an open access journal

RESEARCH ARTICLE

The rank boost by inconsistency in university rankings: Evidence from 14 rankings of Chinese universities

Wenyu Chen[1], Zhangqian Zhu[2], and Tao Jia[1]

[1]College of Computer and Information Science, Southwest University, Chongqing, 400715, P. R. China
[2]Department of National Defense Economy, Army Logistics University of Chinese People's Liberation Army, Chongqing, 500106, P. R. China

Keywords: Chinese universities, rank aggregation, rank-biased overlap, rank boost, rank similarity, university rankings

Citation: Chen, W., Zhu, Z., & Jia, T. (2020). The rank boost by inconsistency in university rankings: Evidence from 14 rankings of Chinese universities. *Quantitative Science Studies*, *2*(1), 335–349. https://doi.org/10.1162/qss_a_00101

DOI:
https://doi.org/10.1162/qss_a_00101

Supporting Information:
https://doi.org/10.1162/qss_a_00101

Received: 10 April 2020
Accepted: 17 August 2020

Corresponding Author:
Tao Jia
tjia@swu.edu.cn

Handling Editor:
Liying Yang

ABSTRACT

University ranking has become an important indicator for prospective students, job recruiters, and government administrators. The fact that a university rarely has the same position in different rankings motivates us to ask: To what extent could a university's best rank deviate from its "true" position? Here we focus on 14 rankings of Chinese universities. We find that a university's rank in different rankings is not consistent. However, the relative positions for a particular set of universities are more similar. The increased similarity is not distributed uniformly among all rankings. Instead, the 14 rankings demonstrate four clusters where rankings are more similar inside the cluster than outside. We find that a university's best rank strongly correlates with its consensus rank, which is, on average, 38% higher (towards the top). Therefore, the best rank usually advertised by a university adequately reflects the collective opinion of experts. We can trust it, but with a discount. With the best rank and proportionality relationship, a university's consensus rank can be estimated with reasonable accuracy. Our work not only reveals previously unknown patterns in university rankings but also introduces a set of tools that can be readily applied to future studies.

1. INTRODUCTION

The rank of a university is playing an increasingly important role not only in the choices by parents and students seeking for education but also in the decisions by funding agencies and policymakers who control where financial support will go (Hazelkorn, 2015; Zhang, Hua, & Shao, 2011). There is also evidence reported on the relationship between the rank of a university and the employment of its students (Bastedo & Bowman, 2011; Sayed, 2019). All of these provide an excellent opportunity for the development of university rankings. There are multiple rankings proposed by different agencies. Some are primarily based on one single index, such as the Nature Index (NI), which only considers the research output in a selective set of journals. But more often, the ranking relies on a comprehensive set of indicators related to the performance of the university. The different weights assigned to the indicators reflect the perspectives that the ranking focuses on. For example, the Academic Ranking of World Universities (ARWU) focuses more on the performance and achievements of research, the Quacquarelli Symonds World University Rankings (QS) is concerned more with a university's reputation and

The MIT Press

internationalization, and University Ranking by Academic Performance (URAP) and Performance Ranking of Scientific Papers for World Universities (NTU) only pay attention to the performance of scientific research (Çakır, Acartürk et al., 2015; Vernon, Balas, & Momani, 2018). These rankings give the relative quality of a university from a certain point of view. Yet, the existence of multiple rankings also naturally gives rise to an interesting question: How similar or how different are these rankings?

There have been intensive studies, either qualitative (Aguillo, Bar-Ilan et al., 2010; Angelis, Bassiliades, & Manolopoulos, 2019; Chen & Liao, 2012; Moed, 2017) or quantitative (Anowar, Helal et al., 2015; Çakır et al., 2015; Robinson-García, Torres-Salinas et al., 2014; Selten, Neylon et al., 2019; Vernon et al., 2018), on the comparisons of university rankings. Nevertheless, we wish to emphasize that the techniques required for a comprehensive understanding of this question are nontrivial. There are four characteristics in the university ranking, making it different from other ranking systems. First, the university ranking is top weighted. The difference between the first and second position is more significant than the difference between the ninety-ninth and the hundredth. Second, the university ranking is incomplete. The elements included in the ranking list are not identical, as a university may appear in some rankings but not all. Moreover, university rankings can be uneven. The length of different ranking lists is not the same when they consider a different set of universities. Finally, the rankings may contain ties, allowing multiple universities to occupy the same rank. All these features make some well-known and also frequently used metrics, such as Spearman's rank correlation (Chen & Liao, 2012; Moed, 2017; Shehatta & Mahmood, 2016; Soh, 2011), Kendall's τ distance (Angelis et al., 2019; Liu, Zhang, et al., 2011; Zhang, Liu, & Zhou, 2011), Spearman's footrule (Abramo & D'Angelo, 2016; Aguillo et al., 2010) incapable of accurately quantifying the similarities or differences among university rankings.

Despite different opinions presented by different studies, there might be one conclusion reached in common: It is very rare, if not impossible, that a university would hold the same position in all rankings. This prompts us to ask the second question: To what extent can a university's rank be raised in different rankings. It is straightforward to find a university's best rank, or the fluctuation of its rank (Liu et al., 2011; Shehatta & Mahmood, 2016; Shi, Yuan, & Song, 2017; Soh, 2011; Zhang et al., 2011). However, to quantitatively measure the boost, we also need a baseline that represents the "average" or consensus rank of this university by combining information on all rankings considered. This technique is called *rank aggregation*. Indeed, while there is a rich body of literature on the methodology of rank aggregation, such a tool is applied to university rankings in very limited cases (Wu, Zhang, & Lv, 2019). Most discussions of a university's rank boost still rely heavily on qualitative descriptions or methods, such as grouping similarly ranked universities (Liu & Liu, 2017; Shi et al., 2017).

In this paper, we aim to answer these two questions using 14 distinct rankings of Chinese universities that are widely accepted by the public of China. There are several reasons why we focus on Chinese universities. One is the number of rankings available. Besides the well-known global rankings such as ARWU and QS that take universities in China into consideration, there are also domestic rankings, such as the Wu Shulian Chinese University Ranking (WSL) and University Ranking of China by the Chinese Universities Alumni Association (CUAA). Rankings based on one single index, such as the Nature Index, are also reported by the mass media and considered by the public. All of them provide a large corpus of rankings to analyze. The focus on Chinese universities also allows us to establish a relatively clear set of subjects and avoid possible bias in some global rankings (e.g., a ranking may systematically put Chinese universities higher or lower on the list). Most importantly, there are few studies of the quantitative comparisons of Chinese university rankings, although the need for such investigations is self-evident.

