Measuring concentration in economic sectors by h-index and g-index

Bartolucci, Francesco

13 August 2012

Measuring concentration in economic sectors by h-index and g-index

Francesco Bartolucci*

August 13, 2012

Abstract

We show that the h-index and the g-index, which are commonly used to measure the research productivity of a scientist, may be seen as concentration indices. For these indices we also propose transformations that make them always ranging between two known limits, which correspond to the situation of null concentration and to that of high concentration. The approach is illustrated by an application to data coming from the bank sector in USA.

Keywords: Bank sector, Income distribution, Inequality.

JEL: C02, D63, G21.

1 Introduction

The h-index (Hirsch, 2005) has become one of the most commonly used indices to measure the research productivity of a scientist, even in Economics (Tol, 2009). Let N denote the number of published articles by this scientist and let y_i, $i = 1, \ldots, N$, denote the number of citations of the i-th most cited article, so that $y_1 \geq \cdots \geq y_N$. Then, this index is defined as the number h such that

$$y_i \geq h,\ i = 1, \ldots, h, \quad \text{and} \quad y_i < h,\ i = h+1, \ldots, N. \tag{1}$$

*Department of Economics, Finance, and Statistics, University of Perugia (IT), bart@stat.unipg.it

In practice, that the h-index is equal to h means that the scientist published h papers with at least h citations each.

The g-index has been proposed by Egghe (2006) as an improvement of the h-index. This index is equal to the maximum value of i such that

$$\sum_{j=1}^{i} y_j \geq i^2. \tag{2}$$

This condition may be reformulated as

$$\frac{1}{i}\sum_{j=1}^{i} y_j \geq i, \tag{3}$$

meaning that, if for a scientist this index is equal to g, then he/she published g papers having, in average, at least g citations. Like the h-index, the g-index is always between 1 (in the non-trivial case that at lest one article as been cited at least once) and $[\sqrt{C}]$, that is, the largest integer which is smaller than or equal to the square root of the total number of citations $C = \sum_{i=1}^{N} y_i$.

In this note we show that the above indices may be used to measure the concentration or inequality in an economic sector (e.g., bank sector) with reference to a certain variable of interest (e.g., amount of deposits). For this aim we take, in place of the number of citations of every published paper, the ratio between the value of the variable of interest for every considered unit and the average of this variable for all units. Moreover, once the h-index (or the g-index) has been computed on the basis of the resulting data, we measure the concentration through the sum of these ratios over the first h (g) units. For the resulting index we discuss some properties from the perspective of the concentration measurement and we illustrate some advantages. The approach is illustrated by an application based on the data about the largest US banks in terms of deposits and assets.

In the following, we first introduce the h-index and the g-index to measure concentration (Section 2) and then we illustrate the application (Section 3).

2 Measuring concentration by the g-index

Let N denote the number of units in the sector of interest and let x_i be the value of the variable of interest for the i-th ordered unit, $i = 1, \ldots, N$, so that $x_1 \geq \cdots \geq x_N$. If the amount of deposits is the variable of interest, this means x_1 is the amount for the largest bank and and x_N is that of the smallest bank. The average value of the variable of interest is obviously $\mu = \sum_i x_i/N$ and, as mentioned above, the approach here proposed consists of first computing

$$y_i = \frac{x_i}{\mu}, \quad i = 1, \ldots, N. \tag{4}$$

Then, for these data the h-index is computed as above on the basis of condition (1), whereas the g-index is computed on the basis of condition (2) or equivalently (3). Note that, since $\sum_i y_i = N$, the maximum of both indices is now $[\sqrt{N}]$.

Coming back to the example about bank deposits, a certain value h of the h-index means that there are h banks having an amount of deposits equal to at least h times the overall average. A similar interpretation may be given to the g-index in the same context. These indices have *per se* an interpretation in terms of concentration. In particular, if the h-index is equal to its maximum, $[\sqrt{N}]$, then the $[\sqrt{N}]$ largest banks have an amount of deposits at least equal to $\mu[\sqrt{N}]^2$, whereas the remaining $N - [\sqrt{N}]$ banks have an amount of deposits at most equal to $\mu(N - [\sqrt{N}]^2)$. The same happens for the g-index. This means that a percentage of $100[\sqrt{N}]/N$ banks have at least $100[\sqrt{N}]^2/N$ percent of the deposits; apart from trivial cases, the first percentage is much smaller than the second, indicating a high level of concentration. See Bartolucci (2012) for a related interpretation in bibliometric analysis.

