The influence of merging on h-type indices

EGGHE, Leo (2008) The influence of merging on h-type indices. In: JOURNAL OF INFORMETRICS, 2(3). p. 252-262.

DOI: 10.1016/j.joi.2008.06.002
Handle: http://hdl.handle.net/1942/8332

The influence of merging on h-type indices

by

L. Egghe

Universiteit Hasselt (UHasselt), Campus Diepenbeek, Agoralaan, B-3590 Diepenbeek, Belgium [1]
and
Universiteit Antwerpen (UA), Stadscampus, Venusstraat 35, B-2000 Antwerpen, Belgium
leo.egghe@uhasselt.be

ABSTRACT

Each information production process has a unique h-index. This paper studies the problem: what are possible h-index values if we merge two or more IPPs ?

First the paper gives examples of IPP mergings. There are at least two types: one where common sources add their number of items and one where common sources get the maximum of their number of items in the two IPPs.

In each case we show that

[1] Permanent address

Key words and phrases: merging, h-index, g-index, R-index, weighted h-index.

$$\max(h_1, h_2) \leq h \leq h_1 + h_2$$

for both types of mergings ($h_i, i = 1, 2$ the h-index in the two IPPs; h = the h-index in the merged one).

We show that the above inequalities cannot be improved (in both merging types). We also show that the same inequalities are true for the g-index and the R-index but that they are false for the weighted h-index. For the R-index we can even refine the above inequality

$$R \leq \sqrt{R_1^2 + R_2^2} < R_1 + R_2$$

I. Introduction

Merging is an important topic in informetrics. Virtually any information production process (IPP) is the result of the merging of several other ones, although it is not always clear to determine (or define) the components. The simplest example is a bibliography, spreading out over several years and were we consider this bibliography as the merged result of the components of the bibliography belonging to (published in) the same year. The bibliography can be anything consisting of sources and items (cf. Egghe (2005)): authors (or journals) and their publications (here the components consist of publications of the same publication year) or papers (of an author or a group of authors) and the citations these papers receive (here the components consist of citations to these papers in a fixed year).

An example of merging not involving time in the merging is the case of an IPP of an author consisting of articles and their received citations. Several author IPPs can be merged that way leading to a paper-citation IPP of the meta-author of the merged IPP. But also one author's IPP (papers-citations) can be considered as the merged one where the component IPPs are on different specific topics. If we merge bibliographies on different topics (regardless of the author) we have another example.

Searches in different databases (e.g. Web of Science (WoS), Scopus, Google Scholar,…) leading to papers and citations received, can be merged into one bibliography. Articles from an author or a group of authors can be considered to be merged where the components consist of papers in a fixed journal. Many other examples, even outside informetrics, can be given: mergings of texts in the sense of word-types (= sources) and tokens (= items) in linguistics, a country consisting of villages and cities (demography), employees of different companies or universities where one considers their productivity or their salaries, and so on.

An obvious but important example is the web (or part of it) which can be considered as the merging of different components. The same goes for any other social network (intranets, citation or collaboration networks,…).

So far, we have discussed the merging of the sources but we did not explain what we should do with their number of items. First of all, we can assume that the sources in the two (or more) IPPs are the same. If this is not the case we add sources with zero items until we reach that all IPPs (to be merged) have the same sources. We now have a situation as in Table 1 (for 2 IPPs)

Table 1. Two IPPs

r_1	$\#_1$	r_2	$\#_2$
1	x_1	1	y_1
2	x_2	2	y_2
3	x_3	3	y_3
4	x_4	4	y_4
.	.	.	.
.	.	.	.
.	.	.	.
N	x_N	N	y_N

The r_is $(i = 1,2)$ denote the source ranks and, as explained above, in both tables the same sources occur (but, of course, possibly on a different ranking). The $\#_i$s $(i = 1,2)$ denote the number of items in the two IPPs of the corresponding source. Based on the above examples we can define two merging types: suppose x_i and y_j are number of items of the same source.

One merging type is to give the value $x_i + y_j$ to this source in the merged IPP. Another merging type is to define $\max(x_i, y_j)$ as the number of items of this source in the merged IPP. If we merge different citing periods of articles of authors we clearly use the sum. If we merge citation scores of the same paper(s) in two different databases (e.g. WoS and Scopus) anything between the maximum and the sum can be the citation score in the merged IPP: consider the set A as the citing paper set of a paper in the first database and set B as the citing paper set of the same paper in the second database – see Fig. 1