Here, we utilize the recently proposed metric, Rank-biased Overlap (RBO) (Webber, Moffat, & Zobel, 2010), to quantify similarities among different rankings. RBO is claimed to be a nice measure for top-weighted, incomplete, and uneven rankings, making it very appropriate for our study. We also use the rank aggregation algorithm (Amodio, D'Ambrosio, & Siciliano, 2016) to identify a consensus ranking from all rankings, which allows us to further quantify the rank boost. We find that these rankings of Chinese universities are in general not similar. On the one hand, rankings lack agreement on the selection of top universities. On the other, they tend to put universities in different positions on the ranking list. If we focus on the universities in the 211 Project and analyze their relative ranks, we obtain an increased similarity, indicating that rankings have some certain foundations. However, the similarities are not uniformly distributed among rankings. In particular, when looking at the 43 universities in the 211 Project, we obtain four clusters by hierarchical clustering, where rankings are more similar to each other inside the cluster than outside. The existence of clusters reveals some previously unknown relationships among rankings. By comparing a university's best rank and its consensus rank, we find that the best rank is roughly 38% higher (towards the top). While a university's consensus rank is not always directly available, its best rank is likely to be used when introducing and advertising the university. Hence we can use this feature to infer a university's consensus rank, which demonstrates good accuracy. The finding implies that a university may find a better position if there are more rankings available, providing a plausible explanation of why there are a large number of university rankings in China. The inconsistency of rankings also raises concerns about the issue of reproducibility, which needs to be further discussed if university ranking is taken as serious science. Finally, we wish to note that the technical approach in this paper is very general and does not apply to Chinese universities only. Hence it has the potential to be applied to the ranking systems of other countries, which may offer new insights into questions related to university rankings.

2. DATA AND METHOD

2.1. Data Set

There are many rankings that include Chinese universities. The rankings used in this study are selected for their influence and public awareness. We also try to balance the number of global and national rankings. We choose 14 rankings, of which 11 are included in the IREG Inventory on International Rankings, and three national ones are commonly used domestically in China. They are ShanghaiRanking's Academic Ranking of World Universities (ARWU), Quacquarelli Symonds World University Rankings (QS), Times Higher Education World University Rankings (THE), U.S. News Best Global Universities Rankings (USNWR), Performance Ranking of Scientific Papers for World Universities (NTU), University Ranking by Academic Performance (URAP), Center for World University Rankings (CWUR), Nature Index (NI), SCImago Institutions Rankings (SIR), Webometrics Ranking of World Universities (WRWU), Universities Ranking of China by Chinese Universities Alumni Association (CUAA), Chinese university ranking by Wu Shulian (WSL), Research Center for Chinese Science Evaluation (RCCSE), and Best Chinese Universities Ranking (BCUR). Details of these rankings can be found in Table 1 and Table S1 in Supplementary Information. For ease of study, we focus on universities on the mainland of China. For international rankings, we consider the relative ranks of Chinese universities.

We collect rankings in the year 2017. Note that ranking agencies publish their results at different times of year, which are also based on indicators measured at a different period of time. To make the comparison fair enough, we use ranking results labeled as "2017" by the agency. For some international rankings, such as QS and WRWU, we select the Chinese universities from the

Table 1. The 14 university rankings applied in this study

Name	Abbr.	Scope	Tie (Y/N)	Number of Chinese institutions
Academic Ranking of World University	ARWU	Global	Y	90
Quacquarelli Symonds World University Rankings	QS	Global	N	82
Times Higher Education World University Rankings	THE	Global	N	63
U.S. News Best Global Universities Rankings	USNWR	Global	N	107
Performance Ranking of Scientific Papers for World Universities	NTU	Global	Y	65
University Ranking by Academic Performance	URAP	Global	Y	376
Center for World University Rankings	CWUR	Global	Y	97
Nature Index	NI	Global	N	493
SCImago Institutions Rankings	SIR	Global	Y	387
Webometrics Ranking of World Universities	WRWU	Global	N	1691
Universities Ranking of China by Chinese Universities Alumni Association	CUAA	Domestic	Y	723
Chinese university ranking (Wu Shulian)	WSL	Domestic	N	400
Research Center for Chinese Science Evaluation	RCCSE	Domestic	Y	136
Best Chinese Universities Ranking	BCUR	Domestic	Y	500

Asia regional rank, which gives a longer ranking list. For THE, which only provides the range of the rank for some universities, we recalculate the score based on the indicators of the ranking system to reach the relative rank. For RCCSE and CUAA, which include a variety of domestic university rankings, we select the overall ranking of Chinese universities' competitiveness for RCCSE, and the top 700 universities in China for CUAA.

Different universities may have different names in different rankings. We manually performed name disambiguation to clean the data. Different rankings contain a different number of Chinese universities as well, yielding different lengths of ranking list. For example, THE has only 63 Chinese universities, while WRWU includes 1691. We choose a top-k list from every ranking list. The similarity measure applied in this paper does not require a uniform length for every list. Hence it is fine if k exceeds the maximum length of the list. We report results based on k = 100 in the main text of the paper. Results based on k = 60, 80 and 120 can be found in the Supplementary Information.

When measuring the pairwise similarity between rankings, we also consider a different choice of data by focusing only on universities in the 211 Project, often called *211-universities*. The 211 Project is initiated by the Chinese Ministry of Education, aiming to build high-level universities in China. There are 116 universities in the 211 Project. However, North China Electric Power University (Beijing) and North China Electric Power University (Baoding), which are two distinct universities in the 211 Project, are sometimes considered as one university in several rankings. So we use 115 universities in this work by combing North China Electric Power University (Beijing) and North China Electric Power University (Baoding) together. If a ranking contains both of them, we choose the one with a higher rank. When considering rankings among 211-universities, we use their relative ranks and ignore other Chinese universities that are not in the 211 Project.

There are a lot of universities appearing only once or twice in one top-100 rank list (Figure S2 and Table S5). While this does not affect the measure of pairwise similarity, it will influence the study of the rank boost. It is less meaningful to analyze a university's best rank and consensus rank if it is contained in only a few lists. For this reason, we filter out universities that appear in fewer than four top k lists, a similar approach to that in Cook, Raviv, and Richardson (2010) that ensures data quality and aggregation results. Consequently, we obtain a new set of rank lists from the top-k lists, which we call *top-k-filtered* lists. The best rank and the consensus rank are based on a university's relative ranks in the top-k-filtered lists (see Supplementary Note 1 for more information). In general, the fit of the data varies only slightly when a different set of lists are analyzed.

2.2. Similarity Measure

As mentioned in the introduction, the university ranking is top weighted, incomplete, uneven, and with ties. Therefore, traditional approaches and indicators for similarity measurement may not work in this scenario (Wang, Ran, & Jia, 2020). For example, Spearman's rank correlation, its variant Spearman's footrule, and Kendall's τ distance only work in ranks with identical size and elements. They do not give a higher weight to the top-ranked elements either.