It is important recalling that there may be some anomalies in directly using the above indices in terms of concentration. This happens, in particular, for the h-index, which may be equal to 1 both in the case of *null concentration* (when $y_i = 1$, $i = 1, \ldots, N$) and *maximum concentration* (when $y_1 = N$ and $y_i = 0$, $i = 2, \ldots, N$). These considerations lead us to preferring, as a measure of concentration, the ratio

$$A_h = \frac{1}{N}\sum_{i=1}^{h} y_i = \frac{\sum_{i=1}^{h} x_i}{\sum_{i=1}^{N} x_i}, \tag{5}$$

which is based on the h-index, instead of directly using the h-index. In our case, $100A_h$ corresponds to the percentage of the deposits belonging to the largest h banks. Similarly we define a measure based on the g-index, which is denoted by A_g. Note that condition (2) is equivalent to $A_g \geq g^2/N$ and then A_g may be simply found from a plot representing two curves: the first is the curve of the partial mean $\sum_{j=1}^{i} y_j/N$ against i, whereas the second curve is that of i^2/N against i.

In the non-trivial case that $N \geq 1$, $x_1 > 0$, and $x_i \geq 0$, $i = 2, \ldots, N$, the index A_h has the following five properties that may be simply proved. Among these properties, some concern transformations of the variable of interest (e.g., scale transformation). In these cases, by h_0 (A_{h0}) we denote the value of the h-index (A_h-index) before the transformation and by h_1 (A_{h_1}) we denote that after such a transformation. The same properties hold for the A_g-index.

1. **Minimum**: the minimum value of the A_h-index is $1/N$, which is reached if and only if there is null concentration;

2. **Maximum**: the maximum value of the A_h-index is 1, which is reached if and only if
$$\sum_{i=1}^{[\sqrt{N}]} y_i = N; \tag{6}$$

3. **Scale transformation**: if every x_i is multiplied by a constant $a > 0$, then the A_h-index does not vary, that is $A_{h1} = A_{h0}$;

4. **Translation**: if a constant $b > 0$ is added to every x_i, then the A_h-index decreases, that is $A_{h1} < A_{h0}$;

5. **Transfer**: if an amount of the variable of interest is transferred from unit j to unit i, then $A_{h1} > A_{h0}$, provided that that $i \leq h_0$ and $j > [\sqrt{N}]$.

These properties allow us to interpret the A_h-index and the A_g-index, computed on the basis of the ratios y_i defined in (4), as a measures of concentration. In particular, according to property 1, the minimum of these indices is reached in a situation of null concentration and this is the only situation in which $A_h = A_g = 1/N$. Property 2 says that the maximum of the A_h-index (and also A_g index), which is equal to 1, is reached in a situation of high concentration in which a reduced group of units (of size $[\sqrt{N}]$) dispose of all the amount of the variable of interest. Obviously, this also happens in the case of maximum concentration. However, this is not the only situation in which $A_h = A_g = 1$.

Property 3 concerns a situation in which the level of concentration does not vary since every x_i is multiplied by the same constant. On the other hand, property 4 concerns a situation in which, hypothetically, the same increase of the variable of interest concerns all the units. This reduces the level of concentration and then it is desirable that an index of concentration decreases. Finally, property 5 concerns the transfer of part of the amount of the variable of interest from a unit associated to a low value of this variable (unit j) to a unit associated to a high value (unit i). Obviously, in this case the concentration level increases and consequently the value of the A_h-index (and also of the A_g-index) increases.

Concluding this section, it is worth noting that the main advantage of the two indices above is that, to be computed, they require a reduced amount of information with respect to traditional indices of concentration. In particular, in order to compute h and g, and then A_h and A_g, we only need to know the amount of the variable of interest for the first $[\sqrt{N}]$ units of the economic sector under study, that is x_i for $i = 1, \ldots, [\sqrt{N}]$, and the average amount for all the sector. Then, the application is facilitated for sectors in which precise data may be acquired only for the largest (in terms of the variable of interest) units, whereas for the other units there may be lack of information or errors in the data. On the other hand, if necessary, the true average may be substituted by some estimated value, which may be easier to obtain than imputed values for many units.

A more obvious advantage of the proposed indices is that, having always the same maximum equal to 1, they may be used to make comparisons between situations corresponding to different values of N. On the other hand, the minimum value $1/N$ may be approximated with 0 when there is a reasonable or large number of units in the sector of interest.