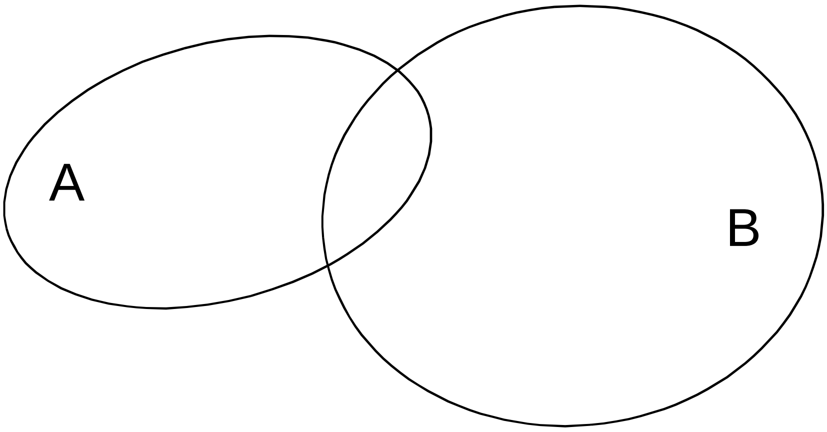

Fig. 1 Citation sets of a paper in two databases

If $A\text{Ì}\ B$ or $B\text{Ì}\ A$ it is clear that we use the maximum in the merged IPP. If $A\text{Ç}B = \text{Æ}$ it is clear we use the sum. In general, however, we will use a value between the maximum and the sum, namely $\#A + \#B - \#(A\text{Ç}B)$ (# denotes the cardinality of the set). Note that indeed

$$\max(\#A, \#B)\text{£}\ \#A + \#B - \#(A\text{Ç}B)\text{£}\ \#A + \#B$$

since $\#(A\text{Ç}B)\text{£}\ \min(\#A, \#B)$. In Vanclay (2007) two citation IPPs are merged using the max device. It concerns citation scores of the author retrieved via Google Scholar and via WoS. Here a device between max and sum would have been more appropriate but it is not so easy to determine the best values. An indication could have been obtained by sampling some papers and check the citations in each database, one by one.

In the sequel we will study the maximum and sum merging devices and from the results we will derive the results for the intermediate scoring devices.

Merging has already been studied in an old paper of Egghe and Rousseau (1988) and in another old one by Rousseau (1989). In Egghe and Rousseau (1988) one demonstrated that the merging of IPPs without a Groos droop (Groos (1967)) can lead to an IPP with a Groos droop. This time our interests are different: we want to know the influence of merging on the values of h-type indices, i.e. given, e.g. the h-index of two IPPs, what different possibilities are there for the h-index of the merged IPP. The same will be done for the g-index, R-index and the weighted h-index h_w. Let us briefly repeat their definitions (see Hirsch (2005), Egghe and Rousseau (2006, 2007), Egghe (2006) and Jin, Liang, Rousseau and Egghe (2007)). Let us have an IPP where the source on rank i has x_i items and where we suppose that the sources are ranked in decreasing order of their number of items, i.e. the x_i are decreasing.

The h-index is the unique (Egghe and Rousseau (2006)) rank h such that $i = h$ is the largest rank for which $x_i \geq i$. Since h does not take the full value of the $x_1,...,x_h$ into account, Egghe (2006) introduced the g-index: an IPP has g-index g if $i = g$ is the largest rank for which $\sum_{j=1}^{i} x_j \geq i^2$, equivalently, g is the largest rank for which the average number of items per source is larger than or equal to g:

$$\bar{x} = \frac{1}{g}\sum_{i=1}^{g} x_i \geq g \qquad (1)$$

(see Schreiber (2007)). Aiming at the same type of improvement of the h-index, Jin, Liang, Rousseau and Egghe (2007) defined the R-index as

$$R = \sqrt{\sum_{i=1}^{h} x_i} \qquad (2)$$

where h is the h-index of the IPP.

Finally, the weighted h-index also takes into account the x_i-values by weighting the ranks with it: an IPP has weighted h-index h_w if

$$h_w = \sqrt{\sum_{j=1}^{i} x_j} \tag{3}$$

for i being the largest rank such that

$$\frac{\sum_{j=1}^{i} x_j}{h} \leq x_i . \tag{4}$$

The paper is organised as follows. The next section studies the h-index in the framework of merging of two IPPs. We will show that, if h_1, h_2 are the h-indices of the two IPPs that will be merged, and if, in the merging, the sum of the item scores per source is applied, the h-index of the merged IPP, denoted h, satisfies

$$\max(h_1, h_2) \leq h \leq h_1 + h_2 \tag{5}$$

We will show that this inequality cannot be improved by giving examples where $\max(h_1, h_2)$ and $h_1 + h_2$ are actually reached.

Section III proves the same inequalities (5) for the g-index and Section IV does the same for the R-index. In Section V we show that (5) is not valid for the h_w-index.

Section VI then studies merging where the maximum of the item scores per source is applied. It is quite trivial from (5) that also in this case (5) is valid for h, g and R. Although this is a weaker result we show that (5) for h and g cannot be improved: we show that $h = \max(h_1, h_2)$ and $h = h_1 + h_2$ can be reached and the same for g. For R, however we are able to improve (5) to

$$\max(R_1, R_2) \leq R \leq \sqrt{R_1^2 + R_2^2} < R_1 + R_2 \tag{6}$$

and show that (6) cannot be improved: $R = \max(R_1, R_2)$ and $R = \sqrt{R_1^2 + R_2^2}$ can be reached.