Bar-Ilan's M measure (Bar-Ilan, Levene, & Lin, 2007) can handle lists with different elements or with different lengths. A top ranked element is also given a higher weight characterized by an inversely proportional function. Overall, it is a very nice measure compared with many others, and is also frequently used in some recent studies, especially in some comprehensive comparisons (Aguillo et al., 2010; Çakır et al., 2015; Selten et al., 2019). One limitation of Bar-Ilan's M measure is that its physical interpretation is not very clear. Moreover, as the preference to the top elements is fixed by the inversely proportional function, one cannot tune the top-weightiness to check the robustness of the conclusion or the sensitivity to the top-weightiness. Because the final value is normalized by the maximum, it is less straightforward to compare M measures of lists with different lengths. Finally, the way to handle ties in the Bar-Ilan's M measure is conceptually tricky, although it is practically convenient.

In this work, we utilize a recently proposed measure (Webber et al., 2010), called Rank-biased Overlap (RBO), which is claimed to be able to sufficiently handle top-weighted, incomplete, and indefinite ranking lists. The definition of RBO can be best illustrated when the length of the list is infinitely long. Assume that S and L are two infinite rankings. Denote $S_{1:d}$ by the set of elements from position 1 to position d. The size of the overlap between lists S and L to depth d can be calculated by the intersection of the two sets as $X_{S,L,d} = |S_{1:d} \cap L_{1:d}|$. The agreement, measured by the proportion of the overlap at depth d, is given by $X_{S,L,d}/d$, namely X_d/d. One can assign different weight w_d to the agreement at depth d forming a similarity measure such that $\mathrm{SIM}(S, L) = \sum_{d=1}^{\infty} w_d X_d/d$.

The RBO takes the weight $w_d = (1 - p)p^{d-1}$, leading to the similarity measure for the infinite lists as

$$RBO(S, L, p) = (1-p)\sum_{d=1}^{\infty} p^{d-1}\frac{X_d}{d}. \tag{1}$$

Note that the form of w_d is the same as the geometric distribution function. Therefore, *RBO* (S, L, p) has a physical meaning. Consider that one selects a top-k list from each of S and L and calculate their agreement X_d/d, where the length k is randomly drawn from a geometric distribution with parameter p. $RBO(S, L, p)$ can thus be interpreted as the expected percentage of the overlap under such comparison. The average length of the top-k list extracted to compare is $1/(1 - p)$. Therefore, the extent of the top-weightiness can be tuned by the parameter p. A small p value means that we tend to select a list with short length, hence giving more weight to the top-ranked items. If p is so large that we are effectively comparing the common elements of the two full lists, the order of these elements can not be quantified. Therefore, p needs to be chosen based on the length of the lists in the analysis. In this work, we choose $p = 0.98$, $p = 0.95$, and $p = 0.9$ for the top-100 lists and the comparison of the lists by 211-universities, corresponding to the average comparison at lengths 50, 20, and 10, respectively. $p = 0.95$ and $p = 0.9$ is used for the last comparison when we select the 43 universities that are included in all of the 14 rankings. As p increases, the RBO measure will be overall higher. But in general our conclusion does not change.

Eq. (1) gives the ideal case when S and L are infinite. But in general their lengths are finite. Let L be the longer list of the two, with length l, and S be the shorter one, with length s. The formula we applied in calculation is

$$RBO(S, L, p) = \frac{1-p}{p} \cdot \left(\sum_{d=1}^{l} \frac{X_d}{d} \cdot p^{d-1} + \sum_{d=s+1}^{l} \frac{X_s \cdot (d-s)}{s \cdot d} \cdot p^{d} \right) + \left(\frac{X_l - X_s}{l} + \frac{X_s}{s} \right) \cdot p^{l} \qquad (2)$$

When the ranking has ties, we need to change X_d/d into $2X_d/(|S_{1:d}| + |L_{1:d}|)$. More details can be found in Webber et al. (2010).

2.3. Hierarchical Clustering Method

Hierarchical clustering is an unsupervised machine learning algorithm that merges similar items into groups. At each step, two units that are closest are merged together, forming a new unit. This process is repeated iteratively until all items in the system are merged together into one unit, giving rise to a dendrogram formed from the bottom to the top. In our work, the distance is calculated based on the pairwise similarity between two university rankings. Assuming there are two units A and B, the distance between them is calculated by averaging the similarities of their elements as

$$D(A, B) = 1 - \frac{\sum_{a \in A} \sum_{b \in B} s_{a,b}}{|A|\,|B|}, \qquad (3)$$

where $s_{a,b}$ is the pairwise similarity between ranking a and b.

2.4. Rank Aggregation Method

Rank aggregation is also known as *Kemeny rank aggregation* (Snell & Kemeny, 1962), *preference aggregation* (Davenport & Kalagnanam, 2004) and *consensus ranking* (Amodio et al., 2016), which aims to integrate multiple rankings into one comprehensive ranking (Wu et al., 2019). It has been applied in recommendation systems (Meila, Phadnis et al., 2012), meta-search (Dwork, Kumar et al., 2001), journal ranking (Cook et al., 2010), and proposal selection (Cook, Golany et al., 2007). There are multiple rank aggregation methods. Some are heuristic, combing rankings based on simple rules of thumb, such as Borda count. Some aim to minimize the average distance (Cohen-Boulakia, Denise, & Hamel, 2011), the number of violations (Pedings, Langville, & Yamamoto, 2012), or to optimize the network structure (Xiao, Deng et al., 2019). While different aggregation algorithms all claim to be superior to existing ones when proposed, the baseline algorithms and the testing samples are all different from case to case. Although there are some reviews or comparisons of aggregation methods (Brancotte et al., 2015; Li, Wang, & Xiao, 2019;

Xiao, Deng et al., 2017), most of them cover only a few algorithms and the conclusions may not be general enough. Indeed, it was unclear which method is most appropriate to aggregate a small number of long lists that are incomplete, uneven, and with ties.

In a recently study, we performed a comprehensive test on nine rank aggregation algorithms. We introduced a variation of Mallows model (Irurozki, Calvo et al., 2016) to generate synthetic ranking lists whose physical property is known and tuned for different circumstances (Chen, Zhu et al., 2020). The synthetic ranking lists provide us the ground truth for comparisons, where we find that the branch and bound algorithm FAST by Amodio et al. (2016) is most appropriate in our task. More details of the FAST algorithm can be found in the original paper and the code can also be downloaded as an R Package "ConsRank" (D'Ambrosio, Amodio, & Mazzeo, 2015).

3. RESULTS

We collect the top-k (k = 100) list from each of the rankings and measure their pairwise similarity using the measurement RBO. Given that nine out of the 14 rankings contain more than 100 Chinese universities, k = 100 is a reasonable choice to gauge the overall similarity. We first choose the parameter p = 0.98 in the RBO measure, meaning that the list overlap is quantified on average at length 50. A larger p value would not be meaningful given that the smallest list length is only 63 (THE). We find that these rankings in general are not similar to each other. Most pairs (90% of them or 82 out of 91) have an RBO similarity below 0.5 and the average RBO is 0.39 (Figure 1a and Table S2 in Supplementary Information), which can be very roughly interpreted, as they on average have only 39% of overlap (Webber et al., 2010). The similarity is even lower if we use p = 0.9, which gives more weights on the top 10 elements (Figure 1b and Table S3 in Supplementary Information). The finding does not change if we vary the length of the top-k list (see results for other k values in Figure S1 of Supplementary Information). Overall, different Chinese university rankings are not very consistent, in line with the claims of other similar studies (Çakır et al., 2015; Liu et al., 2011; Selten et al., 2019; Wu et al., 2019).