Finally, one may object that any index A_k, which is defined as in (5) for an arbitrary k between 1 and N, may be used as a measure of concentration for which we may found interesting properties. For instance, we can consider $A_{[N/10]} = \sum_{i=1}^{[N/10]} x_i / \sum_{i=1}^{N} x_i$ which is the proportion of the amount of the variable of interest that belong to the 10% of the largest units, with respect to the overall amount. However, we think that A_h and A_g have a special role since they are based on finding a subgroup of units (the largest h or the largest g) by rules having a straightforward interpretation even in terms of concentration and that are nowadays well known in the scientific community.

3 Application: concentration of deposits and assets of the largest US banks

In order to illustrate the computation of the indices defined above, we consider the data about the largest 50 US banks[1], according to the *amount of deposits*, as of March 31, 2011. The data are reported in Table 1.

Among these banks, the average amounts of deposits is $\mu = 133,781.57$ (million of dollars). Then, in the table we report the amount of deposits for every bank divided by this mean, denoted by $y_i(50)$, together with the cumulate sum $\sum_{j=1}^{i} y_j(50)$ and its relative counterpart denoted by $A_i(50)$; see equation (5) for the definition of A_i. On the basis of these data (comparing the column of y_i with that of i) we find h=4, meaning that, among the 50 largest bank, there are 4 banks having an amount of deposits at least equal to 4 times the average amount. The corresponding concentration index is $A_h = A_4 = 0.565$. Moreover, from the table (comparing the column of the cumulated y_i with that of i^2) we find that $g = 5$ and $A_g = A_5 = 0.598$.

As noted in the end of the previous section, one of the advantages of the proposed indices is that, to be computed, they need the data only about the largest units. Indeed, we can compute the same indices as above, but referred to the group of the largest 300 banks, only knowing the corresponding mean. The detailed data referred the other 250 banks are not necessary. In particular, we know that the average of the deposits is $\mu = 26,602.50$ (million of dollars) for the group of the 300 largest banks. The corresponding quantities are denoted, in Table 1, by $y_i(300)$, $\sum_{j=1}^{i} y_j(300)$, and $A_i(300)$. On the basis of these results we have $h = 6$, with $A_h = A_6 = 0.525$, and $g = 13$, with $A_g = A_{13} = 0.636$.

[1] data coming from the website: `http://www.relbanks.com/top-us-banks/deposits`

As a comparison, we computed the same indices as above, for the group of the largest 50 banks[1] as of December 31, 2010, in terms of *amount of total assets*, which is the new variable of interest. We again have $h = 4$ and $g = 5$; moreover, we have $A_h = 0.571$ and $A_g = 0.597$. Then, we have very similar levels of concentration of deposits and assets among the largest 50 banks.