Section VII is a concluding section, giving also some advice for further research on this topic.

II. Merging (sum device) and its influence on the h-index

In this section (and the sections III, IV and V) we apply merging using the sum device. We have the following result.

Theorem: In all cases of source matchings we have

$$\max(h_1, h_2) \leq h \leq h_1 + h_2 \qquad (7)$$

where h_1, h_2 are the h-indices of the two IPPs and h is the h-index of the merged one.

Proof: That $h \geq \max(h_1, h_2)$ is trivial since the merged IPP has, on every rank, higher (or equal – zeros are allowed) values than on the corresponding ranks in the two IPPs. To prove the right hand side (RHS) of inequality (7) we note that all sums $x_i + y_j$ with $i \geq h_1 + 1$ and $j \geq h_2 + 1$ yield values that are smaller than or equal to $h_1 + h_2$ (since each $x_i \leq h_1$ and each $y_j \leq h_2$ by definition of the h-index, applied to h_1 and h_2). So, maximally, there are $h_1 + h_2$ sums $x_i + y_j$ that can be larger than or equal to $h_1 + h_2$ (the sums involving $x_1, \ldots, x_{h_1}$ (+ an y-value) and the sums involving $y_1, \ldots, y_{h_2}$ (+ an x-value)). Hence $h \leq h_1 + h_2$. □

Example that $h = \max(h_1, h_2)$ is possible.

r_1	$\#_1$
1	2
2	2
3	1
4	1

r_2	$\#_2$
1	1
2	1
3	1
4	1

merging
®

r	#
1	3
2	3
3	2
4	2

Source matching: permutation of $\{1,2,3,4\}$ $\pi = \mathrm{Id}$ (meaning: source i in the first table is the same source as source i in the second table, $i = 1,2,3,4$). Here $h_1 = 2$, $h_2 = 1$, $h = 2 = \max(h_1, h_2)$.

Example that $h = h_1 + h_2$ is possible.

r_1	$\#_1$
1	3
2	1
3	1
4	1

r_2	$\#_2$
1	3
2	2
3	2
4	1

merging
®

r	#
1	4
2	4
3	3
4	3

Source matching: $\pi(1) = 4$, $\pi(2) = 3$, $\pi(3) = 2$, $\pi(4) = 1$. Here $h_1 = 1$, $h_2 = 2$, $h = 3 = h_1 + h_2$.

These two examples show that inequalities (7) cannot be improved. Of course, both inequalities in (7) can be strict as the next example shows.

$r_1 = r_2$	$\#_1 = \#_2$
1	4
2	3
3	2
4	1

merging
®

r	#
1	8
2	6
3	4
4	2

Both IPPs are the same and source matching is $\pi = \mathrm{Id}$. Now $h_1 = h_2 = 2$ and $h = 3$, hence $\max(h_1, h_2) < h < h_1 + h_2$.

One might think that, adding the highest values of the two IPPs yields extreme high or low values for h. This is not so: the above example shows $h = 3$ which is smaller than when we add via the device $\pi(1) = 4$, $\pi(2) = 3$, $\pi(3) = 2$, $\pi(4) = 1$: now all elements in the table have values 5, hence $h' = 4$ showing that h is not the largest possible value. The next example shows that, adding the highest values of the two IPPs does not yield the smallest possible value for h.

$r_1 = r_2$	$\#_1 = \#_2$		r	#
1	2		1	4
2	2	merging	2	4
3	2	®	3	4
4	2		4	4
5	1		5	2
6	1		6	2

Both IPPs are equal and $\pi = \text{Id}$. Here $h_1 = h_2 = 2$ and $h = 4$ (even the highest possible value here). If we use $\pi(1) = 6$, $\pi(2) = 5$, …, $\pi(6) = 1$ then we have $h' = 3 < h = 4$ which is readily seen. Of course, adding the highest values of the two IPPs can lead to the extreme values: example above: $h = h_1 + h_2$. Now follows an example that $h = \max(h_1, h_2)$.

$r_1 = r_2$	$\#_1 = \#_2$		r	#
1	3		1	6
2	2	merging	2	4
3	1	®	3	2

Here both IPPs are the same, $\pi = \text{Id}$, $h_1 = h_2 = 2$ and $h = 2$.

III. Merging (sum device) and its influence on the g-index

We again use the sum device for the merging. We have the following result.

Theorem: In all cases of source matchings we have

$$\max(g_1, g_2) £ g £ g_1 + g_2 \qquad (8)$$

where g_1, g_2 are the g-indices of the two IPPs and g is the g-index of the merged one.