Two factors can be associated with the low similarity among rankings. On the one hand, the rankings lack an agreement on the selection of top candidates (Angelis et al., 2019; Soh, 2011). If we consider the union of all these 14 top-100 lists, there are a total of 167 different universities (Table S5). For example, NTU contains only 65 Chinese universities. But the top

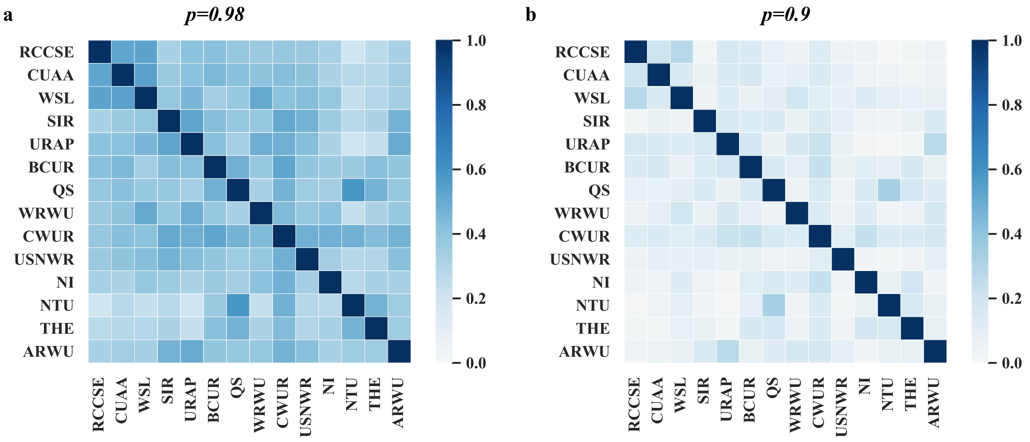

Figure 1. The heat map of the pairwise RBO similarity among the top-100 lists of 14 university rankings. Panel a: p = 0.98; Panel b: p = 0.9.

65 universities of NTU are not fully included in the top 100 of CUAA or NI. Hence, even though the relative ranks of two universities are the same in two rankings (i.e., A ranks higher than B), their actual positions can have a drastic difference (e.g., B ranks 65 in QS and 100 in WSL), giving rise to a low similarity measure between the two. On the other hand, these rankings indeed rank universities in a different manner, which is a more inherent reason for the rank inconsistency. For instance, USNWR and URAP have 86 universities in common in their choice of top-100 (Table 2), showing a relatively good consensus on the top candidates. But their similarity is below 0.5 when $p = 0.98$ and below 0.1 when $p = 0.9$ (Tables S2 and S3 in Supplementary Information).

The lack of agreement on the top-k universities motivates us to perform another comparison by fixing the set of universities to analyze. Here we choose universities in the 211 Project and use their relative ranks in the 14 different rankings. The corresponding RBO measure increases overall, implying that ranking agencies have more agreement on the relative rank of the 211-universities. The increased similarities are not uniformly distributed: some rankings become very close but some remain distant from each other (Figure 2). To identify the underlying structure of the similarity relationship, we perform hierarchical clustering on the similarity matrix (see Section 2) . The obtained dendrogram suggests the existence of two clusters. One consists of RCCSE, WSL, BCUR, CUAA, WRWU, NI, SIR, and URAP. The other consists of NTU, THE, QS, ARWU, USNWR, and CWUR. Rankings are more similar to each other inside the cluster than outside.

It is noteworthy that the numbers of 211-universities in each ranking are different. Indeed, the clustering effect coincides with the length difference of the ranking lists. The eight rankings in the same cluster contain more than 100 universities, while the six rankings in the other cluster have far fewer. Therefore, it is unclear if the clusters are a result of the preference of ranking

Table 2. The number of overlapping elements between two rankings

	CUUA	WSL	SIR	URAP	BCUR	QS	WRWU	CWUR	USNWR	NI	NTU	THE	ARWU
RCCSE	87	86	75	78	79	76	81	76	78	74	58	57	70
CUAA		89	76	78	79	75	81	75	79	73	58	55	71
WSL			81	84	77	75	88	81	84	79	61	58	74
SIR				89	77	70	83	86	87	81	64	56	80
URAP					78	70	86	85	86	79	61	54	83
BCUR						74	78	77	78	73	60	58	72
QS							73	66	73	66	56	53	65
WRWU								82	83	80	60	59	73
CWUR									85	83	64	56	77
USNWR										77	63	56	77
NI											60	54	71
NTU												49	61
THE													55

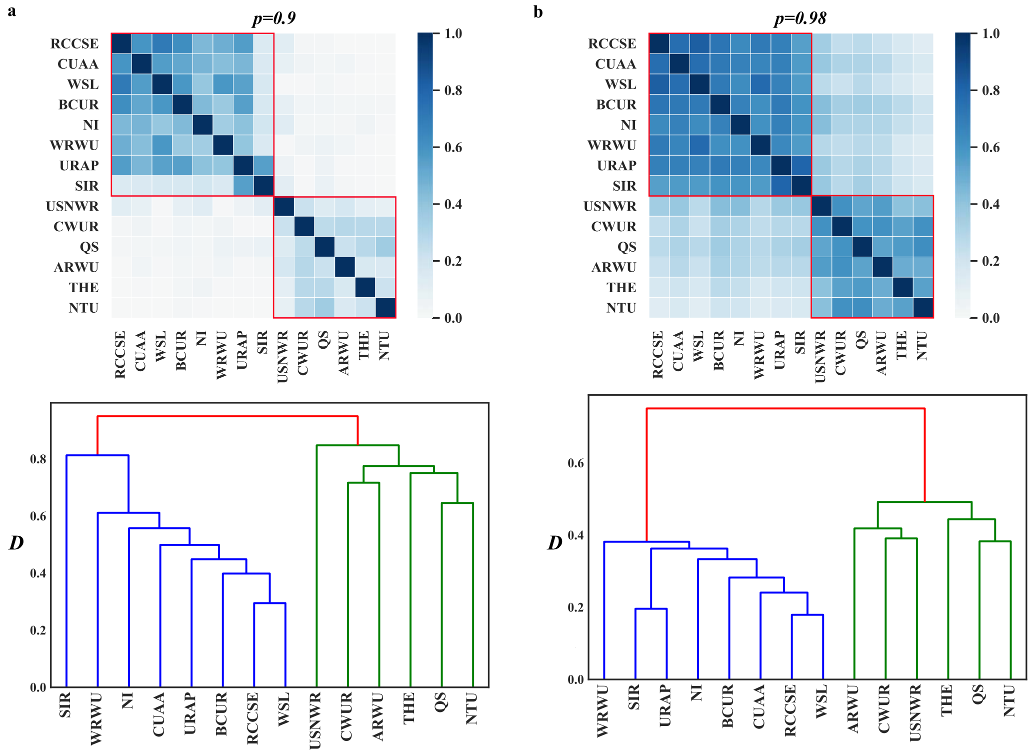

Figure 2. The heat map of similarities based on lists by 211-universities and the dendrogram by hierarchical clustering on the similarity matrix. Panel a: p = 0.9; Panel b: p = 0.98.

