Secure H-Numbers

P. Gaudry*

No Institute Given

Abstract. Some security flaws of the h-number indicator are revealed and countermeasures are proposed.

Introduction

In the past few years we have seen a constant increase in the need of indicators to assess the quality of research. There are indicators for journals (impact factors), for universities (ranking of Shanghai[1]) and for researchers (Erdos number, H-number). Though very popular, Erdos number is slightly controversial as a measure of the quality of the research of a person, since mathematicians, and in particular researchers in number theory might gain a small advantage. In order to facilitate the work of the funding agencies that need precise indicators to distribute grants, J. E. Hirsch [2] has recently proposed a new indicator that he calls the h-index. Following the Erdos number terminology, this quickly became the h-number, for Hirsch number.

In this paper, we study the resistance of h-number to malicious users: a researcher who wants to artificially increase his h-number, or a university who wants a large average h-number for its members. In the following, we shall freely assume that a Trusted Third Party is available. This is clearly realistic, since large funding agencies could take this role.

The paper is organized as follows: in Section 1 we recall the definition of the h-number and the way to compute it efficiently. In Section 2, we give evidence that there are some security problems leading to absurd values of h. In Section 3 we propose various counter-measures, based on off-the-shelf and highly reputed cryptographic tools.

1 Generalities on H-numbers

1.1 Definition

It is widely admitted that in order to assess the quality of the publications of a researchers, counting the number of published papers is not enough. Various

* The author thanks the PCRI (rest in peace) for having required him to write such a footnote, and nether providing any funding.

[1] `http://ed.sjtu.edu.cn/ranking.htm`. This ranking is often used in the news media to prove the decrepitude of European universities.

solutions have been proposed; but most of them require some non verifiable input, like the impact factor of the journals, that can be used as a multiplicative coefficient for each publication. On the other hand, the proposition of J. Hirsch is elegant by its simplicity.

Definition 1. *The h-number of an individual is the number of papers with citation number higher or equal to h.*

Let us give an example. A researcher R has published 45 papers. Among them, 15 are really bad and are nether cited (and will nether be); 4 of them are very good papers that have been cited dozens of time; the rest if average quality, with 2 papers cited 7 times, 11 papers cited 6 times and the rest being cited 5 times or less.

Then the h-number of R is 6. Indeed he has got 6 papers that have been cited at least 6 times each. To increase his h-number, R must convince somebody to cite one of his 11 papers that already have 6 citations. This would give him 7 papers with at least 7 citations and therefore a h-number of 7.

1.2 Properties

The main advantage of this indicator is that it varies consistently with time. For instance, if an author publishes 3 very nice papers when he is young, and then stops doing research, his citation number (the number of papers citing him) will continue to grow during his all life (and even after!). This will not happen with the h-number: this author will stay at 3. On the other hand, someone who publishes during his whole career with a regular quality and quantity will have a h-number that grows linearly with time (in a reasonable model), whereas with a classical citation number the growth will be quadratic.

The h-number is also meaningful when compared to the number of papers written by a researcher: if the h-number is almost equal to the number of published papers, then this author should probably publish more, even if he sometimes finds that an idea is too stupid to worth being written. On the other hand, somebody with 200 papers and a h-number of 3 should really consider preserving the trees that are used to make the paper on which his articles are printed.

In [2], the interested reader will find many other arguments demonstrating the superiority of the h-number over all previously proposed indicators.

The h-number has been introduced very recently, so that its use is not yet widely established. However, there is a growing interest in using it. For instance, several recent preprints propose improved versions of h-numbers [3, 5]. There are automatic tools to compute h-numbers, one of them based on Google Scholar [4] and another integrated to the Thomson-ISI Web of Knowledge database[2]. Finally, it has been reported that this index was an important input for the evaluation report of the INRIA French institute.

[2] http://isiknowledge.com

1.3 Efficient computation

At first sight, the definition is recursive, and one could be afraid that some simulated annealing method, conjugate gradient approximation, or some other fix-point determination algorithm is to be used. Fortunately, in his seminal work, J. E. Hirsch has given an efficient, deterministic polynomial time algorithm to compute h-numbers. We recall it here for completeness.

Algorithm 1.
Input: A list P of pairs (p_i, c_i), where p_i is a paper and c_i the number of citations of paper p_i.
Output: h-number associated to P.

1. Make a copy Q of P in order not to destroy the input;
2. Sort Q in decreasing order of the second coordinate of the entries; this sorting should be done *in place*;
3. Append a fake paper $(p_\infty, 0)$ at the end of Q;
4. Set h to 0;
5. While the $(h+1)$-th entry of Q has a second coordinate larger or equal to $h+1$, increment h;
6. Return h;

The proof of the correctness is slightly technical so we leave it to the reader. Note however the elegant solution of the fake paper to ensure the termination of the while loop.

