Ordinal ranking aggregation in bibliometric analysis

Wohlrabe, Klaus and Friedrich, Elisabeth

Ifo Institute

1 February 2016

Ordinal Ranking Aggregation in Bibliometric Analysis

Elisabeth Friedrich
University of Würzburg

Klaus Wohlrabe*
Ifo Institute – Leibniz-Institute
for Economic Research
at the University of Munich e.V.
Poschingerstr. 5
81679 Munich, Germany
Phone: +49(0)89/9224-1229
wohlrabe@ifo.de

* Corresponding author.

Abstract

This paper reviews standard ranking aggregation approaches in bibliometric analysis. These include the arithmetic and the harmonic mean. We also present two less-well known aggregation schemes, lexicographic and graphicolexic, which are based on the order of the rankings. Finally, we introduce two recently proposed ranking aggregation approaches which are based on stochastic aggregation. We describe all approaches and give a small illustrative and an empirical example to highlight the differences.

JEL Code: A12, A14
Keywords: ranking aggregation, stochastic aggregation, RePEc

1 Introduction

Rankings play an important role in bibliometric analysis. Typical rankings in bibliometrics cover journals, scientists or institutions. Rankings have gained more interest, visibility and importance recently. Well ranked scientists have a higher probability of getting tenure or research funding and improving their reputation. The number of rankings has increased in recent years, which might be both due to better data availability, the increased competition amongst the science community and the need for a permanent research evaluation. Finally, people seem to be fascinated by rankings.

In order to compare the different features of various items, one can find many categories in which they could be ranked. Take, for example, the research performance of academic staff: The most obvious indicators would be the number of publications as well as the number of citations for a particular author. These, however, can be adjusted for several aspects like the quality of the journal in which the author published, how many authors were involved or who cited the work.[1] Another example are rankings of universities which are meant to give an overview of the quality of research and teaching in higher education.[2] Finally, journals are ranked according to different criteria like the impact factor, citation count or influence.[3]

With many rankings for several indicators available, finding an overall ranking becomes a rather complex issue. Not only does one face the problem of which indicators to rank, but also the problem of which method to use for the ranking aggregation, i.e., the combined ranking. Our aim is to provide a systematic overview of existing ranking aggregation approaches for ordinal rankings. To the best of our knowledge, we are the first to present such a review. Our review includes easily computable and well-known standard aggregation schemes like the arithmetic and the harmonic mean. We also present two less-well known aggregation schemes, lexicographic and graphicolexic, which are based on the ordering of the ranks. Finally, we introduce two recently proposed ranking aggregation approaches which are based on stochastic aggregation. These two have never been applied in bibliometrics before. Each method will,

[1]See rankings of scientists in the area of economists (Zimmermann (2013), Ben-David (2010) or Franses (2014)).

[2]Examples include the Times Higher Education Ranking or the Shanghai Ranking (see Marginson (2014)).

[3]See Wohlrabe (2016) for a survey of rankings of economics journals.

in general, produce a different overall ranking. We compare the procedures regarding the final output using a simple example and outline the reasons for the differences. Finally, we give an illustrative example using data taken from Research Papers in Economics (RePEc) for 50 economists. Among others, RePEc offers several rankings for economists, concerning, for example, the number of works or citations. We use data from 31 different rankings. We show how the ranking of the best 20 economists vary across different aggregation approaches.

The paper is structured as follows: Section 2 describes the different approaches available and includes the illustrative example. In section 3, we apply these approaches to economists rankings provided by RePEc. Finally, we conclude.

2 Ranking aggregation approaches

We review eight ranking aggregation approaches in total. These apply to ordinal rankings.[4] Furthermore, we abstract from missing values, i.e. we focus on complete rankings.[5] We assume N different rankings r_i. Let $r_i(a)$ be the rank given to individual a in ranking r_i. The most obvious way to aggregate the N ranks is to take their mean. Our starting point is therefore the generalized mean which is given by

$$M_p(a) = \left(\frac{1}{N} \sum_{i=1}^{N} w_i r_i(a)^p \right)^{\frac{1}{p}}. \tag{1}$$

In our analysis, the weights w_i are set to one, i.e. all rankings have the same weight. The scores $M_p(\cdot)$ for all ranked individuals are then ordered from lowest to highest and form the aggregate ranking.