i	i^2	Institution Name	Deposits (×1,000,000$)	y_i (50)	y_i (300)	$\sum_{j=1}^{i} y_i$ (50)	$\sum_{j=1}^{i} y_i$ (300)	A_i (50)	A_i (300)
1	1	JPMorgan Chase Bank	1093004	8.170	41.087	8.170	41.087	0.163	0.137
2	4	Bank of America	1047013	7.826	39.358	15.996	80.444	0.320	0.268
3	9	Wells Fargo Bank	843237	6.303	31.698	22.299	112.142	0.446	0.374
4	16	Citibank	799179	5.974	30.042	28.273	142.183	0.565	0.474
5	25	U.S. Bank National Association	215206	1.609	8.090	29.882	150.273	0.598	0.501
6	36	PNC Bank	188397	1.408	7.082	31.290	157.355	0.626	0.525
7	49	The Bank of New York Mellon	158103	1.182	5.943	32.472	163.298	0.649	0.544
8	64	TD Bank	141389	1.057	5.315	33.529	168.613	0.671	0.562
9	81	HSBC Bank USA	138812	1.038	5.218	34.566	173.831	0.691	0.579
10	100	SunTrust Bank	128212	0.958	4.820	35.525	178.651	0.710	0.596
11	121	State Street Bank and Trust Company	114736	0.858	4.313	36.382	182.964	0.728	0.610
12	144	Branch Banking and Trust Company	106265	0.794	3.995	37.177	186.958	0.744	0.623
13	169	Regions Bank	99341	0.743	3.734	37.919	190.692	0.758	0.636
14	196	Capital One	98286	0.735	3.695	38.654	194.387	0.773	0.648
15	225	FIA Card Services	94234	0.704	3.542	39.358	197.929	0.787	0.660
16	256	Fifth Third Bank	84394	0.631	3.172	39.989	201.102	0.800	0.670
17	289	ING Bank, fsb	81640	0.610	3.069	40.599	204.171	0.812	0.681
18	324	RBS Citizens	74134	0.554	2.787	41.154	206.957	0.823	0.690
19	361	KeyBank National Association	63203	0.472	2.376	41.626	209.333	0.833	0.698
20	400	The Northern Trust Company	61436	0.459	2.309	42.085	211.643	0.842	0.705
21	441	Union Bank	59010	0.441	2.218	42.526	213.861	0.851	0.713
22	484	Morgan Stanley Bank	56690	0.424	2.131	42.950	215.992	0.859	0.720
23	529	Charles Schwab Bank	51285	0.383	1.928	43.333	217.920	0.867	0.726
24	576	Manufacturers and Traders Trust Company	50696	0.379	1.906	43.712	219.825	0.874	0.733
25	625	Citibank (South Dakota), N.A.	50124	0.375	1.884	44.087	221.710	0.882	0.739
26	676	Sovereign Bank	47330	0.354	1.779	44.441	223.489	0.889	0.745
27	729	Compass Bank	46567	0.348	1.750	44.789	225.239	0.896	0.751
28	784	USAA Federal Savings Bank	43167	0.323	1.623	45.112	226.862	0.902	0.756
29	841	Comerica Bank	42436	0.317	1.595	45.429	228.457	0.909	0.762
30	900	The Huntington National Bank	42033	0.314	1.580	45.743	230.037	0.915	0.767
31	961	Bank of the West	40645	0.304	1.528	46.047	231.565	0.921	0.772
32	1024	Chase Bank USA	38539	0.288	1.449	46.335	233.014	0.927	0.777
33	1089	M&I Marshall and Ilsley Bank	37187	0.278	1.398	46.613	234.411	0.932	0.781
34	1156	Harris National Association	36911	0.276	1.388	46.889	235.799	0.938	0.786
35	1225	Ally Bank	36534	0.273	1.373	47.162	237.172	0.943	0.791
36	1296	Discover Bank	35161	0.263	1.322	47.425	238.494	0.948	0.795
37	1369	Goldman Sachs Bank USA	32281	0.241	1.213	47.666	239.707	0.953	0.799
38	1444	Capital One Bank (USA)	32063	0.240	1.205	47.906	240.913	0.958	0.803
39	1521	E*TRADE Bank	31441	0.235	1.182	48.141	242.095	0.963	0.807
40	1600	UBS Bank USA	27569	0.206	1.036	48.347	243.131	0.967	0.810
41	1681	Hudson City Savings Bank	25629	0.192	0.963	48.538	244.094	0.971	0.814
42	1764	Citizens Bank of Pennsylvania	25516	0.191	0.959	48.729	245.054	0.975	0.817
43	1849	Deutsche Bank Trust Company Americas	25169	0.188	0.946	48.917	246.000	0.978	0.820
44	1936	Synovus Bank	23213	0.174	0.873	49.091	246.872	0.982	0.823
45	2025	RBC Bank (USA)	21457	0.160	0.807	49.251	247.679	0.985	0.826
46	2116	Banco Popular de Puerto Rico	20752	0.155	0.780	49.406	248.459	0.988	0.828
47	2209	New York Community Bank	20535	0.153	0.772	49.560	249.231	0.991	0.831
48	2304	American Express Bank, FSB.	20336	0.152	0.764	49.712	249.995	0.994	0.833
49	2401	First Republic Bank	20029	0.150	0.753	49.861	250.748	0.997	0.836
50	2500	City National Bank	18552	0.139	0.697	50.000	251.446	1.000	0.838

Table 1: *Distribution of the deposits among the largest 50 US banks as of March 31, 2011; (50) means that the data are based on the average of the deposits for the largest 50 banks and (300) means that they referred to the average for the largest 300 banks.*

References

Bartolucci, F. (2012). On a possible decomposition of the h-index (letter to the editor). *Journal of the American Society for Information Science and Technology*, (to appear).

Egghe, L. (2006). Theory and practise of the g-index. *Scientometrics*, 69:131–152.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Science*, 102:16569–16572.

Tol, R. S. J. (2009). The h-index and its alternatives: An application to the 100 most prolific economists. *Scientometrics*, 80:317–324.