Shifting data sources in the rankings of universities

Chun-Kai (Karl) Huang[1], Cameron Neylon[1], Lucy Montgomery[1], Katie Wilson[1], Richard Hosking[1], Alkim Ozaygen[1], Chloe Brookes-Kenworthy[1]

[1]Curtin University, Perth, Australia, email: karl.huang@curtin.edu.au

DESCRIPTION

This study examines the effects of shifting data sources in the rankings of universities. In particular, universities are ranked according to their levels of average citation counts (ACC) and open access (OA) for publications indexed by each of three different bibliographic sources, namely Web of Science (WoS), Scopus and Microsoft Academic (MSA). Metadata on ACC and OA are retrieved from two external databases, i.e., Unpaywall and OpenCitations, to avoid internal biases. The results demonstrate how changing the data source can have significant impact on the perceived performance of individual universities. This work forms part of the Curtin Open Knowledge Initiative (COKI) project [1], spearheaded by Curtin University.

MAIN FINDINGS AND CONTRIBUTION

There is limited literature on comparing the coverages (in terms of both publications and citations) of WoS, Scopus and MSA. The scale of these studies ranges from publications by a single researcher [2], publications associated with a group of academics [3, 4], to outputs that are affiliated to one university [5]. The current work extends the above by expanding the target set to several universities and also adding OA as an additional metric of interest, with a focus on how shifting data sources can affect the rankings of universities. A further innovation is the use of external data sources (i.e., Unpaywall and OpenCitations) for measuring ACC and OA to avoid internal biases.

The results show that, while overall correlations may be high across data sources, there remains a significant number of cases where the ranking of a university drastically changes from source to source, i.e., Simpson's paradox. This implies universities can potentially choose sources to their advantage in small scale comparisons, and the reliance on a single source for a large scale evaluation is also notionally flawed.

METHODOLOGY

Figure 1 provides a quick summary of our data collection and processing framework. Firstly, identifiers for a set of 155 universities are manually collected from WoS, Scopus and MSA and matched against their corresponding identifiers in the Global Research Identifier Database (GRID). Subsequently, an automated process (through APIs) is used to query lists of publications associated with each university from the three bibliographic databases. This produces a list of unique GRID-DOI pairs. These are then matched against data dumps of Unpaywall and OpenCitations. For the purpose of this study, all publications are from 2016, as per date of publication recorded in each bibliographic data source.

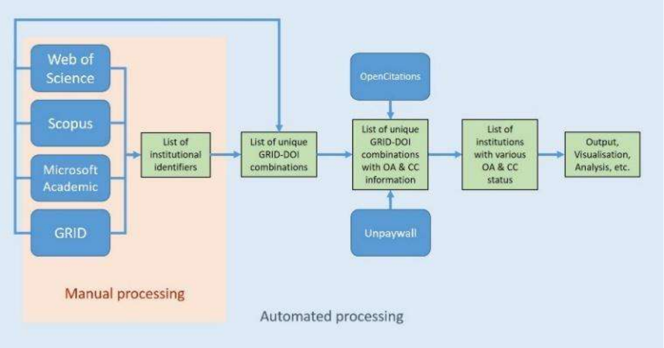
Web of Science
Scopus
Microsoft Academic
GRID
List of institutional identifiers
List of unique GRID-DOI combinations
OpenCitations
List of unique GRID-DOI combinations with OA & CC information
Unpaywall
List of institutions with various OA & CC status
Output, Visualisation, Analysis, etc.
Manual processing
Automated processing

Figure 1: Summary of data collection and processing

RESULTS

Figures 2 and 3 present some of the main results in our analysis. As an example, we select 15 universities[1] (ranging in geography, prestige and size) to illustrate how universities shift in ranks according the source of DOIs. This is presented

[1] The sample of 15 universities include: Cairo University, Curtin University, Dalian University of Technology (DUT), Indian Institute of Science Bangalore (IISC), Institut Teknologi Bandung (ITB), Loughborough University (LU), Massachusetts Institute of Technology (MIT), Moscow State University (MSU), National Autonomous University of Mexico (UNAM), University College London (UCL), University of Cape Town (UCT), University of Giessen, University of Sao Paulo (USP), University of Tokyo, Wayne State University (WSU).

in Figure 2. Evidently, positions of the top 3 ranked universities in either ranking remain unchanged across WoS, Scopus and MSA. However, there are definite cases of universities shifting significantly in ranks (for both ACC and OA) across the sources. In fact, this mimics what we also observe for the larger set of 155 universities, where the most noticeable changes are linked to medium to lower ranked universities in both rankings. Figure 3 shows the distributions of the sizes of rank shifts (as per ACC and OA levels) per pair of bibliographic databases (e.g., the first boxplot in each chart represent sizes of shifts in rank when the source of DOIs changes from WoS to Scopus). All boxplots are characterised by high central peak and long tails, indicating a significant number of extreme cases. This is also evidence for the potential existence of separate groupings of universities (those that are highly affected by shift in sources, and those that are less impacted).