2.1 Arithmetic, harmonic and geometric mean

Setting $p = 1$ in (1) we obtain the arithmetic mean of the ranks for individual a:

[4] A different approach is to aggregate the underlying scores. This is done, inter alia, in the Shanghai-Ranking or the Times Higher Education Ranking (THE). See Seiler and Wohlrabe (2012) for further details and references.

[5] For the aggregation of incomplete rankings see Cook et al (2010) and references therein.

$$M_1(a) = \frac{1}{N} \sum_{i=1}^{N} r_i(a) \ . \tag{2}$$

Note that the arithmetic mean is not robust to outliers: $M_1(a)$ can attain a very high absolute value if only one value $|r_i(a)|$ is very large. As rankings only comprise natural numbers, it is not possible to balance out a large positive value with a negative one. This means that the arithmetic mean is particularly sensitive to lower ranks (i.e., high values for $r_i(a)$) and will not favor a high rank in equal measure. Thus, it will rank those higher who show consistently good results.

Putting $p = -1$ leads to the harmonic mean. It favors good ranks and therefore, a higher aggregated rank can be achieved by being ranked high just a few times, even if the other ranks are rather low. This can be seen using (1): $p = -1$ results in

$$M_{-1}(a) = N \frac{1}{\sum_{i=1}^{N} \frac{1}{r_i(a)}} \ .$$

Thus, a high aggregated rank (equivalent to a low value for $M_{-1}(\cdot)$) is achieved if the denominator is large. This is accomplished by having small values for $r_i(a)$, which means good ranks in the individual rankings. Still, a good rank has more weight than a bad one. To see this, note that the harmonic mean of the N ranks $r_i(a)$ is equal to the inverse of the arithmetic mean of the inverse ranks, i.e.

$$M_{-1}(a) = \left(\frac{1}{N} \sum_{i=1}^{N} \frac{1}{r_i(a)} \right)^{-1} \ .$$

As the arithmetic mean punishes lower ranks (i.e. higher values for $\frac{1}{r_i(a)}$ or, equivalently, low values for $r_i(a)$), the expression in parentheses will be large for good rankings. In this case, computing the inverse yields a smaller value for $M_{-1}(\cdot)$ and thus a better aggregated ranking.

For $p \to 0$ we obtain the geometric mean:

$$M_0(a) = \sqrt[N]{\prod_{i=1}^{N} r_i(a)}.$$

It balances both arithmetic and harmonic mean: Low ranks are punished and high ones accentuated.

The advantage of these three aggregation methods clearly is their good computability. They are easily implemented and computed in a short time, even for many large rankings. Furthermore, it is possible to emphasize good or low ranks by choosing either the harmonic or the arithmetic mean, depending on the desired result.

2.2 Median rank

The median rank as described by Sculley (2007) takes a different approach to aggregate N different rankings.

The aggregated ranks are computed as follows:

1. For every ranked individual a a score $M(a) = 0$ is initialized. A counter n is set to 1.
2. Count the times that individual a is ranked n^{th} in any of the rankings. This number is then added to $M(a)$.
3. n is incremented by 1. Start again at step 2.

The first item that fulfills $M(a) > \theta$ (with θ being set to $\frac{N}{2}$ in the standard case) is given rank 1, the second rank 2 and so on. The idea is that in order to appear in a top spot of the aggregate ranking, an individual has to be ranked high in at least θ rankings. The handling of ties is arbitrary, i.e., if a tie occurs, the higher rank can be assigned to either of the individuals and the lower to the other one. We chose to make our implementation of the Median rank stable and thus to get the same results for every run of the algorithm. We had an initial list of all the individuals to be ranked which was used to iterate over all the individuals. If individuals

a and b both passed the threshold θ for a certain n and thus tied, the initial ordering of a and b was kept in the aggregate ranking.

