477 | B. A. Nau | 350   | Donald Knuth | 325   |
| Yue Cao | 1464 | Michael Franklin | 349 | H. V. Jagadish | 324   |
| Michael Stonebraker | 1361 | Leslie Lamport | 347 | Donald E. Knuth | 311   |
| J. Widom | 1353 | Samir Seshia | 347 | Richard Karp  | 309   |
| Jennifer Widom | 1336 | Martin Abadi | 347 | Richard Karp | 307   |
| Nancy Lynch | 1334 | Michael Franklin | 346 | H. V. Jagadish | 306   |
| Robert Tarjan | 1314 | Phillip S. Yu | 345 | Robert Tarjan | 302   |
| Hector Garcia-Molina | 1301 | Thomas S. Huang | 343 | Michael Franklin | 301   |
| Jennifer Widom | 1238 | Michael Franklin | 342 | Robert Tarjan | 298   |
| Peter Norvig | 1227 | Michael Franklin | 341 | Jennifer Widom | 297   |
| Jennifer Widom | 1225 | Michael Franklin | 340 | Robert Tarjan | 295   |
| Jennifer Widom | 1224 | Michael Franklin | 339 | Jennifer Widom | 293   |
| Jennifer Widom | 1224 | Michael Franklin | 338 | Jennifer Widom | 292   |
| Jennifer Widom | 1224 | Michael Franklin | 337 | Jennifer Widom | 290   |
| Jennifer Widom | 1224 | Michael Franklin | 336 | Jennifer Widom | 289   |

Table 1. Computer scientists’ ranking based on $h$-index.

Then, we applied the new indicators $f_{s_2}$ and $f_{s_3}$ and the results appear in Table 2. Both indicators cause changes in the ranking provided by the $h$-index. As expected, the values of the $f_{s_2}$ index are significantly different than the respective $h$-index values. It is important to note, that these differences (and their size) appear in any position, independently of the value of the $h$-index. If these differences concerned only the scientists with the largest $h$-index, then we could (safely) argue that for someone who has written a lot of papers and each paper has received a large number of citations, then some overlap citations and some self-citations are unavoidable. This is not the case though, and it seems that there is a deeper, latent explanation.

Seeking this explanation, we calculated the differences in ranking positions for each scientist when ranked with $h$-index versus when they are ranked with the $f_{s_2}$. The results are illustrated in Table 3.

The general comment is that the scientists who climb up the largest number of positions

---

3Scientists with the same $h$-index have the same ranking position. For instance, J. Widom and S. Shenker each is ranked 6-th in the $h$-index ranking. The same holds for the ranking based on $f_{s_2}$. 

---
are those whose work can “penetrate” (and thus benefit) large “audiences”. For instance, the research results by Lixia Zhang and John A. Stankovic, who work on sensors now, are cited in communities like databases, networking, communications. Other scientists whose works is used by large audiences are those working on “computer organization”, e.g., M. Frans Kaashoek, Barbara Liskov, Andrew S. Tanenbaum, etc. Notice here, that scientists’ age has nothing to do with the ranking relocation, since both younger researchers (e.g., Lixia Zhang) can climb up positions, just like elder scientists (e.g., Andrew S. Tanenbaum).

Another important question concerns whether the particular area of expertise of a researcher could help him/her acquire a larger reputation. Undoubtedly, the research area plays its role, but it is not the definitive factor. Consider for instance, the case of data mining which is a large area and has attracted an even larger number of researchers. We see that George Karypis has earned four positions in the ranking provided by $f_{s_2}$. If the area of expertise was the only rational explanation for that, then why Rakesh Agrawal, who founded the field, is among the scientists that lost the most number of positions in the ranking provided by $f_{s_2}$? The answers lies in the particularities of the research subfields; George Karypis contributed some very important results useful also in the field of bioinformatics. To strengthen this, we can mention the case of Jiawei Han. He is a data-mining expert whose work penetrates to other communities like mining, databases, information retrieval, artificial intelligence, and his is ranked second, based either on $h$-index, or on $f_{s_2}$ or on $f_{s_3}$.

Examining the scholars with the largest loses, we see that scientists who have made ground-breaking contributions and offered some unique results, e.g., Mihalis Yannakakis, and Moshe Y. Vardi, drop in the ranking provided by the $f_{s_2}$. This has nothing to do with the theoretical vs. practical sides of the computer science; contrast the cases of M. Yannakakis and M. Vardi, versus A. Zisserman and R. Agrawal. It is due to the nature of the scientific results that do not “resound” to other communities.