agency or the different number of universities. To eliminate the length difference, we perform another test by using only the 211-universities that are included in all of the 14 rankings. We end up with 14 lists containing 43 universities. The pairwise similarity is measured and the hierarchical clustering on the similarity matrix is performed (Figure 3). We have four clusters emerging from the overall increased RBO value. Given that different rankings choose different sets of indicators in their methodology, and that the values of these indicators are not generally publicly available, in addition to the correlations among indicators, it is impossible for us to give a quantitative explanation of why we end up with four clusters, not five or three. It is also hard to explicitly explain why the two rankings are in the same cluster, not with others. Nevertheless, we can still find some clues using the general information of the ranking. It is not surprising that WSL, RCCSE, and CUAA are in the same cluster, because they are the major domestic rankings focusing only on Chinese universities. Some indicators, such as the teaching quality and discipline of the university, are only used by them. SIR and URAP are in the same cluster likely because they all rely on the volume of scientific output, such as the number of papers published and the total number of citations received. NTU and ARWU are in the same cluster because they consider indicators based on more selective scientific output, such as papers in prestigious journals, with high citations, and faculties with international awards. Although BCUR focuses on Chinese universities, it is not in the same cluster as the other three domestic rankings. This may be related to the fact that BCUR uses a comprehensive set of indicators, including academic research, talent training, and social services.

So far, we have performed two types of measurements on university rankings. One gauges the overall similarity and the other considers the similarity for a fixed set of universities. The

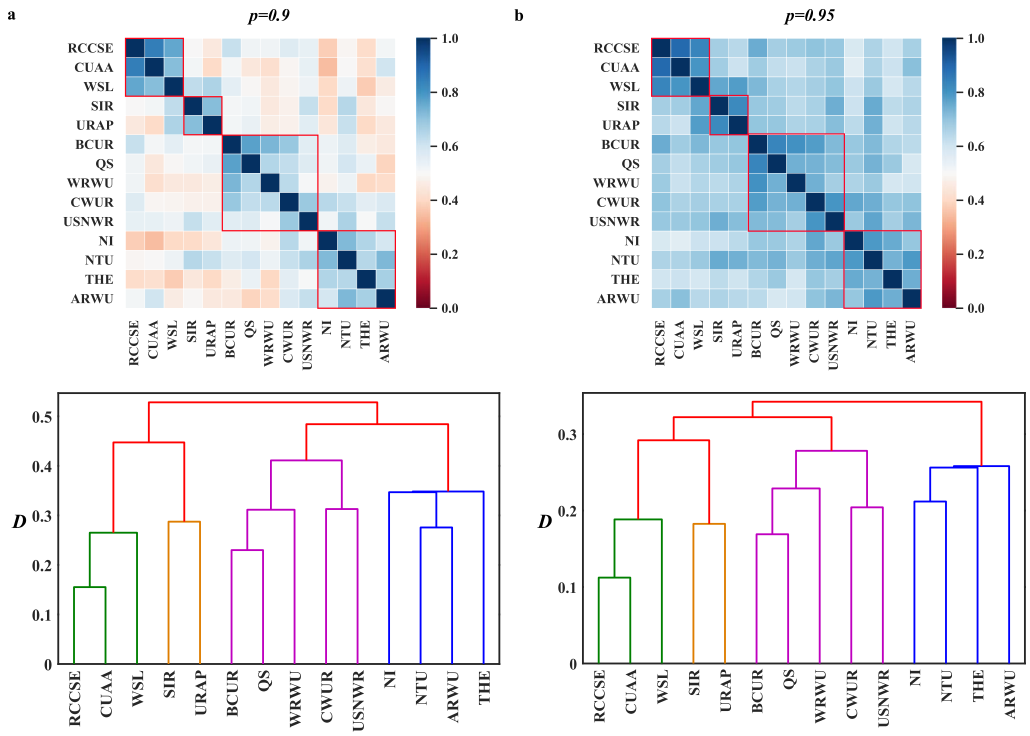

Figure 3. The heat map of similarities based on the 43 universities in the 211 Project that are included in all14 rankings, and the dendrogram by hierarchical clustering. The clusters obtained are the same for different parameters of RBO metric. Panel a: $p = 0.9$; Panel b: $p = 0.95$.

results show that the relative position of the 211-universities is relatively stable. The 14 rankings fall into different clusters within which they are very similar to each other. However, other candidates can fill in the relative rankings of 211-universities in different ways. Therefore, a Chinese university's national rank can vary significantly and the top-*k* lists of the university rankings are not consistent. To show an example, we list the top 20 universities by CUAA and their ranks in other rankings (Table S4 in the Supplementary Information). The rank fluctuation in different rankings is nonnegligible. Renmin University of China, for instance, ranks 8th in the CUAA but 130th in the SIR. The fact that a university can rank higher or lower in different rankings prompts us to ask another interesting question: To what extent could a Chinese university's national rank be raised?

Although it is easy to identify a university's best rank or the range of the rank fluctuation, a quantitative answer to the above question is nontrivial. To calculate the rank boost, we need not only the best but also a "true" rank of a university as the reference. Any rank alone is insignificant to represent the collective information of the multiple rankings. To cope with this issue, we apply the rank aggregation technique (see Section 2) FAST to generate an aggregated or consensus rank as the baseline. We then calculate the rank boost of a university as

$$\Delta = P_{\mathrm{AR}} - P_{\mathrm{Best}}, \tag{4}$$

where P_{AR} is the position of a university in the aggregated ranking and P_{best} is the best rank in the 14 university rankings.