The advantage of the Median rank is its easy implementation and the fact that it is robust to outliers. As it is a recursive algorithm, however, large data sets might cause problems such that the computation of the aggregate ranking takes a long time. In this case, there exists a trade-off between the time and effort it takes to compute the aggregate ranking with the Median rank and its robustness to outliers.

2.3 Lexicographic and graphicolexic rank

The lexicographic and the graphicolexic rank offer a fairly simple aggregation. They are oriented at the ordering of words in a dictionary, hence their names. They were introduced into bibliometrics by Zimmermann (2013).

For the lexicographic rank, all the ranks of an item are sorted, starting with the best rank. For example, if item A was ranked 2^{nd}, 5^{th} and 2^{nd} in three individual rankings, the ordering of the ranks for A would be (2, 2, 5). Then, all those items with a first rank as their best rank are ordered according to their second best rank, or, in case of ties, the third best and so on. Once all those with best rank 1 have been considered, one moves on to those with their best rank being second and so forth. Hence, an item gets a higher aggregated rank if it performs really well in many individual rankings.

The graphicolexic rank approaches the problem the other way around: Here, all the ranks for an item are sorted from the worst to the best. For item A mentioned above, this would be (5, 2, 2). The first spot in the aggregated ranking will then be given to the item whose worst rank is the best (i.e., the item having the lowest value as a first entry in the list of its ordered ranks), with the second worst being the tiebreaker and so on. Assuming a second item B that is always ranked 3^{rd}, leading to (3, 3, 3) as the list of ordered ranks, B would then be ranked higher than A in the overall ranking because B's worst rank is better than A's worst one. With this method, it pays off not to have any outliers at the bottom but to rank well consistently across the rankings.

Again, these two algorithms are easily implemented and computed. If the aggregate ranking should have a focus on really accentuating high or punishing low ranks, then the lexicographic or the graphicolexic rank, respectively, is the ideal choice.

2.4 Stochastic rank aggregation

The previously discussed methods all combine the individual rankings in the various ways presented and return an aggregate ranking only at the very end (a ranking which does not change in another run of the algorithm). Throughout the aggregation process, a final list is produced and no other list is considered as the possible output. A different procedure is given by a stochastic aggregation (SA) algorithm.

For the SA, the starting point is a randomly selected ordering of the elements to be ranked. This ordering is seen as one possible aggregate ranking. How well it represents the underlying rankings is then determined using a distance function. The original random ordering is used as the basis for further iterations. Hence, many possibilities for the aggregate ranking are considered and, in an iterative process, the optimal list is returned as the final ranking. This algorithm is non-deterministic, i.e., the resulting aggregate ranking will generally not be the same for two runs of the algorithm. In addition to this, the SA can be seen as trying to find a 'compromise ranking' because the final ranking is meant to represent all the underlying rankings as well as possible. The other methods, however, all accentuate some feature (like the harmonic mean favoring good ranks). The SA was introduced by Pihur et al (2009) in the area of bioinformatics.

An essential part of the SA is the already mentioned distance function $d(r_i, r_j)$ which measures the distance between two rankings r_i and r_j. Two main approaches exist to compute such a function. Firstly, one can take into account the actual ranks and sum up the absolute differences between the ranks $r_i(a)$ and $r_j(a)$ of item a in ranking r_i and r_j:

$$d_S(r_i, r_j) := \sum_{a \in r_i \cup r_j} |r_i(a) - r_j(a)| \tag{3}$$

$d_S(\cdot, \cdot)$ is called the Spearman footrule distance.