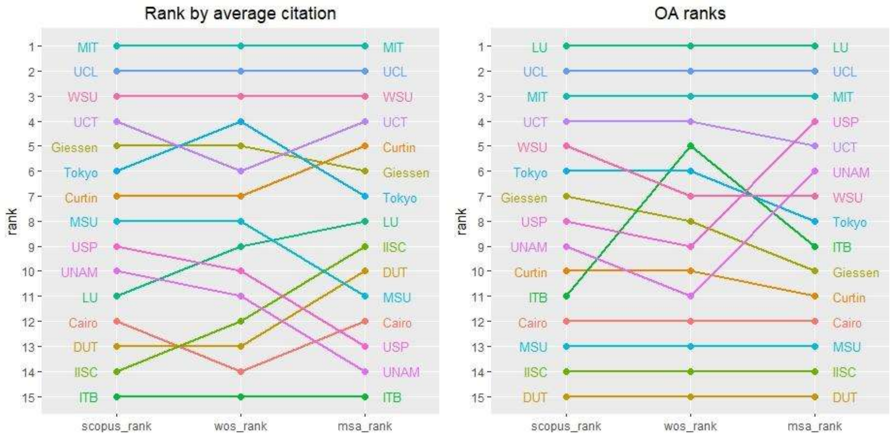

Figure 2: Changes to ACC (left) and OA (right) ranks for a sample of 15 universities.

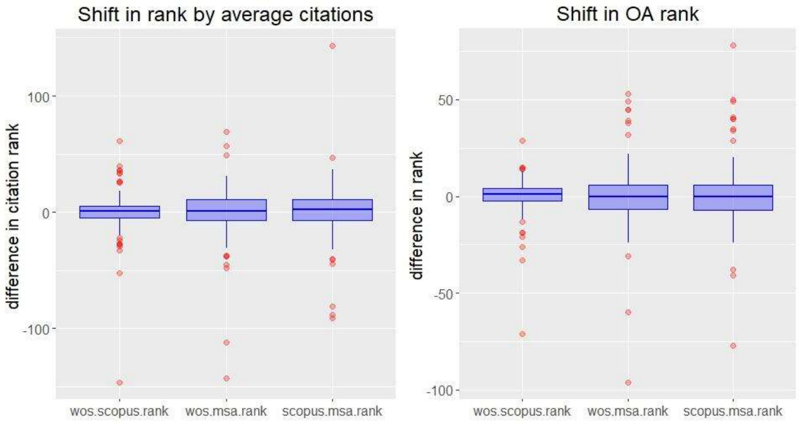

Figure 3: Distributions of rank changes in ACC (left) and OA (right) for 155 universities.

CONCLUSION AND FURTHER WORK

Our analysis finds that the choice of bibliographic source can have substantial impact on the perceived performance of a significant number of universities, in terms of both ACC and OA ranks. In particular, there is evidence that such effects are most pronounced for universities that are medium to lower ranked, non-European and non-English. This signals the need to combine and supplement data sources for more robust and fairer evaluation metrics and frameworks.

REFERENCES

1. Montgomery L, Hartley J, Neylon C, Gillies M, Gray E, Hermann-Pillath C, Huang C-K, Leach J, Potts J, Ren X, Skinner K, Sugimoto CR & Wilson K (2018) *Open Knowledge Institutions: Reinventing Universities*. Work in Progress, MIT Press. https://wip.mitpress.mit.edu/oki
2. Harzing AW (2016) Microsoft Academic (Search): a phoenix arisen from the ashes? Scientometrics 108(3): 1637-1647. https://doi.org/10.1007/s11192-016-2026-y
3. Harzing AW & Alakangas S (2017) Microsoft Academic: is the phoenix getting wings? Scientometrics 110(1): 371-383. https://doi.org/10.1007/s11192-016-2185-x
4. Harzing AW & Alakangas S (2017) Microsoft Academic is one year old: the phoenix is ready to leave the nest. Scientometrics 112(3): 1887-1894. https://doi.org/10.1007/s11192-017-2454-3
5. Hug SE & Brändle MP (2017) The coverage of Microsoft Academic: analyzing the publication output of a university. Scientometrics 113(3): 1551-1571. https://doi.org/10.1007/s11192-017-2535-3