We find that the rank boost Δ is not constant. Instead, it is linearly correlated with a university's aggregated rank P_{AR} (Figure 4a, $R^2 = 0.9$). On average, the rate of proportionality is 0.38. The

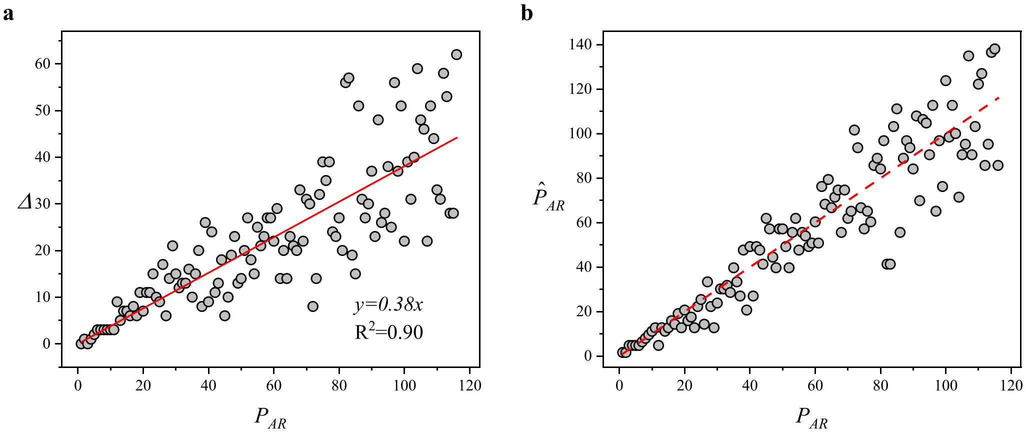

Figure 4. Panel a: The relationship between a university's consensus rank P_{AR} and its rank boost Δ. The red line corresponds to the linear fit. Panel b: The estimate of consensus rank $\hat{P}_{\mathrm{AR}}$ is very close to the true consensus rank P_{AR}. The dashed line is $y = x$.

linear dependence pattern is robustly observed in different tests and the slope varies only slightly from 0.38 (Figures S3 and S4 in the Supplementary Information). In other words, a university can find itself in a more preferred position if there are multiple rankings to choose. Such a rise, however, is not unbounded or a fixed value. Instead, it is proportional to a university's collective position given by multiple rankings. A university's best rank reflects its consensus rank, but is 38% higher (towards the top).

The observation of the rank boost not only reveals an important pattern in the collective information drawn from different ranking systems but also leads to practical applications. Indeed, a university or a researcher may follow multiple rankings and be able to calculate the consensus ranking, but such information is generally unknown to the public. The information usually displayed on the front page of a university is its best rank. Using the correlation discovered, one can do an inverse calculation and estimate a university's consensus rank from its best rank using the equation $\hat{P}_{\mathrm{AR}} = P_{\mathrm{Best}}/0.62$. Indeed, the estimate agrees very well with the true consensus rank, with a mean percentage error of less than 3% (Figure 4b). It is noteworthy that a university's consensus rank relies on not only its positions in all rank lists but also the positions of its peers. In other words, one cannot directly find a university's consensus rank alone, but must find the consensus rank list first. But using the pattern uncovered, we can estimate a university's consensus rank using only the best rank of this university.

4. CONCLUSION

To summarize, we analyze 14 Chinese university rankings, which are some of the largest of their kind. Using RBO as the similarity measure, we find that these university rankings are not similar to each other in general. This discrepancy is caused by the lack of agreement on the choice of top universities, and that the top selected universities are ranked differently. But this does not mean those rankings do not have any foundation. When we focus on the 211-universities and use their relative ranks, we find that those rankings are more consistent. If we compare the whole ranking list, the 14 rankings fall into two clusters roughly divided by the number of 211-universities a ranking contains. If we focus on the 43 universities included in all of the 14 rankings, the pairwise similarity indicates that these rankings fall into four clusters. In general, Chinese university rankings have a certain degree of consensus on which 211-universities should be ahead of others. But there are a different number of other universities among the relative ranking of 211-universities, and in different orders. Eventually, these rankings become dissimilar.

Given the inconsistency of rankings, a university may rank higher or lower depending on which ranking is considered, which prompts us to further explore the extent to which the rank is raised. We apply the rank aggregation method to generate the consensus ranking by combining information in the 14 rankings. Using it as the baseline, we measure the rank boost, quantified as the difference between a university's best rank and consensus rank. The rank boost is linearly correlated with the consensus rank. The statistics tell that a university should be able to find itself on a preferred rank list where its position is, on average, 38% better. The rank information on the front page of a university website is not nonsense at all. Rather, it adequately reflects the collective opinion of experts. We should trust it, but with a discount. With the best rank and proportionality relationship, we can estimate a university's consensus rank with good accuracy.

Our findings provide some new perspectives. We may ask: Is the university ranking a science or a business? On the one hand, the ranking is performed by experts in the area, built on carefully and reasonably chosen indicators that are quantitatively measured. There are also intensive scientific studies on the framework of the ranking, the comparison of different rankings, and the validity of the indicators. There must be science in it. On the other hand, reproducibility, a crucial element in science, is largely missing. While the relative ranks among some universities are stable to a certain extent, which serves as the cornerstone of the ranking system, a university's actual position is not consistent in different rankings. It is fine that an individual ranking selects a different set of indicators, aiming to reveal a unique aspect of ground truth. Yet, it is still awkward to notice that two rankings rarely reach a consensus. This makes us hypothesize that the university ranking is also a business (Vernon et al., 2018). Indeed, the workload required to select and collect information for over 1,000 universities worldwide cannot be easily carried out by a small group of scientists. The tremendous effort devoted by the ranking agencies makes it necessary to draw the public's attention, which favors distinction rather than similarity. This makes university rankings different from pure science. A scientist is willing to reproduce existing results with a different set of analyses, which that confirms the validity of the known findings, from where new questions can be explored. Yet, ranking agencies are reluctant, if at all, to reproduce the same or just similar enough ranking to one already proposed. Nor can anyone risk even mentioning that the new ranking is generally in line with another. This hypothesis is further supported by our finding of the rank boost, which implies that different rankings are in favor of different universities.

In the context of the recent focus on the development and challenge of science in China (Liu, Yu et al., 2021; Yang, Fukuyama, & Song, 2018), our findings on Chinese university rankings provide new insights into this topic. But this work also has a few limitations that may be addressed in future work. First, while we observe the clustering among multiple rankings, it is relatively unclear why rankings are in the same cluster or how the number of rankings emerges. Indeed, multiple factors can affect the clustering result. The different clusters in Figures 2 and 3 already demonstrate the impact of ranking length. But how other factors, such as the set of indicators or the weights of indicators, are related to the clustering needs further investigation. The ranking data we collected are labeled as "2017," presumably reflecting the rankings in the year 2017. But this may not be entirely true. For example, WSL2017 was proposed in 2017, but QS2017 was introduced in 2016. It would be interesting to study the alignment of different rankings by different agencies proposed in different years. We also note that the ranking may change between years, including not only the rank of a university but also the set of universities included in the ranking. The evolution of ranking over time may merit further investigations (Garcia-Zorita, Rousseau et al., 2018). We accept that the value 38% for the rank boost would definitely change when including a new ranking list or using the data from another year. Hence it is meaningful to perform similar analyses in different years to identify the range of the boost. It would also be meaningful to check if the pattern of the rank boost remains the same in countries other than China. The results obtained may lead to the discovery of some universal patterns in university rankings. This is of particular importance given recent research on the "science of science," which uncovers many universal patterns underlying how science is performed and organized (Fortunato, Bergstrom et al., 2018; Jia, Wang, & Szymanski, 2017; Pan, Petersen et al., 2018; Wu, 2019; Wu, Wang, & Evans, 2019). Taken together, the work not only presents new patterns in rankings of Chinese universities but also introduces a set of tools that have not been utilized in related studies. These tools and technical approaches are very general and are not restricted to Chinese universities only, and can be easily applied to a variety of ranking systems and problems (Liao, Mariani et al., 2017).