A different approach is given by Kendall's tau distance. Instead of considering the actual ranks, Kendall's tau uses the relative ordering of the elements. To illustrate this, take elements a and b that appear in the two rankings r_i and r_j. If a is ranked higher than b in both rankings (or the other way around), the distance between them will be 0 as the two rankings agree on the relative ordering of a and b. If a and b are ranked differently in the two rankings (i.e., $r_i(a) > r_i(b)$ and $r_j(a) < r_j(b)$ or vice versa), then this disagreement will account for the overall distance to rise by 1. The total distance between rankings r_i and r_j is then defined by

$$d_K(r_i, r_j) := \sum_{a,b \in r_i \cup r_j} K(a, b) \tag{4}$$

with

$$K(a,b) := \begin{cases} 0 & \text{if } r_i(a) < r_i(b),\ r_j(a) < r_j(b) \text{ or } r_i(a) > r_i(b),\ r_j(a) > r_j(b) \\ 1 & \text{if } r_i(a) < r_i(b),\ r_j(a) > r_j(b) \text{ or } r_i(a) > r_i(b),\ r_j(a) < r_j(b) \end{cases}$$

Both distance functions imply an objective function $\Phi(R, r_1, ..., r_N)$ which measures the overall distance between an aggregated ranking R and the N individual rankings r_i:

$$\Phi(R, r_1, ..., r_N) := \sum_{i=1}^{N} d_m(R, r_i), \quad m = S, K \tag{5}$$

Finding a minimum of this objective function is equivalent to finding a ranking that represents the individual rankings best, given the distance function. Pihur et al (2009) propose two estimation routines to perform the SA: the Genetic algorithm (GA) and the Cross-Entropy Monte Carlo method. As the latter one is computationally very intensive, we focus on the former one. The basic procedure for the GA is the following: An initial ranking R is randomly selected and contains the elements from the r_i, but in an arbitrary order. Then, $\Phi(R, r_1, ..., r_N)$ is computed and R is updated to $\tilde{R}$. In every iteration, the objective function $\Phi(\cdot)$ is evaluated. If its value does not change for 30 iterations (default settings), the current ranking is returned as the optimal ranking.

2.5 An illustrative example

To illustrate the differences between the aggregation algorithms, consider the following example. Table 1 shows seven different rankings for some artificial scientists A, B, C, D and E. Each column constitutes a single ranking. Every one of them could represent a different feature to be ranked, or the opinion of one person on what the ordering should look like.

Table 1: Illustrative Ranking Example

Rank	r_1	r_2	r_3	r_4	r_5	r_6	r_7
1.	A	A	C	B	E	E	A
2.	B	C	A	A	C	A	B
3.	C	B	B	E	A	D	E
4.	D	D	E	D	B	C	C
5.	E	E	D	C	D	B	D

Notes: This table reports seven different rankings (r_i) for five artificial scientists (A to E).

The rankings are very heterogeneous. While item A is in the top 3 for every list, item E occupies the last rank (Ranking 1, 2) as well as the first one (Ranking 5, 6).

Table 2: Results of standard aggregation schemes

Rank	Arithm. mean	Harm. mean	Geom. mean	Median	Lexicographic	Graphicolexic
1.	A	A	A	A	A	A
2.	B	E	B	B	E	B
3.	C	B	E	C	B	C
4.	E	C	C	E	C	E
5.	D	D	D	D	D	D

In Table 2 we state the results of the standard aggregation schemes. The arithmetic mean punishes the two last ranks for scientist E and E reaches the 4^{th} rank, while E reaches the second place when applying the harmonic mean which rewards the two first ranks in Ranking 5 and 6. The median rank offers a fairly stable aggregation: It pays off to have high ranks, but a lower one is not immediately punished. Thus, scientist B, which is ranked mostly in the top 3 but has two outliers at the bottom (4^{th} and 5^{th} in Ranking 5 and 6, respectively), reaches the second rank. In this particular example, median and arithmetic rank return the same aggregate ranking. A reason for this is, apart from the sample size, the handling of ties described in 2.2. In this particular example, the lexicographic and the graphicolexic rank lead to the same results as the harmonic and the arithmetic mean, respectively. This might be due to the small size of the example, but the basic idea behind the algorithms is similar: Both lexicographic rank and harmonic mean favor good ranks, while graphicolexic rank and arithmetic mean punish lower ranks.