| Scientist | $f_{s_2}$ | Scientist | $f_{s_2}$ | Scientist | $f_{s_2}$ |
|-----------|-----------|-----------|-----------|-----------|-----------|
| Hector Garcia-Molina | 68 – 53 | Donald E. Knuth | 61 – 52 | Geoffrey B. Hinton | 57 – 37 |
| Jaewo Han | 57 – 43 | Philip S. Yu | 41 – 36 | Teuvo Kohonen | 36 – 39 |
| David E. Goldberg | 50 – 42 | Andreas G. Dimakis | 46 – 38 | Sunil Jajodia | 36 – 41 |
| Robert Tarjan | 56 – 41 | Luca Cardelli | 40 – 36 | Ankur Jain | 36 – 40 |
| Scott Shenker | 55 – 59 | Ronald Fagin | 40 – 35 | Joseph Goguen | 35 – 40 |
| Jennifer Widom | 54 – 43 | Vivien Z矛 | 40 – 34 | Michael Stonebraker | 34 – 38 |
| Jeffrey D. Ullman | 53 – 55 | Didier Dubois | 40 – 34 | Philip Wadler | 35 – 38 |
| David Culler | 52 – 53 | Alex Pentland | 40 – 33 | Amst Sheth | 35 – 39 |
| Deborah Estrin | 51 – 56 | Thomas S. Huang | 40 – 32 | Nancy Lynch | 35 – 42 |
| Rakesh Agrawal | 51 – 60 | Sally Floyd | 40 – 32 | Leonard Kleinrock | 35 – 38 |
| David A. Gondek | 50 – 52 | Helmut Mats | 39 – 32 | Peter Finkbeiner | 35 – 37 |
| Richard Karp | 49 – 55 | M. Frans Kaashoek | 40 – 41 | John A. Stankovic | 35 – 37 |
| David J. DeWitt | 48 – 51 | Carl Kesselman | 40 – 40 | Saul Greenberg | 34 – 37 |
| Huip Hanrahan | 45 – 49 | Menic V. Vardi | 39 – 38 | Steven Feiner | 34 – 37 |
| Anil K. Jain | 47 – 50 | Martin Abadi | 39 – 38 | Raghu Ramakrishnan | 34 – 40 |
| Angru Parisi | 47 – 52 | Christian Faloutsos | 39 – 37 | Krishnan Ramamohan | 34 – 38 |
| Rakesh Kasturi | 46 – 49 | Michael Yanakakis | 39 – 36 | Joe Hellerstein | 34 – 36 |
| Randy H. Katz | 46 – 51 | Michel Beller | 39 – 35 | Ramesh Govindan | 33 – 36 |
| Arnav A. Khosla | 45 – 48 | David Goldberg | 39 – 34 | Rogerio Medaglia | 33 – 34 |
| Don Towsley | 45 – 49 | Garcia Luna-Aceves | 39 – 33 | India Pearl | 32 – 36 |
| Stefan Abesop | 45 – 52 | Kai Li | 39 – 31 | Richard Lipton | 32 – 35 |
| David E. Johnson | 45 – 48 | Barbara Babcock | 39 – 30 | Ronald Fagin | 32 – 34 |
| Ken Kennedy | 44 – 49 | David Pompino | 39 – 30 | Victor Lauti | 32 – 35 |
| Rajeev Motwani | 44 – 48 | Henry Levy | 39 – 30 | Andrew S. Tanenbaum | 32 – 34 |
| Hyoung Heo | 44 – 48 | Michael Franklin | 39 – 29 | Amiel Blum | 32 – 34 |
| Ben Shneiderman | 44 – 48 | Wen Kun | 38 – 22 | Jose Moreira | 31 – 37 |
| Francesco Pugliese | 44 – 48 | Monica S. Lam | 38 – 22 | David Alb | 31 – 35 |
| W. Bruce Croft | 43 – 46 | Vipin Kumar | 38 – 21 | Willy Zwaan \textsuperscript{3} | 31 – 34 |
| Chi-H. Papadimitriou | 43 – 47 | Victor Lesser | 37 – 21 | Al. Sangiovanni-Vincentelli | 30 – 34 |
| Brendan Kehoe | 43 – 48 | James D. Tassone | 37 – 21 | Andrew S. Tanenbaum | 30 – 34 |
| Michael Stonebraker | 42 – 45 | Misha Shario | 37 – 20 | Herbert Edelsbrunner | 29 – 34 |
| Jack Dongarra | 42 – 48 | Olivier Faugeras | 37 – 20 | Tom Finin | 29 – 30 |
| Craig Chambers | 42 – 46 | Mario Szegedy | 37 – 20 | Liang Yang | 28 – 29 |
| Douglas C. Schmidt | 42 – 46 | Dimos Tzassopoulos | 37 – 20 | Maja Matari | 27 – 30 |
| Michael A. Naylor | 42 – 46 | David A. Patterson | 37 – 20 | Anthony Kline | 27 – 26 |
| Pat Hanrahan | 42 – 44 | George Karypis | 37 – 18 | John McCarthy | 26 – 29 |