Proof:

Again $g \geq \max(g_1, g_2)$ is trivial since the merged IPP has, on every rank, higher (or equal-zeros are allowed and sometimes needed in the calculation of the g-index – see Egghe (2006)) values than on the corresponding ranks in the two IPPs.

Independent of the source identification in both IPPs, we obtain the highest possible g-index for the merged IPP if we add the highest values in both IPPs, i.e. if we apply $\pi = \mathrm{Id}$. Hence it suffices to prove the RHS of inequality (8) for this merging. This will be proved if we can show that

$$\sum_{i=1}^{g_1+g_2+1} (x_i + y_i) < (g_1 + g_2 + 1)^2 \tag{9}$$

By definition of g_1 and g_2 we have

$$\sum_{i=1}^{g_1+1} x_i < (g_1 + 1)^2 \tag{10}$$

$$\sum_{i=1}^{g_2+1} y_i < (g_2 + 1)^2 \tag{11}$$

But in (10) and (11) we have, since all numbers are natural numbers, that the difference between the LHS and the RHS is at least 1. So if we add (10) and (11), the difference is at least 2. So we can subtract 1 and still have a strict inequality:

$$\sum_{i=1}^{g_1+1} x_i + \sum_{i=1}^{g_2+1} y_i < (g_1 + 1)^2 + (g_2 + 1)^2 - 1 \tag{12}$$

Further

$$\sum_{i=g_1+2}^{g_1+g_2+1} x_i \leq g_1 g_2 \qquad (13)$$

since this sum has g_2 numbers all smaller than or equal to g_1 since $i \geq g_1 + 2 > h_1$ (see Egghe (2006)) hence $x_i \leq h_1 \leq g_1$ (see Egghe (2006)).

Similarly

$$\sum_{i=g_2+2}^{g_1+g_2+1} y_i \leq g_1 g_2 \qquad (14)$$

since this sum has g_1 numbers all smaller than or equal to g_2 since $i \geq g_2 + 2 > h_2$, hence $y_i \leq h_2 \leq g_2$ (see Egghe (2006)).

Adding (12), (13) and (14) yields

$$\sum_{i=1}^{g_1+g_2+1} (x_i + y_i) = \sum_{i=1}^{g_1+1} x_i + \sum_{i=1}^{g_2+1} y_i + \sum_{i=g_1+2}^{g_1+g_2+1} x_i + \sum_{i=g_2+2}^{g_1+g_2+1} y_i$$

$$< (g_1 + 1)^2 + (g_2 + 1)^2 - 1 + 2g_1 g_2$$

$$= (g_1 + g_2 + 1)^2,$$

Proving (9), hence the theorem. □

Note:

This highest possible value of g (see the proof above) does not always yield $g_1 + g_2$ as the next example shows.

$r_1 = r_2$	$\#_1 = \#_2$
1	2
2	2
3	2
4	1

merging
®

r	#
1	4
2	4
3	4
4	2

Both IPPs are the same and $\pi = \mathrm{Id}$. Here $g_1 = g_2 = 2$ and $g = 3 < g_1 + g_2$.

Example that $g = \max(g_1, g_2)$ is possible

r_1	$\#_1$
1	3
2	3
3	3
4	1

r_2	$\#_2$
1	1
2	1
3	1
4	1

merging
®

r	#
1	4
2	4
3	4
4	2

$\pi = \mathrm{Id}$. Here $g_1 = 3$, $g_2 = 1$, $g = 3 = \max(g_1, g_2)$.

Example that $g = g_1 + g_2$ is possible.

$r_1 = r_2$	$\#_1 = \#_2$
1	2
2	2
3	2
4	2

merging
®

r	#
1	4
2	4
3	4
4	4

Both IPPs are equal, $\pi = \mathrm{Id}$. Here $g_1 = g_2 = 2$, $g = 4 = g_1 + g_2$.

Hence, the inequalities (8) cannot be improved.

Note: Unlike the case of the h-index, we now have that, adding the highest values of the two IPPs yields the highest possible value for g (but, as seen above, not always equal to $g_1 + g_2$). This is trivially seen since, at each rank of the merged IPP we have the highest possible values.

IV. Merging (sum device) and its influence on the R-index

We again use the sum device for the merging. We have the following result.

Theorem: In all cases of source matchings we have

$$\max(R_1, R_2) \le R \le R_1 + R_2 \tag{15}$$

where R_1, R_2 are the R-indices of the two IPPs and R is the R-index of the merged one.