Table 3: Results for the stochastic rank aggregation

Rank	Spearman's footrule distance			Kendall's tau		
1.	A	A	A	A	A	A
2.	B	B	B	B	B	B
3.	E	E	E	E	E	C
4.	C	C	C	C	C	E
5.	D	D	D	D	D	D
$\Phi(\cdot)$	5.143			5.143		

Notes: Our algorithm of choice was the Genetic algorithm with both Spearman's footrule distance and Kendall's tau. $\Phi(\cdot)$ is the value of the objective function, evaluated for the rankings in the columns.

Table 3 shows the results for the stochastic ranking aggregation using the Genetic algorithm (GA) with Spearman's footrule distance and Kendall's tau. As this approach is stochastic, the results may vary with every run of the algorithm.[6] To get a first impression, we ranked the given lists three times for each chosen distance function. Each column represents a resulting aggregate ranking. Here, the three runs for the GA with Spearman's footrule distance all returned the same ranking, while for Kendall, the three rankings are the same but for one swap of E and C in the last one. The last row gives the value of the objective function $\Phi(R, r_1, \ldots, r_7)$ which was the same for the three iterations of each algorithm. As $\Phi(R, r_1, \ldots, r_7)$ returns the overall distance between the aggregated ranking R and the individual r_i ($i = 1, \ldots, 7$), this means that all three runs of each algorithm returned rankings that were equally optimal, regarding the chosen distance function.

[6] A seed can be set to ensure the reproducibility of the results, but we chose to use the default setting of no seed.

3 An application to economists

3.1 The database

In economics, RePEc (Research Papers in Economics, www.repec.org) has become an essential source for the spread of knowledge and ranking of individual authors and academic institutions. RePEc is based on the 'active participation principle', i.e., authors, institutions and publishers have to register and to provide information to the network. This approach has the main advantage that a clear assignment of works and citations to authors and articles is possible.[7] Indeed, the RePEc story has become a success, with more than 45,000 registered authors with listed works and 13,000 institutions in economic sciences worldwide as of April 2015. RePEc offers a large variety of individual rankings for economists and institutions. Based on all available bibliographic information within the network, RePEc calculates 37 different bibliometric indicators for registered authors and institutions every month. Table 4 provides an overview of these measures. There are six main categories: number of (published) works, citations, citation indices, citing authors, journal pages, and RePEc access statistics. Each of these main categories can be combined with different weighting schemes: simple or recursive impact factors, number of authors and combination of them. For the category 'distinct number of works' different version of a paper are counted only once. Published work is only counted if first, the publisher provides the meta data to RePEc and second, the author assigns this work to his/her account. Currently there are more than 2,100 journals and about 4,100 working papers listed in RePEc and the list is constantly expanding. To the best of our knowledge no major journal or working paper series is missing in RePEc. The indicators are not publicly available on the web page, RePEc only reports the bibliometric scores for the top 5% listed authors for each category. Therefore, only for authors belonging to the top 5% list in each category a complete record can be established. RePEc provides all scores with its corresponding worldwide rank for each author every month via email. Table 4 reveals that there is a focus on citations both directly and indirectly. In 14 out of 37 rankings citations are

[7] For instance, Google Scholar as a source for citation analysis potentially suffers from the problem of clear identification of citations which can lead to overestimation of citations, see Harzing and van der Wal (2009).

counted with quality and time adjustments. The indirect channels are the different impact factors. For further details on RePEc, see Zimmermann (2013) or Seiler and Wohlrabe (2012). Given the 37 individual RePEc rankings, we used 31 (excluding number of works, the Wu-Index, Closeness, Betweenness, Strength of Students and NEP-Cites). To ensure complete rankings, we only used the authors who were among the top 5% in each of the 31 categories. We downloaded the data in March 2015 and it refers to the February ranking in RePEc.