Table 2. Computer scientists’ ranking based on $f_{s_2}$. The $f_{s_3}$ value is represented too.
| Scientist | h-rank | earned pos. in fig. | Scientist | h-rank | lost pos. in fig. |
|-----------|--------|-------------------|-----------|--------|------------------|
| David Haussler | 32 | 32 | +6 | Rakesh Agrawal | 62 | 5 | -2 |
| Carl Kesselman | 42 | 24 | +5 | Amir Pnueli | 56 | 9 | -2 |
| Geoffrey E. Hinton | 39 | 26 | +5 | Jesper Jøsang | 46 | 16 | -2 |
| Lixia Zhang | 49 | 16 | +4 | Shahro Yassaki | 48 | 17 | -2 |
| M. Frans Kaashoek | 43 | 23 | +4 | Didier Dubois | 49 | 16 | -2 |
| Tomas Poggio | 42 | 23 | +4 | Mihalis Yannakakis | 48 | 17 | -2 |
| Henry Levy | 42 | 23 | +4 | Oded Goldreich | 48 | 17 | -2 |
| Craig Chambers | 40 | 25 | +4 | Andrew Zisserman | 45 | 20 | -2 |
| Damek Dendukuri | 40 | 25 | +4 | Jose Meseguer | 40 | 25 | -2 |
| George Karypis | 40 | 25 | +4 | George Abello | 40 | 25 | -2 |
| Vern Paxson | 38 | 27 | +4 | Moshe Vardi | 49 | 19 | -3 |
| John A. Stankovic | 38 | 27 | +4 | Moshe Shai | 47 | 18 | -3 |
| Victor Basili | 35 | 30 | +4 | Nance Lynch | 45 | 20 | -3 |
| Andrew S. Tanenbaum | 35 | 30 | +4 | Tim Finin | 31 | 33 | +4 |
| Demetri Terzopoulos | 40 | 25 | +4 | Total | 35 | 30 | +4 |

Table 3. Largest relocations w.r.t. rank position. (Left) Most positions up. (Right) Most positions down.

Discussion

When measuring science we should always have in mind the principle which says that “not everything that can be counted counts”. On the other hand, we believe in the power of numbers and we side with Lord Kelvin which stated that “When you can measure what you are speaking about, and express it in numbers, you know something about it. But when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind: It may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science.”

We argue that instead of anathematizing each and every scientometric indicator, we should strive to develop the correct set of them. David Parnas did an excellent job in recording a number of existing and significant problems with current publication methodologies. Along the spirit of his ideas, we describe for the first time here, another dimension of publication methodologies, the existence of *scientospam* and set forth an effort to discover the spamming patterns in citation networks.

The astute reader will have realized by now that in our battle against the scientospam, we have in our arsenal the research works dealing with Web link spam [2], e.g., TrustRank, BadRank and so on. Unfortunately, the situation is radically difficult in citation networks, because they consist of entities richer than the Web pages and the Web links encountered in Web spam. Each node i.e., a citing article, in a citation network consists of entities i.e., co-authors, which form a complex overlay network above the article citation network.

We believe that the detection of spamming patterns in citation networks is quite a difficult procedure, and the cooperation of the authors is mandatory. Maybe the scientific community should set some rules about citing, rules not only ethical, but practical as well. For instance, we could have sections in the “References” section of each published article, to describe which citations involve only relevant work, which citations refer to earlier work done by the authors of the article, which citations refer to works implemented as competing works in the article, and so on. Apart from these organizational categories, others could be devised as well; whether the citing article’s results contradict or support the results of the cited articles and many other.

In any case, we believe that scientometric indicators are not a panacea, and we should work a lot before applying a set of them to characterize the achievements of a scholar. Indicators do have their significance, but some methodologies, both ethical and practical should change in order to have reliable and automated measurements of science.

References

[1] J. H. Fowler and D. W. Aksnes. Does self-citation pay? *Scientometrics*, 72(3):427–437, 2007.
[2] Z. Gyöngyi and H. Garcia-Molina. Spam: It’s not just for inboxes anymore. *IEEE Computer*, pages 28–34, October 2005.

[3] I. Hellsten, R. Lambiotte, and A. Scharnhorst. Self-citations, co-authorships and keywords: A new approach to scientists’ field mobility. *Scientometrics*, 72(3):469–486, 2007.

[4] J. E. Hirsch. An index to quantify an individual’s scientific research output. *Proceedings of the National Academy of Sciences*, 102(46):16569–16572, 2005.

[5] D. T. Parnas. Stop the numbers game: Counting papers slows down the rate of scientific progress. *Communications of the ACM*, 50(11):19–21, 2007.

[6] J. Ren and R. N. Taylor. Automatic and versatile publications ranking for research institutions and scholars. *Communications of the ACM*, 50(6):81–85, 2007.

[7] M. Schreiber. Self-citation corrections for the Hirsch index. *Europhysics Letters*, 78(3), 2007.

[8] A. Schubert, W. Glänzel, and B. Thijs. The weight of author self-citations. A fractional approach for self-citation counting. *Scientometrics*, 67(3):503–514, 2006.

[9] A. Sidiropoulos, D. Katsaros, and Y. Manolopoulos. Generalized Hirsch h-index for disclosing latent facts in citation networks. *Scientometrics*, 72(2):253–280, 2007.