Proof:

Since $h \ge \max(h_1, h_2)$, by the theorem in Section II and since the merged IPP has, on every rank, higher values (or equal if zeros are allowed) than on the corresponding ranks in the two IPPs, we have that

$$R \ge \max(R_1, R_2)$$

If we add the highest values in both IPPs we do not necessarily generate the highest possible R (as was true for g). This is so because h is not necessarily the highest possible value (see Section II) (see an example after this proof). But by the theorem in Section II we know that $h \le h_1 + h_2$. So if we add the highest values in both IPPs and if we use $h_1 + h_2$ for h in the formula (2) of R, we have an upper bound for R of the merged IPP. Hence

$$R^2 \le \sum_{i=1}^{h_1+h_2} (x_i + y_i)$$

$$= \sum_{i=1}^{h_1+h_2} x_i + \sum_{i=1}^{h_1+h_2} y_i$$

$$= \sum_{i=1}^{h_1} x_i + \sum_{i=h_1+1}^{h_1+h_2} x_i + \sum_{i=1}^{h_2} y_i + \sum_{i=h_2+1}^{h_1+h_2} y_i \tag{16}$$

Now

$$R_1^2 = \sum_{i=1}^{h_1} x_i \tag{17}$$

$$R_2^2 = \sum_{i=1}^{h_2} y_i \tag{18}$$

$$\sum_{i=h_1+1}^{h_1+h_2} x_i \le h_1 h_2 \tag{19}$$

since this sum contains h_2 numbers each smaller than or equal to h_1 since $i \ge h_1 + 1$ and by definition of h_1. Similarly, using the definition of h_2:

$$\sum_{i=h_2+1}^{h_1+h_2} y_i \le h_1 h_2 \tag{20}$$

Putting (17), (18), (19) and (20) in (16) we have

$$R^2 \le R_1^2 + R_2^2 + 2h_1 h_2 \tag{21}$$

But

$$h_1 \le \sqrt{\sum_{i=1}^{h_1} x_i} = R_1 \tag{22}$$

since $x_i \ge h_1$ for all $i = 1,\ldots,h_1$, by definition of h_1. Similarly, using the definition of h_2 we have

$$h_2 \le \sqrt{\sum_{i=1}^{h_2} y_i} = R_2 \tag{23}$$

Now (22) and (23) in (21) yields

$$R^2 \le R_1^2 + R_2^2 + 2R_1R_2$$

$$= (R_1 + R_2)^2$$

hence the proof of the RHS of (15). □

If no zeros are allowed then $R > \max(R_1, R_2)$. Indeed denote by π_1, π_2 the permutions of $\{1,...,N\}$ such that the merged table has values on rank i: $x_{\pi_1(i)} + y_{\pi_2(i)}$ then we have:

$$R^2 = \sum_{i=1}^{h} \left(x_{\pi_1(i)} + y_{\pi_2(i)}\right)$$

$$\ge \sum_{i=1}^{\max(h_1,h_2)} \left(x_{\pi_1(i)} + y_{\pi_2(i)}\right)$$

$$> \max\left(\sum_{i=1}^{h_1} x_i, \sum_{i=1}^{h_2} y_i\right) = \max(R_1, R_2)$$

since $h \ge \max(h_1, h_2)$ and by definition of merging using sums and since no x_i of y_i is zero: for each i: $x_{\pi_1(i)} + y_{\pi_2(i)} \ge x_i$ and $x_{\pi_1(i)} + y_{\pi_2(i)} \ge y_i$.

If zeros are allowed then R can be $\max(R_1, R_2)$ as the next example shows

r_1	$\#_1$	r_2	$\#_2$		r	#
1	3	1	1		1	3
2	2	2	0	merging	2	2
3	1	3	0	→	3	2

Here we use the source match: $\pi(1) = 3$, $\pi(2) = 2$, $\pi(3) = 1$. Note that $h_1 = 2$, $R_1 = \sqrt{5}$, $h_2 = 1 = R_2$ and $h = 2$, $R = \sqrt{5} = \max(R_1, R_2)$.

Example showing that $R = R_1 + R_2$ is possible.

$r_1 = r_2$	$\#_1 = \#_2$
1	2
2	2
3	2
4	2

merging
®

r	#
1	4
2	4
3	4
4	4

Both IPPs are equal, $\pi = \text{Id}$. Here $h_1 = h_2 = 2$, $R_1 = R_2 = \sqrt{4} = 2$, $h = 4$, $R = \sqrt{4 + 4 + 4 + 4} = 4 = R_1 + R_2$.

The example in Section III showing that we do not always get $g_1 + g_2$ as highest possible value, is also useable here: $R_1 = R_2 = 2$ and $R = \sqrt{12} < R_1 + R_2$.

In the proof of the above Theorem we mentioned that, adding the highest values in both IPPs does not always generate the highest possible R (as was the case for g). We can give an example of this:

$r_1 = r_2$	$\#_1 = \#_2$
1	4
2	3
3	2
4	1

merging
®

r	#
1	8
2	6
3	4
4	2

Here both IPPs are the same and $\pi = \text{Id}$. We have $h = 3$ and hence $R = \sqrt{18}$. But if we apply $\pi(1) = 4$, $\pi(2) = 3$, $\pi(3) = 2$, $\pi(4) = 1$ to the IPPs we get that all sources have 5 items, so $h = 4$ and hence $R = \sqrt{20} > \sqrt{18}$. This shows that the precaution in the proof of the above Theorem was in order.