AUTHOR CONTRIBUTIONS

Wenyu Chen: Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing—Original draft. Zhangqian Zhu: Methodology, Software, Formal analysis, Investigation, Data curation. Tao Jia: Conceptualization, Methodology, Supervision, Writing—Original draft, Writing—Review & editing.

COMPETING INTERESTS

The authors have no competing interests.

FUNDING INFORMATION

The work is supported by the National Natural Science Foundation of China (No. 61603309) and Fundamental Research Grant of Educational Science of the Ministry of Education in 2020.

DATA AVAILABILITY

The rankings used in this study are obtained from public data, whose urls are provided in Table Sl of the Supplementary Information. The list of universities and their ranks are given in Table S5 of the Supplementary Information, which are sufficient to reproduce our results.

REFERENCES

Abramo, G., & D'Angelo, C. A. (2016). A comparison of university performance scores and ranks by MNCS and FSS. *Journal of Informetrics*, *10*(4), 889–901. **DOI:** https://doi.org/10.1016/j.joi.2016.07.004

Aguillo, I., Bar-Ilan, J., Levene, M., & Ortega, J. (2010). Comparing university rankings. *Scientometrics*, *85*(1), 243–256. **DOI:** https://doi.org/10.1007/s11192-010-0190-z

Amodio, S., D'Ambrosio, A., & Siciliano, R. (2016). Accurate algorithms for identifying the median ranking when dealing with weak and partial rankings under the Kemeny axiomatic approach. *European Journal of Operational Research*, *249*(2), 667–676. **DOI:** https://doi.org/10.1016/j.ejor.2015.08.048

Angelis, L., Bassiliades, N., & Manolopoulos, Y. (2019). On the necessity of multiple university rankings. *COLLNET Journal of Scientometrics and Information Management*, *13*(1), 11–36. **DOI:** https://doi.org/10.1080/09737766.2018.1550043

Anowar, F., Helal, M. A., Afroj, S., Sultana, S., Sarker, F., & Mamun, K. A. (2015). A critical review on world university ranking in terms of top four ranking systems. In *New trends in networking, computing, e-learning, systems sciences, and engineering* (pp. 559–566). Cham: Springer. **DOI:** https://doi.org/10.1007/978-3-319-06764-3_72

Bar-Ilan, J., Levene, M., & Lin, A. (2007). Some measures for comparing citation databases. *Journal of Informetrics*, *1*(1), 26–34. **DOI:** https://doi.org/10.1016/j.joi.2006.08.001

Bastedo, M. N., & Bowman, N. A. (2011). College rankings as an interorganizational dependency: Establishing the foundation for strategic and institutional accounts. *Research in Higher Education*, *52*(1), 3–23. **DOI:** https://doi.org/10.1007/s11162-010-9185-0

Brancotte, B., Yang, B., Blin, G., Cohen-Boulakia, S., Denise, A., & Hamel, S. (2015). Rank aggregation with ties: Experiments and analysis. *Proceedings of the VLDB Endowment*, *8*(11). **DOI:** https://doi.org/10.14778/2809974.2809982

Çakır, M. P., Acartürk, C., Alaşehir, O., & Çilingir, C. (2015). A comparative analysis of global and national university ranking systems. *Scientometrics*, *103*(3), 813–848. **DOI:** https://doi.org/10.1007/s11192-015-1586-6

Chen, K.-H., & Liao, P.-Y. (2012). A comparative study on world university rankings: A bibliometric survey. *Scientometrics*, *92*(1), 89–103. **DOI:** https://doi.org/10.1007/s11192-012-0724-7

Chen, W.-Y., Zhu, Z.-Q., Wang, X.-M., & Jia, T. (2020). Comparison of performance of rank aggregation algorithms in aggregating a small number of long rank lists (in Chinese). *Acta Physica Sinica*, *69*(8), 080201. **DOI:** https://doi.org/10.7498/aps.69.20191584

Cohen-Boulakia, S., Denise, A., & Hamel, S. (2011). Using medians to generate consensus rankings for biological data. In *Scientific and statistical database management* (pp. 73–90). Cham: Springer. **DOI:** https://doi.org/10.1007/978-3-642-22351-8_5

Cook, W. D., Golany, B., Penn, M., & Raviv, T. (2007). Creating a consensus ranking of proposals from reviewers' partial ordinal rankings. *Computers & Operations Research*, *34*(4), 954–965. **DOI:** https://doi.org/10.1016/j.cor.2005.05.030

Cook, W. D., Raviv, T., & Richardson, A. J. (2010). Aggregating incomplete lists of journal rankings: An application to academic accounting journals. *Accounting Perspectives*, *9*(3), 217–235. **DOI:** https://doi.org/10.1111/j.1911-3838.2010.00011.x

Davenport, A., & Kalagnanam, J. (2004). A computational study of the Kemeny rule for preference aggregation. In *Proceedings of the 19th National Conference on Artificial Intelligence* (pp. 697–702). Palo Alto, CA: AAAI Press. https://dl.acm.org/doi/abs/10.5555/1597148.1597260

Dwork, C., Kumar, R., Naor, M., & Sivakumar, D. (2001). Rank aggregation methods for the web. In *Proceedings of the 10th International Conference on World Wide Web* (pp. 613–622). **DOI:** https://doi.org/10.1145/371920.372165

D'Ambrosio, A., Amodio, S., & Mazzeo, G. (2015). ConsRank: Compute the median ranking(s) according to the Kemeny's axiomatic approach. *R package version*, *1*(2). https://rdrr.io/cran/ConsRank/

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., … Barabási, A. L. (2018). Science of science. *Science*, *359*(6379), eaao0185. **DOI:** https://doi.org/10.1126/science.aao0185, **PMID:** 29496846, **PMCID:** PMC5949209

Garcia-Zorita, C., Rousseau, R., Marugan-Lazaro, S., & Sanz-Casado, E. (2018). Ranking dynamics and volatility. *Journal of Informetrics*, *12*(3), 567–578. **DOI:** https://doi.org/10.1016/j.joi.2018.04.005

Hazelkorn, E. (2015). Rankings and quality assurance: Do rankings measure quality. *Policy Brief*, *4*. Washington, DC: CHEA International Quality Group.