2 Evidences of Security Flaws

In order to check the validity of the h-number, we picked random names in a volume of proceedings of some obscure cryptographic conference. Here are their h-numbers according to [4]:

Rivest, Ron	42
Shamir, Adi	40
Adleman, Leonard	24

These numbers are terribly high: according to J. E. Hirsch's research, the typical h-number for a Nobel prize is between 35 and 39. This is really strange since those names are known only to the few people working in the field. On the other hand, we can check the h-numbers of some famous cryptographers, that form the editorial board of a reputed journal:

Berson, Tom	7
Smart, Nigel	20
Phan, Raphael C.-W.	4
Dunkelman, Orr	7
Page, Dan	16

Let us also have a look at the author's h-number:

I think that these numbers speak for themselves: this is inconceivable that the almost anonymous R., S. and A. can reach a h-number which is more than the double of the h-numbers of so eminent researchers BSPDP and even higher than the typical h-number for a Nobel laureate. Some security must be added to the system in order to avoid such problems.

3 Propositions for Securing H-Numbers

3.1 Associating a unique ID to each paper

The main issue with the notion of h-number is the validity of the input of Algorithm 1 for a given researcher. The key is that the same paper can create several entries of the list: a typical case is when the title of the work has been changed due to some remarks of the referees. This is easy to overcome, and we suggest to forbid this kind of practice. However, there is still a risk due to variations of the way a paper is cited: for instance, I don't remember how to type properly Erdos within LaTeX, so I have decided to remove the strange symbols on the o. To continue the illustration, imagine that the title of the present work was "Secure Erdos number". For sure, a non-negligible proportion of the authors who will want to cite it would use the proper spelling (there are many maniacs about typography, around there!), thus creating a fake entry in the list. Some experiments have been made, demonstrating that the h-number could change up to $\pm 42.17\%$.

To prevent such problems, D. Bernstein [1] (a good researcher: $h = 33$) suggested to associate to each paper a unique chain of symbols (let us call it an ID-tag). Usually he puts ID-tag in a footnote. Our claim is that this is of no effect in protecting h-number, since despite his efforts, nobody ever reads (nor write) them. We therefore propose to put the ID-tag in the middle of the title. The title of our paper would then become something like

Secure 04e55c19cc6755cee8a711de7fa9fab1 H-Numbers.

At first sight, this is not so comfortable for the reader. But we are sure that after some time everybody will be used to it; and in fact it is good to recall to the reader and the writer that h-number and more generally indicators are very important, and improving their own score should be their first goal.

With our solution, the problem of approximate reproduction of title is overcome, since now it suffices to compare the ID-tag. Furthermore, since these ID-tags are part of the title, everybody will reproduce them when citing papers.

3.2 Man-in-the middle attack

The mechanism proposed in the previous section is not yet perfect: it is subject to the classical man-in-middle attack. We recall it for completeness. Nowadays,

most of the papers are downloaded online, instead of retrieved from an old-fashioned library. Imagine that during the transfer a malicious user intercepts the communication and changes the ID-tag (for instance to put the ID-tag of a paper by himself that nobody wants to cite). Then, the agency who computes h-numbers will be fooled.

To counter this attack, various techniques are possible. The most obvious is to use a secure channel during the transfer of the papers. However, such things are incredibly costly to put in practice, since it requires to install SSL/TLS on the appropriate computers. We propose a much better method, based on a cryptographically secure hash function. This hash function should be collision-resistant, and therefore we suggest to use MD5[3].

Then the protocol is as follows:

1. Write the title of your paper in a string S;
2. Compute the ID-tag as the MD5 hash of S;
3. Insert the ID-tag at a random place inside S.

Please note that the hash value corresponds to the *original* title, without the hash value included. Computing an ID-tag which is the hash value of the title with the ID-tag included is a difficult problem in general. Constructing cryptographic hash functions that allow this is the subject of active research by the author.

3.3 Certification

In the previous section, we have addressed the problem of a unique malicious user. However one can easily imagine several researchers making an alliance to cheat the system and increase the h-number of all of them. This is easily made by producing a lot of fake papers, with only a title and a list of references. Unfortunately it is hard to overcome, and the classical peer-review refereeing process does not give a satisfactory answer: the hypothetical above-mentioned alliance could easily create a fake journal of which they would form the editorial board.

The only entities that could really do the job of verifying that the papers are real ones are the funding agencies that in the end are the users of h-numbers. Therefore I suggest to replace the current system of peer-reviewed articles by a system where each article is sent to a central agency that decides whether or not this is a valuable scientific work. Of course, the agency would need some experts to help it in this evaluation; but these experts should *not* be researchers themselves. Otherwise an alliance is again possible.

On acceptance, the agency would then provide a digital signature of the title (which contains the ID-tag above). This digital signature could be put in a footnote of the paper, so that it could be verified by the person who wants to compute h-numbers.

[3] There are rumors that in some cases collisions can be found in MD5, but this just proves that this function has been very well studied and is then very strong.

4 Conclusion

We have proposed several counter-measures to ensure the reliability of the h-number indicator. We urge the funding agencies not to take the h-number into account in their evaluations for distributing grants, before all these important security measures are implemented.

References

1. D. Bernstein. Document IDs. http://cr.yp.to/bib/documentid.html.
2. J. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the USA*, 102:16569–16572, 2005. arXiv:physics/0508025.
3. J. Iglesias and C. Pecharroman. Scaling the h-index for different scientific ISI fields. arXiv:physics/0607224, July 2006.
4. M. Schwartzbach. H-number'o'matic. `http://www.brics.dk/~mis/hnumber.html`.
5. A. Sidiropoulos, D. Katsaros, and Y. Manolopoulos. Generalized h-index for disclosing latent facts in citation networks. arXiv:cs.DL/0607066, July 2006.