Table 4: Bibliometric measures in RePEc

		Without any further weighting	Simple Impact Factor	Recursive Impact Factor	Number of Authors	Number of Authors + Simple Impact Factor	Number of Authors + Recursive Impact Factor
Works	Overall	X					
	Distinct	X	X	X	X	X	X
Citations	Overall	X	X	X	X	X	X
	Discounted by citation year	X	X	X	X	X	X
Citing Authors	Overall	X					
	Weighted by authors rank	X					
Journal Pages		X	X	X	X	X	X
Access via RePEc	Abstract Views	X			X		
	Downloads	X			X		
Indices	h-Index	X					
	Wu-Index	X					
	Closeness	X					
	Betweenness	X					
	Strength of Students	X					
	NEP-Cites	X					

Notes: See Zimmermann (2013) or Seiler and Wohlrabe (2012) for further details.

3.2 Estimation issues for the stochastic rank aggregation

For the rank aggregation using the Genetic algorithm, every ranking had to be surrendered as a list. For example, if in category A the resulting ranking places a before b before c, GA uses (a, b, c) as an input. Ties in the scores cause a problem in this approach: It is not clear whether a is actually ranked higher than b or if, in fact, they reached the same score. To ensure comparability with the other algorithms, we adjusted the data for the other methods accordingly. That means that if in a certain category, two authors had the same score, they were given two different ranks, assigning the lower one if the author was the first one in the RePEc ranking with that score.[8]

As the stochastic algorithm is computationally intensive for large data sets, we first ranked the authors with complete rankings, using the arithmetic mean, and then selected the top 50 for further analysis.[9] This means that first, we aggregated the complete rankings, using the arithmetic mean. Then, we took the first 50 of these, computed the new individual ranks[10] and aggregated the shortened rankings. Our list contains several Nobel Prize winners such as James Heckman, Jean Tirole, Paul Krugman and Gary Becker.

A further difficulty is that the algorithm generally did not converge within the given iteration limit of 1000 iterations (i.e. the value of the objective function did not stay the same for the default setting of 30 iterations in a row, but varied). The resulting ranking will, in general, not be the optimal list. To compensate for this, we started 1000 runs of both variations of the GA (with Spearman's footrule and Kendall's tau, respectively) and averaged the rankings, using the arithmetic mean.

3.3 Results

Table 5 shows the results for the different ranking aggregation approaches. Our presentation is based on the arithmetic mean, i.e. we documented the ranks for the scientists who were in the top 20 according to the arithmetic mean.

Taking a look at the top 3 (James Heckman, Joseph Stiglitz and Andrei Shleifer), their ranks seem rather consistent when applying the different methods. Exceptions are the lexicographic rank for James Heckman (due to him being ranked first only a few times, while Andrei Shleifer occupies many first ranks) and the graphicolexic rank for Andrei Shleifer, whose worst rank in the original rankings is 24.[11]

[8] This is due to the implementation for ordering the scores for the GA.

[9] We chose the arithmetic mean as a starting point because it is the most intuitive form of ranking aggregation.

[10] For example, James J. Heckman is in the top 50. In the first ranking, he occupied rank 14. Using only the top 50, this rank improved to 6 because several authors who were ranked better in this category didn't make the top 50.