V. Merging (sum device) and its influence on h_w

We show by example that merging (sum device) does not even guarantee that h_w increases.

r_1	$\#_1$
1	8
2	4
3	3

r_2	$\#_2$
1	1
2	0
3	0

merging
$\rightarrow$

r	#
1	9
2	4
3	3

Here $\pi = \mathrm{Id}$. We have $h_1 = 3$, $h_{w,1} = \sqrt{12}$, $h_2 = 1$, $h_{w,2} = 1$, $h = 3$, $h_w = \sqrt{9} = 3 < \sqrt{12} = h_{w,1}$. Hence $h_w \geq \max(h_{w,1}, h_{w,2})$ is not always valid. In Egghe and Rousseau (2007) this bad property is explained as a discrete aberration. We will not consider h_w further on.

VI. Merging (maximum device) and its influence on the h-index, the g-index and the R-index

As said in the Introduction we will now study merging where we use the maximum device for the item values: if we have two IPPs with the same sources (if not we add enough sources with 0 items) and if source i in the first IPP is the same as source j in the second IPP, then this source will have the item score $\max(x_i, y_j)$, where x_i, y_j are the item scores of this source in the first and second IPP respectively.

Theorem: In all cases of source matchings we have

$$\max(h_1, h_2) \leq h \leq h_1 + h_2 \qquad (24)$$

$$\max(g_1, g_2) \leq g \leq g_1 + g_2 \qquad (25)$$

$$\max(R_1, R_2) \leq R \leq R_1 + R_2 \qquad (26)$$

Proof:
Also in this merging device we have that the merged IPP has, on every rank, higher values than on the corresponding ranks in the two IPPs, hence $h\,^{3}\ \max(h_1, h_2)$ and $g\,^{3}\ \max(g_1, g_2)$. The same argument, together with $h\,^{3}\ \max(h_1, h_2)$ yield that also $R\,^{3}\ \max(R_1, R_2)$.

Since all values of the merged IPP using the max device are smaller than or equal to all values of the merged IPP using the sum device we trivially have that the h- and g-index of the merged IPP (max device) are smaller than or equal to the h- and g-index of the merged IPP (sum device). Hence (24) and (25) follow from this and the RHSs of (7) and (8). The same argument and using that the h-index of the merged IPP (max device) is smaller than or equal to the h-index of the merged IPP (sum device) proves the RHS of (26), now using (15). □

For R, a considerable improvement of the RHS of (26) is possible as is seen in the next Theorem.

Theorem: In all cases of source matchings we have

$$R \pounds \sqrt{R_1^2 + R_2^2} < R_1 + R_2 \tag{27}$$

Proof:
Let π_1, π_2 be the two permutations of $\{1, \ldots, N\}$ such that the source $\pi_1(i)$ in the first IPP is the same source as $\pi_2(i)$ in the second IPP and such that this source is on rank i in the merged (max device) IPP. By definition of R, the R-index of the merged IPP, we have

$$R^2 = \overset{h}{\underset{i=1}{\text{å}}} \max\left(x_{\pi_1(i)}, y_{\pi_2(i)}\right) \tag{28}$$

where h is the h-index of the merged IPP.

We consider three cases

I. Let $h_1 = h_2 = \max(h_1, h_2) = h$.

Then

$$R^2 = \sum_{i=1}^{h_1 = h_2} \max\left(x_{\pi_1(i)}, y_{\pi_2(i)}\right) \tag{29}$$

$$\leq \sum_{i=1}^{h_1} x_i + \sum_{j=1}^{h_2} y_j = R_1^2 + R_2^2$$

since $x_1, \ldots, x_{h_1}, y_1, \ldots, y_{h_2}$ are the largest values that can occur in (29) and since the maximum is smaller than the sum.

II. Let $h = \max(h_1, h_2)$.

We can suppose $h_1 > h_2$ (otherwise reverse the order of the two IPPs). If $j > h_2 + 1$ then, by definition of h_2: $y_j \leq h_2 < h_1 = h$. By definition of h, these $y_j s$ are not used in the calculation of R. Hence

$$R^2 = \sum_{i=1}^{h_1} \max\left(x_{\pi_1(i)}, y_{\pi_2(i)}\right)$$

$$\leq \sum_{i=1}^{h_1} x_i + \sum_{j=1}^{h_2} y_j = R_1^2 + R_2^2$$

since the values $y_{h_2+1}, \ldots, y_N$ are not used in the calculation of R^2.