Irurozki, E., Calvo, B., Lozano, J. A., et al. (2016). Permallows: An R package for mallows and generalized mallows models. *Journal of Statistical Software*, *71*(12), 1–30. **DOI:** https://doi.org/10.18637/jss.v071.i12

Jia, T., Wang, D., & Szymanski, B. K. (2017). Quantifying patterns of research-interest evolution. *Nature Human Behaviour*, *1*(4), 1–7. **DOI:** https://doi.org/10.1038/s41562-017-0078

Li, X., Wang, X., & Xiao, G. (2019). A comparative study of rank aggregation methods for partial and top ranked lists in genomic applications. *Briefings in Bioinformatics*, *20*(1), 178–189. **DOI:** https://doi.org/10.1093/bib/bbx101, **PMID:** 28968705, **PMCID:** PMC6357556

Liao, H., Mariani, M. S., Medo, M., Zhang, Y.-C., & Zhou, M.-Y. (2017). Ranking in evolving complex networks. *Physics Reports*, *689*, 1–54. **DOI:** https://doi.org/10.1016/j.physrep.2017.05.001

Liu, L., & Liu, Z. (2017). On the rank changes of world-class universities in the past 10 years and its enlightenment (in Chinese). *Research in Higher Education of Engineering*, *2017*(3), 179–182.

Liu, L., Yu, J., Huang, J., Xia, F., & Jia, T. (2021). The dominance of big teams in China's scientific output. *Quantitative Science Studies*. Advance publication. **DOI:** https://doi.org/10.1162/qss_a_00099

Liu, Z., Zhang, S., Gao, Y., & Zhou, X. (2011). A correlation analysis on the sequences of the three world university rankings–WRWU, QS & ARWU: An empirical study based on twenty-two China's domestic universities ranking results (in Chinese). *Education Science*, *27*(1), 40–45.

Meila, M., Phadnis, K., Patterson, A., & Bilmes, J. A. (2012). Consensus ranking under the exponential model. *arXiv preprint* arXiv:1206.5265. https://arxiv.org/abs/1206.5265

Moed, H. F. (2017). A critical comparative analysis of five world university rankings. *Scientometrics*, *110*(2), 967–990. **DOI:** https://doi.org/10.1007/s11192-016-2212-y

Pan, R. K., Petersen, A. M., Pammolli, F., & Fortunato, S. (2018). The memory of science: Inflation, myopia, and the knowledge network. *Journal of Informetrics*, *12*(3), 656–678. **DOI:** https://doi.org/10.1016/j.joi.2018.06.005

Pedings, K. E., Langville, A. N., & Yamamoto, Y. (2012). A minimum violations ranking method. *Optimization and Engineering*, *13*(2), 349–370. **DOI:** https://doi.org/10.1007/s11081-011-9135-5

Robinson-García, N., Torres-Salinas, D., López-Cózar, E. D., & Herrera, F. (2014). An insight into the importance of national university rankings in an international context: the case of the I-UGR rankings of Spanish universities. *Scientometrics*, *101*(2), 1309–1324. **DOI:** https://doi.org/10.1007/s11192-014-1263-1

Sayed, O. H. (2019). Critical treatise on university ranking systems. *Open Journal of Social Sciences*, *7*(12), 39–51. **DOI:** https://doi.org/10.4236/jss.2019.712004

Selten, F., Neylon, C., Huang, C.-K., & Groth, P. (2019). A longitudinal analysis of university rankings. *Quantitative Science Studies*, *1*(3), 1109–1135. **DOI:** https://doi.org/10.1162/qss_a_00052

Shehatta, I., & Mahmood, K. (2016). Correlation among top 100 universities in the major six global rankings: Policy implications. *Scientometrics*, *109*(2), 1231–1254. **DOI:** https://doi.org/10.1007/s11192-016-2065-4

Shi, Y., Yuan, X., & Song, G. (2017). A comparative and empirical study of world university ranking system based on ARWU (in Chinese). *Library and Information Service*, *61*(15), 95–103.

Snell, J. L., & Kemeny, J. (1962). *Mathematical models in the social sciences*. Boston, MA: Ginn.

Soh, K. C. (2011). World university rankings: Take with a large pinch of salt. *European Journal of Higher Education*, *1*(4), 369–381. **DOI:** https://doi.org/10.1080/21568235.2012.662837

Vernon, M. M., Balas, E. A., & Momani, S. (2018). Are university rankings useful to improve research? A systematic review. *PLOS ONE*, *13*(3). **DOI:** https://doi.org/10.1371/journal.pone.0193762, **PMID:** 29513762, **PMCID:** PMC5841788

Wang, X., Ran, Y., & Jia, T. (2020). Measuring similarity in co-occurrence data using ego-networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *30*(1), 013101. **DOI:** https://doi.org/10.1063/1.5129036, **PMID:** 32013468

Webber, W., Moffat, A., & Zobel, J. (2010). A similarity measure for indefinite rankings. *ACM Transactions on Information Systems*, *28*(4), 1–38. **DOI:** https://doi.org/10.1145/1852102.1852106

Wu, J. (2019). Infrastructure of scientometrics: The big and network picture. *Journal of Data and Information Science*, *4*(4), 1–12. **DOI:** https://doi.org/10.2478/jdis-2019-0017

Wu, J., Zhang, Y., & Lv, X. (2019). On comprehensive world university ranking based on ranking aggregation (in Chinese). *Journal of Higher Education Research*, *2011*(1).

Wu, L., Wang, D., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. *Nature*, *566*(7744), 378. **DOI:** https://doi.org/10.1038/s41586-019-0941-9, **PMID:** 30760923

Xiao, Y., Deng, H.-Z., Lu, X., & Wu, J. (2019). Graph-based rank aggregation method for high-dimensional and partial rankings. *Journal of the Operational Research Society*, 1–10. **DOI:** https://doi.org/10.1080/01605682.2019.1657365

Xiao, Y., Deng, Y., Wu, J., Deng, H.-Z., & Lu, X. (2017). Comparison of rank aggregation methods based on inherent ability. *Naval Research Logistics*, *64*(7), 556–565. **DOI:** https://doi.org/10.1002/nav.21771

Yang, G., Fukuyama, H., & Song, Y. (2018). Measuring the inefficiency of Chinese research universities based on a two-stage network DEA model. *Journal of Informetrics*, *12*(1), 10–30. **DOI:** https://doi.org/10.1016/j.joi.2017.11.002

Zhang, S., Liu, Z., & Zhou, X. (2011). Correlation analysis on domestic and foreign university rankings: An empirical study of China's forty universities' sequence in the four major rankings (in Chinese). *China Agricultural Education*, *2011*(2), 8–12.

Zhang, W., Hua, X., & Shao, Y. (2011). The influence of university ranking on students' choice of schools and university enrollment—taking the university rankings of US News and World Report as an example (in Chinese). *Higher Education Exploration*, *5*, 44–47.