[11] Still, Andrei Shleifer excels when applying the harmonic mean, due to his large number of top spots. This

Table 5: Top 20 based on the arithmetic mean

Name	Arithm.	Harm.	Geom.	Med.	Lex.	Graphic.	GA S	GA K
James J. Heckman	1	2	1	1	5	1	1	1
Joseph E. Stiglitz	2	4	3	4	4	2	3	4
Andrei Shleifer	3	1	2	2	1	6	2	2
Robert J. Barro	4	3	4	3	3	5	4	3
Daron Acemoglu	5	6	5	5	7	3	5	5
Jean Tirole	6	9	6	6	10	4	6	6
Olivier J. Blanchard	7	14	11	9	24	7	11	12
Kenneth S. Rogoff	8	10	8	7	13	8	7	7
John Y. Campbell	9	11	9	8	17	12	8	8
Peter C. B. Phillips	10	5	7	15	2	17	10	9
Thomas J. Sargent	11	15	14	16	15	19	13	13
Paul R. Krugman	12	16	15	17	12	21	18	18
Robert E. Lucas Jr.	13	7	10	11	8	50	9	10
David E. Card	14	26	24	20	29	9	21	21
Gary S. Becker	15	12	12	10	11	41	12	11
Lawrence H. Summers	16	25	21	14	36	16	16	16
Ben S. Bernanke	17	18	16	12	18	22	14	14
Maurice Obstfeld	18	29	26	23	33	10	23	23
Ross Levine	19	17	17	18	20	34	17	17
Alan B. Krueger	20	33	30	27	40	11	25	26

Notes: *Arithm.*: Arithmetic mean, *Harm.*: Harmonic mean, *Geom.*: Geometric mean, *Med.*: Median rank, *Lex.*: Lexicographic, *Graphic.*: Graphicolexic, *GA S*: Genetic algorithm with Spearman's footrule distance, *GA K*: Genetic algorithm with Kendall's tau.

The widest ranges in the individual ranks are given by the lexicographic and the graphicolexic rank. Take, for example, Peter C. B. Phillips: Aggregating the rankings with the lexicographic rank, he reaches the second place, while he only occupies rank 17 when applying the graphicolexic rank. The reason for this is his rather high number of first ranks (six in total), but also his wide range of ranks (with the worst rank being 45). This illustrates the problem in these two approaches: There might be too much weight on good ranks for the lexicographic and on low ranks for the graphicolexic rank. Consider the following (extreme) example: Take scientist A, who is ranked first twice and 20^{th} in eight other rankings, and scientist B, who is ranked first once and second elsewhere. The lexicographic rank will put A before B, even though overall B has the better results and should therefore be assigned the higher aggregate rank.

is done in the official RePEc ranking.

Table 6: Rank distribution for the stochastic algorithm using Spearman

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
James J. Heckman	298	220	201	149	70	38	15	8	1	0
Joseph E. Stiglitz	131	165	179	151	161	102	61	36	2	12
Andrei Shleifer	379	173	97	88	141	65	28	18	2	7
Robert J. Barro	167	171	147	128	162	69	70	47	10	23
Daron Acemoglu	16	110	139	150	126	133	124	90	26	66
Jean Tirole	1	24	63	88	45	151	145	156	113	136
Olivier J. Blanchard	0	3	2	10	0	9	33	40	87	79
Kenneth S. Rogoff	0	17	42	62	25	77	111	112	110	105
John Y. Campbell	0	18	34	56	20	67	93	121	105	106
Peter C. B. Phillips	7	59	60	61	62	77	54	69	69	68

Note: This table gives the number of times that the top 10 economists (according to the arithmetic mean) held a rank between 1 and 10 in the 1000 runs of the SA with Spearman's footrule distance.

Table 7: Rank distribution for the stochastic algorithm with Kendall's tau

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
James J. Heckman	306	232	175	142	73	43	18	8	1	2
Joseph E. Stiglitz	138	160	195	157	143	93	56	35	16	6
Andrei Shleifer	356	161	134	137	80	66	34	14	10	7
Robert J. Barro	158	215	177	138	115	77	62	33	13	8
Daron Acemoglu	32	95	124	146	159	133	108	84	53	38
Jean Tirole	1	17	40	56	81	153	169	166	132	102
Olivier J. Blanchard	0	1	1	3	7	20	27	40	68	98
Kenneth S. Rogoff	0	13	24	38	71	77	114	112	124	105
John Y. Campbell	0	11	18	29	56	89	101	113	128	94
Peter C. B. Phillips	7	61	54	61	69	70	58	70	50	75

Note: This table gives the number of times that the top 10 economists (according to the arithmetic mean) held a rank between 1 and 10 in the 1000 runs of the SA with Kendall's tau as a distance function.