III. Let now $h > \max(h_1, h_2)$.

If $i \geq h_1 + 1$ then, by definition of h_1: $x_i \leq h_1 \leq \max(h_1, h_2) < h$. Hence, in (28), these $x_i s$ are not used, by definition of h. Similarly, if $j \geq h_2 + 1$ then, by definition of h_2: $y_j \leq h_2 \leq \max(h_1, h_2) < h$. Hence, in (28), these $y_j s$ are not used, by definition of h. Further, in (28): $h \leq h_1 + h_2$ (by (24)) and, by the above, only (part) of the values $x_1, \ldots, x_{h_1}, y_1, \ldots, y_{h_2}$ can be used. Hence

$$R^2 \le \sum_{i=1}^{h_1} x_i + \sum_{j=1}^{h_2} y_j = R_1^2 + R_2^2$$

Hence

$$R^2 \le R_1^2 + R_2^2 \tag{30}$$

proving (27): indeed

$$\sqrt{R_1^2 + R_2^2} < R_1 + R_2$$

(strict inequality) since $R_1 R_2 \neq 0$. □

This also shows that $R = R_1 + R_2$ in (26) cannot be reached. We will now show that all other inequalities (24), (25), (26) (LHS) and (27) cannot be improved by showing that the LHSs and RHSs can be reached. For the LHSs of (24), (25) and (26) this follows from the corresponding examples given in the previous sections since merging using the max device gives h, g and R values that are smaller than or equal to their corresponding values in the merging case using the sum device.

Another set of examples showing that the LHSs of (24), (25) and (26) can actually be reached is by presenting any two identical IPPs !

It is surprising that, in this merging case using the max device, the RHSs of (24) and (25) cannot be improved. Indeed: take the example

$r_1 = r_2$	$\#_1 = \#_2$		r	$\#$
1	2		1	2
2	1	merging	2	2
3	1	→	3	1

Here both IPPs are the same and the source matching is $\pi(1) = 2$, $\pi(2) = 1$, $\pi(3) = 3$. Here $h_1 = h_2 = 1$ and $h = 2 = h_1 + h_2$! Also: $g_1 = g_2 = 1$ and $g = 2 = g_1 + g_2$.

The same example also shows that $R_1^2 = R_2^2 = 2$ and that $R^2 = 4 = R_1^2 + R_2^2$. This shows that (27) cannot be improved and hence that (27) is optimal.

General Corollary:
If, by merging, we adopt a device between the sum and the max then it follows from (7), (8), (15), (24), (25) and (26) that these inequalities are also true in this general “intermediate” case. As remarked in the introductory Section I, this case is applied when one merges citation scores of the same paper(s) in different databases.

We close this section on merging with the max device with a result that is valid when $\pi = \mathrm{Id}$, i.e. when, for each $i = 1,...,N$, the source on rank i in the first IPP is the same as the source on rank i in the second IPP.

Proposition: When $\pi = \mathrm{Id}$ and in case we apply merging with the max device, then

$$h = \max(h_1, h_2) \qquad (31)$$

, using the same notation as above.

Proof:
We can suppose h_1 ³ h_2 (otherwise change the order of the IPPs). Let us first assume $h_1 > h_2$. Then

$$\max_{i = h_2+1,...,h_1} (x_i, y_i) = x_i \qquad (32)$$

since y_i £ $h_2 < h_1$ and x_i ³ h_1 since i £ h_1. Also

$$\max_{i^3\ h_1+1} (x_i, y_i) \text{£}\ h_1 \qquad (33)$$

since $h_1 > h_2$. From (32) and (33) it follows that $h = h_1 = \max(h_1, h_2)$, since $\pi = \mathrm{Id}$.

If $h_1 = h_2$, we have that x_i £ $h_1 = h_2$ for all i ³ $h_1 + 1$ and y_i £ $h_1 = h_2$ for all i ³ $h_2 + 1 = h_1 + 1$.

Hence

$$\max_{i^3\ h_1+1}(x_i, y_i) \text{£}\ h_1 = h_2$$

Consequently, since $\pi = \mathrm{Id}$,

$$h\ \text{£}\ h_1 = \max(h_1, h_2) \tag{34}$$

By (24) and (34), we have

$$h = \max(h_1, h_2) \qquad \square$$

<u>Corollary</u>:
If we maximize the highest scores then we have the minimal h-index value.

Note that we proved in Section II that this is not true in case we <u>add</u> the highest scores.

Let us illustrate the above Proposition, thereby also yielding a counterexample to (31) for the g-index and the R-index.

r_1	$\#_1$	r_2	$\#_2$		r	#
1	5	1	7		1	7
2	4	2	4	merging	2	4
3	3	3	2	®	3	3
4	2	4	2		4	2

Here, according to the proposition, $\pi = \mathrm{Id}$. We have $h_1 = 3$, $h_2 = 2$ while $h = 3 = \max(h_1, h_2)$, according to the above proposition. But $g_1 = 3$, $g_2 = 3$ and $g = 4 > \max(g_1, g_2)$. Also $R_1 = \sqrt{12}$, $R_2 = \sqrt{11}$ and $R = \sqrt{14} > \max(R_1, R_2)$.