To get an impression of the variations in the 1000 runs of the stochastic algorithm, we present Tables 6 and 7. They show how often which scientist held which rank.[12] For example, James J. Heckman was first in 298 runs of the stochastic algorithm (Spearman) and in 306 runs of the SA with Kendall.

In Table 8, the correlations between the different resulting rankings are presented, computed with Spearman's rank correlation coefficient. While there are some highly correlated rankings (for example, arithmetic and geometric mean, or the two versions of the stochastic

[12] As an overview, we only chose the top 10 scientists, according to the arithmetic mean, and only reported the first ten ranks.

algorithm, GA Spearman and GA Kendall), others show a lower correlation, like lexicographic and graphicolexic rank. These results should not be surprising, considering the different approaches taken by the algorithms above. As outlined in Section 2, some of the approaches are similar (like arithmetic and geometric mean in punishing lower ranks), while others value opposite aspects (like lexicographic and graphicolexic rank).

Table 8: Spearman rank correlations between the rankings

	Arithm.	Harm.	Geom.	Median	Lexicogr.	Graphicol.	GA S.	GA K.
Arithm.	1							
Harm.	0.672	1						
Geom.	0.724	0.650	1					
Median	0.724	0.551	0.749	1				
Lexicographic	0.493	0.514	0.525	0.479	1			
Graphicolexic	0.677	0.551	0.548	0.647	0.458	1		
GA Spearman	0.784	0.562	0.706	0.604	0.523	0.564	1	
GA Kendall	0.743	0.518	0.734	0.784	0.450	0.586	0.711	1

4 Conclusion

Given many different rankings to aggregate, we have presented eight approaches to do so. These can be classified into deterministic and stochastic methods. Depending on whether the aim is to have an easily computable method, or whether, for example, one wants to favor good or punish low ranks, the method of choice will vary. Due to its long running time, the stochastic algorithm is impractical for large data sets. Still, because of its properties of attempting to determine a 'compromise ranking' and of offering the distance and objective function as a comprehensible way to measure the quality of that compromise, it is a valuable approach. If the ranking aggregation problem is not too large, the stochastic algorithm should definitely be considered as an alternative. The lexicographic and graphicolexic rank should be treated with caution as they might only value a few really good or low ranks and thus lead to a somewhat distorted result.

References

Ben-David D (2010) Ranking Israel's economists. Scientometrics 82(2):351–364

Cook WD, Raviv T, Richardson AJ (2010) Aggregating incomplete lists of journal rankings: An application to academic accounting journals. Accounting perspectives 9(3):217–235

Franses PH (2014) Trends in three decades of rankings of dutch economists. Scientometrics 98(2):1257–1268

Harzing A, van der Wal R (2009) A Google Scholar h-index for journals: An alternative metric to measure journal impact in economics and business. Journal of the American Society for Information Science and Technology 60(1):41–46

Marginson S (2014) University rankings and social science. European Journal of Education 49(1):45–59

Pihur V, Datta S, Datta S (2009) Rankaggreg, an r package for weighted rank aggregation. BMC bioinformatics 10(1):62

Sculley D (2007) Rank aggregation for similar items. In: SDM, SIAM, pp 587–592

Seiler C, Wohlrabe K (2012) Ranking economists on the basis of many indicators: An alternative approach using repec data. Journal of Informetrics 6(2):389–402

Wohlrabe K (2016) Taking the temperature: A meta-ranking of economics journals. CESifo Working Paper 5726

Zimmermann C (2013) Academic Rankings with RePEc. Econometrics 1(3):249–280