Final note:

The above inequalities apply to the merging (in different devices) of 2 IPPs. These inequalities can be extended to the merging of, say n, IPPs, $n \in \mathbb{N}$. Let us illustrate this on the inequalities of the type

$$\max(k_1, k_2) \leq k \leq k_1 + k_2 \tag{35}$$

(k stands for h, g or R). For the merging of n of these IPPs and denoting by $k_i\ (i = 1, \ldots, n)$ the h-type index of IPP i and denoting by k the h-type index of the merged IPP (and device described above), we have:

$$\max_{i=1,\ldots,n} k_i \leq k \leq \sum_{i=1}^{n} k_i \tag{36}$$

Proof:

From (35), the result is true for $n = 2$. Now we proceed by complete induction. Suppose (36) is true for $n \in \mathbb{N}$. Merging this IPP with an $(n+1)^{\mathrm{th}}$ IPP, yielding the h-type index k^* gives, by (35)

$$\max(k, k_{n+1}) \leq k^* \leq k + k_{n+1}$$

By (36) we have

$$\max\left(\max_{i=1,\ldots,n} k_i, k_{n+1}\right) \leq k^* \leq \sum_{i=1}^{n} k_i + k_{n+1}$$

Hence

$$\max_{i=1,\ldots,n+1} k_i \leq k^* \leq \sum_{i=1}^{n+1} k_i$$

This finishes the complete induction argument and hence the proof of this result. The same argument can be used for other inequalities, e.g. the inequality (27). □

VII. Conclusions and suggestions for further research

In this paper we studied mergings of two types (and all their intermediate values): corresponding sources receive the sum of their scores (= item values) or corresponding sources receive the maximum of their scores. For the sum device we show that

$$\max(k_1, k_2) \leq k \leq k_1 + k_2 \tag{37}$$

where k is any of the h-type indices h, g or R. We also show by example that the extreme values $\max(k_1, k_2)$ and $k_1 + k_2$ can be realized so that (37) cannot be improved. We show that (37) is false for h_w.

Merging with the max device also yields (37) for h, g and R, but we show that the RHS of (37) can be improved for R:

$$R \leq \sqrt{R_1^2 + R_2^2} < R_1 + R_2 \tag{38}$$

We show that, also in this max device case, (37) cannot be improved for h, g and that the LHS of (37) cannot be improved for R as well as (38), by giving examples that the extreme values can be reached.

Finally, when the source rankings in the first IPP are the same as the source rankings in the second IPP, we show that, in case of merging with the max device:

$$h = \max(h_1, h_2) \tag{39}$$

and we prove by example that this equality is false for g and R.

As said in the introduction, all merging devices between the maximum and the sum have applications. Of course, since (37) and (38) are inequalities, no exact values of the merged h, g or R-indices can be given here. For this, concrete IPPs must be available.

We leave some open problems.

1. Any two IPPs with the same N sources can be merged in N! ways (regarding source matching π, being a permutation of $\{1,...,N\}$). Describe the distribution of h-, g- and R-values. Preliminary calculations of this problem indicated that h-values occur in an even quantity, for which we do not have an explanation.
2. Characterise the mergings (any type), yielding for h, g and R, the minimal and maximal values, i.e. characterize the permutations π of $\{1,...,N\}$ yielding these extreme values.
3. Characterise the two IPPs (with the same number (N)sources) yielding, in any type or merging, the same h-, g- or R-values.

References

L. Egghe (2005). Power Laws in the Information Production Process: Lotkaian Informetrics. Elsevier, Oxford, UK, 2005.

L. Egghe (2006). Theory and practice of the g-index. Scientometrics 69(1), 131-152, 2006.

L. Egghe and R. Rousseau (1988). Reflections on a deflection: a note on different causes of the Groos droop. Scientometrics 14(5-6), 493-511, 1988.

L. Egghe and R. Rousseau (2006). An informetric model for the Hirsch index. Scientometrics 69(1), 121-129, 2006.

L. Egghe and R. Rousseau (2007). An h-index weighted by citation impact. Information Processing and Management 44(2), 770-780, 2008.

O.V. Groos (1967). Bradford’s law and the Keenan-Atherton data. American Documentation 18, 46, 1967.

J.E. Hirsch (2005). An index to qualify an individual’s scientific research output. Proceedings of the National Academy of Sciences of the United States of America 102, 16569-16572, 2005.

B. Jin, L. Liang, R. Rousseau and L. Egghe (2007). The R- and AR-indices: complementing the h-index. Chinese Science Bulletin 52(6), (March 2007), 855-863, 2007.

R. Rousseau (1989). Merging data sets. Scientometrics 15(3-4), 305-308, 1989.

M. Schreiber (2007). The influence of self-citation corrections on Egghe’s g index. arXiv:0707.4577v1[physics.soc-ph], 31 July 2007.

J.K. Vanclay (2007). On the robustness of the h-index. Journal of the American Society for Information Science and Technology 58(10), 1547-1550, 2007.